TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN
BỘ MÔN ĐIỀU KHIỂN KINH TẾ
NGUYỄN CAO VĂN (Chủ biên) - TS. TRẦN THÁI NINH

GIÁO TRÌNH

LÝ THUYẾT XÁC SUẤT & THỐNG KÊ TOÁN

NHÀ XUẤT BẢN THỐNG KÊ

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

GIÁO TRÌNH

LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN

(IN LẦN THỨ HAI)

NHÀ XUẤT BẢN THỐNG KÊ

LỜI NÓI ĐẦU

"Giáo trình Lý thuyết xác suất và thống kê toán" được biên soạn cho sinh viên kinh tế sau khi đã được trang bị các kiến thức cơ bản về toán cao cấp bao gồm giải tích cổ điển và đại số tuyến tính.

Mục đích của giáo trình là trang bị cho các nhà kinh tế tương lai phần đảm bảo về toán học cho quá trình thu thập và xử lý thông tin kinh tế - xã hội sẽ được tiếp tục nghiên cứu trong các giáo trình khác như Lý thuyết thống kê, Dự báo kinh tế, Dân số học, Marketing... Nó cũng chuẩn bị các kiến thức cho sinh viên tiếp thu các giáo trình Mô hình toán kinh tế sẽ nghiên cứu ở các năm sau như Kinh tế lượng, Lý thuyết phục vụ công cộng, Lý thuyết quản lý dự trữ...

Ra đời từ thế kỷ 17, lý thuyết xác suất nghiên cứu quy luật của các hiện tượng ngẫu nhiên. Dựa vào các thành tựu của lý thuyết xác suất, thống kê toán xây dựng các phương pháp ra quyết định trong điều kiện thông tin không đầy đủ. Hơn 300 năm phát triển, đến nay nội dung và các phương pháp xác suất và thống kê toán rất phong phú, được ứng dụng rộng rãi trong nhiều lĩnh vực tự nhiên và xã hội khác nhau. Do khuôn khổ có hạn, giáo trình chỉ đề cập những nội dung cơ bán nhất mà nhà kinh tế hoặc kinh doanh không thể thiếu trong hành trang của mình. Những vấn đề không được

đề cập như lý thuyết các quá trình ngẫu nhiên, phương pháp phân tích nhân tố, phương pháp thành phần chính... bạn đọc có thể tiếp tục nghiên cứu ở các tài liệu đầy đủ hơn về xác suất và thống kê toán.

Giáo trình được viết theo quan điểm thực hành, chú trọng việc áp dụng các phương pháp của xác suất, thống kê toán trong nghiên cứu kinh tế hơn là trình bày dưới dạng thuần túy toán học. Mỗi khái niệm, vấn đề hay phương pháp đều được minh họa bằng nhiều thí dụ trong những lĩnh vực thực tế khác nhau nhằm giới thiệu khả năng ứng dụng rộng rãi của các phương pháp đó, đồng thời chứng tỏ ưu thế của việc sử dụng các phương pháp toán nói chung và xác suất thống kê nói riêng trong việc giải quyết các vấn đề thực tiễn. Riêng đối với sinh viên kinh tế thì điều này lại càng có ý nghĩa, nhất là khi nước ta đang chuyển mạnh sang nền kinh tế thị trường.

Để tạo điều kiện cho các nhà kinh tế tương lai sử dụng các phương pháp thống kê toán thuận lợi trong điều kiện được trang bị các phương tiện xử lý thông tin hiện đại, trong giáo trình đã đưa thêm phần đảm bảo chương trình cho các phương pháp được xét. Phần mềm này được viết bằng ngôn ngữ thông dụng và có thể sử dụng được ở mọi loại máy vi tính đang phổ biến ở Việt Nam hiện nay.

Mặc dù đối tượng phục vụ của giáo trình là sinh viên kinh tế, nó vẫn có thể có ích cho tất cả những ai trong công việc hoặc trong nghiên cứu phải tiến hành thu thập và xử lý một khối lượng lớn thông tin, số liệu.

Tương ứng với giáo trình này chúng tôi đã xuất bản một tuyển tập các bài tập. Việc phân công biên soạn giáo trình này như sau:

- PGS. TS. NGUYỄN CAO VĂN : Chủ biên và viết các chương I, II, III, IV, V, VI, VII và VIII.

- TS. TRẦN THÁI NINH : Viết các chương IX và X.

Trong lần xuất bản này chúng tôi đã nhận được nhiều ý kiến đóng góp quý báu của các đồng nghiệp ở Bộ môn Điều khiển kinh tế - trường Đại học Kinh tế Quốc dân và ở nhiều trường Đại học khác. Chúng tôi chân thành cám ơn tất cả những đóng góp đó. Tuy vậy chắc chắn giáo trình không tránh khỏi những hạn chế và thiếu sót. Chúng tôi mong tiếp tục nhận được các ý kiến nhận xét, phê bình của bạn đọc để tiếp tục hoàn thiện nội dung của giáo trình.

CÁC TÁC GIẢ

Phần thứ nhất

LÝ THUYẾT XÁC SUẤT

Lý thuyết xác suất là bộ môn toán học xác lập những quy luật tất nhiên ẩn dấu sau những hiện tượng mang tính ngẫu nhiên khi nghiên cứu một số lớn lần lặp lại cùng các hiện tượng ấy. Việc nắm bắt các quy luật này sẽ cho phép dự báo các hiện tượng ngẫu nhiên đó sẽ xảy ra như thế nào.

Các phương pháp của lý thuyết xác suất được ứng dụng rộng rãi trong việc giải quyết các bài toán thuộc nhiều lĩnh vực khác nhau của khoa học tự nhiên, kỹ thuật và kinh tế – xã hội.

Chương I

BIẾN CỐ NGẪU NHIÊN VÀ XÁC SUẤT

§1. PHÉP THỬ VÀ CÁC LOẠI BIẾN CỐ

Trong tự nhiên và xã hội, mỗi hiện tượng đều gắn liền với một nhóm các điều kiện cơ bản và các hiện tượng đó chỉ có thể xảy ra khi nhóm các điều kiện cơ bản gắn liền với nó được thực hiện. Do đó, khi muốn nghiên cứu một hiện tượng, ta cần thực hiện nhóm các điều kiện cơ bản ấy. Chẳng hạn, nếu muốn quan sát việc xuất hiện mặt sấp hay mặt ngửa của một đồng xu, ta phải tung đồng xu xuống đất; còn để xét xem viên đạn trúng bia hay trượt, ta phải bắn các viên đạn; khi muốn nghiên cứu chất lượng của một lô sản phẩm, ta phải lấy ngẫu nhiên một hoặc một số sản phẩm của lô sản phẩm đó v.v...

Việc thực hiện một nhóm các điều kiện cơ bản để quan sát một hiện tượng nào đó có xảy ra hay không được gọi là thực hiện một phép thử, còn hiện tượng có thể xảy ra trong kết quả của phép thử đó được gọi là biến cố.

Thí dụ 1. Tung một con xúc xắc xuống đất là một phép thử, còn việc lật lên một mặt nào đó là biến cố.

Thí dụ 2. Bắn một phát súng vào bia. Việc bắn súng là phép thử, còn việc trúng vào một miền nào đó của bia là biến cố.

Thí dụ 3. Từ một lô sản phẩm gồm chính phẩm và phế phẩm lấy ngẫu nhiên một sản phẩm. Việc lấy sản phẩm là phép thử, còn việc lấy được chính phẩm hay phế phẩm là biến cố.

Như vậy, ta thấy rằng một biến cố chỉ có thể xảy ra khi một phép thử gắn liền với nó được thực hiện. Trong thực tế có thể xảy ra các loại biến cố sau đây:

+ *Biến cố chắc chắn:* Là biến cố nhất định sẽ xảy ra khi thực hiện một phép thử. Biến cố chắc chắn được ký hiệu là U.

Thí dụ 4. Thực hiện phép thử tung một con xúc xắc. Gọi U là biến cố "Xuất hiện mặt có số chấm nhỏ hơn hoặc bằng 6", U là biến cố chắc chắn.

+ *Biến cố không thể có:* Là biến cố nhất định không xảy ra khi thực hiện phép thử. Biến cố không thể có được ký hiệu là V.

Thí dụ 5. Tung một con xúc xắc, gọi V là biến cố "Xuất hiện mặt có 7 chấm", V là biến cố không thể có.

+ *Biến cố ngẫu nhiên:* Là biến cố có thể xảy ra hoặc không xảy ra khi thực hiện một phép thử. Các biến cố ngẫu nhiên được ký hiệu là A, B, C... hoặc $A_1, A_2, ..., A_n, B_1, B_2, ..., B_m$.

Thí dụ 6. Tung một con xúc xắc, gọi A là biến cố "Xuất hiện mặt 1 chấm", A là biến cố ngẫu nhiên.

Thí dụ 7. Bắn một phát đạn vào bia, gọi B là biến cố "Trúng vòng 10", B là biến cố ngẫu nhiên.

Tất cả các biến cố mà chúng ta gặp trong thực tế đều thuộc về một trong ba loại biến cố kể trên. Tuy nhiên, các biến cố ngẫu nhiên là các biến cố thường gặp hơn cả.

§2. XÁC SUẤT CỦA BIẾN CỐ

Như trên đã thấy, việc biến cố ngẫu nhiên xảy ra hay không xảy ra trong kết quả của phép thử là điều không thể đoán trước được. Tuy nhiên, bằng trực quan ta có thể nhận thấy các biến cố ngẫu nhiên khác nhau có những khả năng xảy ra khác nhau. Chẳng hạn biến cố "Xuất hiện mặt sấp" khi tung một đồng xu sẽ có khả năng xảy ra lớn hơn nhiều so với biến cố "Xuất hiện mặt một chấm" khi tung một con xúc xắc. Hơn nữa, khi lặp đi lặp lại nhiều lần cùng một phép thử trong những điều kiện như nhau, người ta thấy tính chất ngẫu nhiên của biến cố mất dần đi và khả năng xảy ra của biến cố sẽ được thể hiện theo những quy luật nhất định. Từ đó ta thấy có khả năng định lượng (đo lường), khả năng khách quan xuất hiện một biến cố nào đó.

Xác suất của một biến cố là một con số đặc trưng khả năng khách quan xuất hiện biến cố đó khi thực hiện phép thử.

Ta chú ý rằng đây là khả năng khách quan, do những điều kiện xảy ra của phép thử quy định chứ không tùy thuộc vào ý muốn chủ quan của con người.

Như vậy, bản chất xác suất của một biến cố là một con số xác định. Để tính xác suất của một biến cố, người ta xây dựng các định nghĩa và định lý sau đây:

§3. ĐỊNH NGHĨA CỔ ĐIỂN VỀ XÁC SUẤT

3.1. Thí dụ

Giả sử thực hiện một phép thử là tung một con xúc xắc đều đặn và đồng chất. Gọi A là biến cố "Xuất hiện mặt chẵn chấm". Ta phải xác định xác suất của biến cố A.

Khi tung một con xúc xắc đều đặn và đồng chất, ta thấy có 6 trường hợp có thể xảy ra là: Xuất hiện các mặt 1 chấm, 2 chấm, ..., 6 chấm. Những trường hợp này thỏa mãn hai điều kiện: Trước hết chúng duy nhất, tức là trong kết quả của phép thử xảy ra một và chỉ một trong các trường hợp đó. Sau nữa đây là những trường hợp có khả năng xảy ra như nhau. Các trường hợp thỏa mãn hai điều kiện nói trên được gọi là các trường hợp (kết cục) duy nhất đồng khả năng.

Trong số 6 kết cục duy nhất đồng khả năng đó ta thấy chỉ có 3 kết cục mà nếu xảy ra thì biến cố A sẽ xảy ra, đó là những kết cục được mặt 2 chấm. 4 chấm và 6 chấm. Những kết cục làm cho biến cố xảy ra được gọi là các kết cục thuận lợi cho biến cố.

Như vậy đứng về mặt trực quan ta thấy khả năng xảy ra của biến cố A là 3 phần 6, tức là 1 phần 2. Đó chính là cách xác định xác suất của biến cố theo quan điểm cổ điển.

3.2. Định nghĩa

Xác suất xuất hiện biến cố A trong một phép thử *là tỉ số giữa số kết cục thuận lợi cho A và tổng số các kết cục duy nhất đồng khả năng có thể xảy ra khi thực hiện phép thử đó.*

Nếu ký hiệu P(A) là xác suất của biến cố A, m là số kết cục thuận lợi cho biến cố A, n là số kết cục duy nhất đồng khả năng của phép thử, ta có công thức sau:

$$P(A) = \frac{m}{n} \qquad (1.1)$$

3.3. Các tính chất của xác suất

Từ định nghĩa cổ điển về xác suất ta có thể suy ra các tính chất sau đây:

a. Xác suất của biến cố ngẫu nhiên là một số dương nằm trong khoảng giữa 0 và 1.

$$0 < P(A) < 1$$

Thật vậy, vì số kết cục thuận lợi cho một biến cố ngẫu nhiên luôn luôn thỏa mãn $0 < m < n$ do đó:

$$0 < \frac{m}{n} < 1$$

từ đó: $0 < P(A) < 1$.

b. Xác suất của biến cố chắc chắn bằng một.

$$P(U) = 1$$

Thật vậy, nếu U là biến cố chắc chắn thì tất cả các kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử đều thuận lợi cho biến cố xảy ra. Do đó $m = n$ và ta có:

$$P(U) = \frac{m}{n} = 1.$$

c. Xác suất của biến cố không thể có bằng không.

$$P(V) = 0$$

Thật vậy, nếu V là biến cố không thể có thì trong số các kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử không có kết cục nào thuận lợi cho biến cố xảy ra. Như vậy m = 0, do đó:

$$P(V) = \frac{m}{n} = 0$$

Như vậy, xác suất của một biến cố bất kỳ luôn thỏa mãn điều kiện:

$$0 \le P(A) \le 1 \qquad (1.2)$$

Đối với các tính chất trên ta chú ý rằng mệnh đề đảo của hai tính chất b và c chưa chắc đã đúng. Tức là, nếu một biến cố có xác suất bằng 1 thì chưa chắc đã là biến cố chắc chắn và nếu một biến cố có xác suất bằng 0 thì chưa chắc đã là biến cố không thể có.

3.4. Các phương pháp tính xác suất bằng định nghĩa cổ điển

1. Phương pháp suy luận trực tiếp: Nếu số các kết cục trong phép thử là khá nhỏ và việc suy đoán là khá đơn giản thì có thể sử dụng phương pháp suy luận trực tiếp.

Thí dụ 1. Trong bình có a quả cầu trắng và b quả cầu đen. Lấy ngẫu nhiên một quả cầu. Tìm xác suất để lấy được quả cầu trắng.

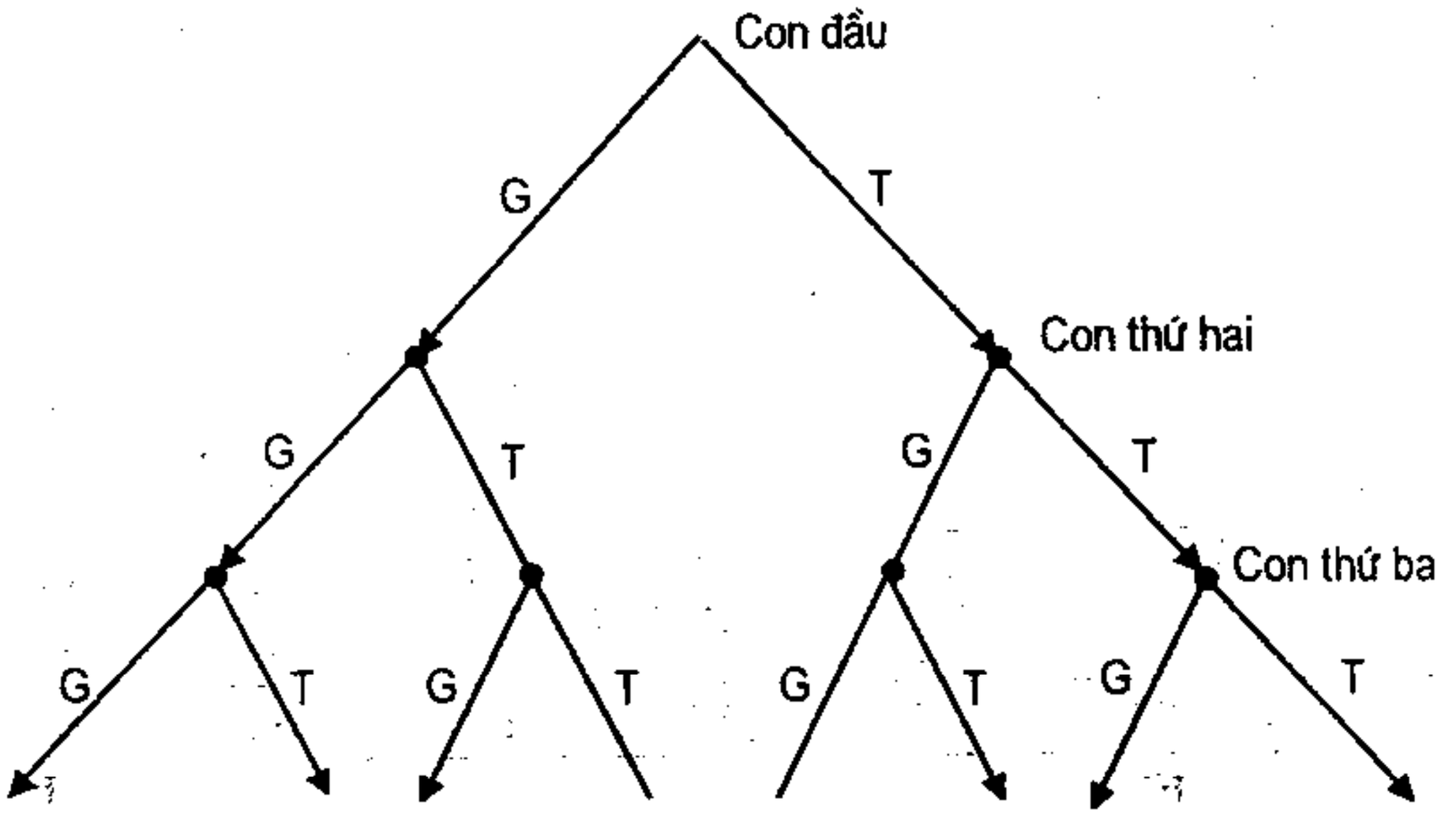
Giải. Gọi A là biến cố "Lấy được quả cầu trắng". Khi lấy ngẫu nhiên từ bình ra một quả cầu, ta có thể lấy được bất kỳ quả nào trong số a + b quả cầu có trong bình. Như vậy số kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử n = a + b.

Biến cố A sẽ xảy ra khi ta lấy được một trong số a quả cầu trắng. Như vậy số kết cục thuận lợi m = a. Từ đó theo định nghĩa cổ điển về xác suất, ta có:

$$P(A) = \frac{m}{n} = \frac{a}{a+b}$$

2. Phương pháp dùng sơ đồ Venn: Khi số kết cục là khá lớn và việc suy đoán phức tạp hơn thì có thể dùng sơ đồ Venn, tức là mô tả các kết cục của phép thử dưới dạng sơ đồ để dễ nhận biết hơn. Trong thực tế có thể dùng các loại sơ đồ sau:

a. Sơ đồ hình cây



Hình 1.1. **Sơ đồ hình cây**

Thí dụ 2. Giả sử xác suất sinh con trai và con gái là như nhau. Một gia đình có 3 con. Tìm xác suất để gia đình đó có 2 con gái.

Giải. Gọi A là biến cố "Gia đình đó có 2 con gái". Số kết cục đồng khả năng có thể suy ra từ sơ đồ trên hình 1.1.

Như vậy tổng số ta có n = 8 kết cục đồng khả năng là GGG, GGT, GTG, GTT, TGG, TGT, TTG và TTT. Trong đó có 3 kết cục thuận lợi để có 2 con gái.

Vậy $P(A) = \frac{3}{8}$.

b. Sơ đồ dạng bảng

Thí dụ 3. Tung một con xúc xắc hai lần. Tìm xác suất để trong đó có một lần được 6 chấm.

Giải. Gọi A là biến cố "Trong 2 lần tung con xúc xắc có 1 lần được mặt 6 chấm". Số kết cục đồng khả năng của phép thử có thể mô tả dưới dạng bảng sau (hình 1.2):

I \ II	1	2	3	4	5	6
1	11	12	13	14	15	16
2	21	22	23	24	25	26
3	31	32	33	34	35	36
4	41	42	43	44	45	46
5	51	52	53	54	55	56
6	61	62	63	64	65	66

***Hình 1.2.* Sơ đồ dạng bảng**

Như vậy ta có n = 36 kết cục đồng khả năng. Trong đó có m = 10 kết cục thuận lợi. Vậy:

$$P(A) = \frac{m}{n} = \frac{10}{36} = \frac{5}{18}$$

c. Sơ đồ dạng tập hợp

Thí dụ 4. Trong một lớp 50 học sinh có:

20 người chơi bóng đá

15 người chơi bóng chuyền

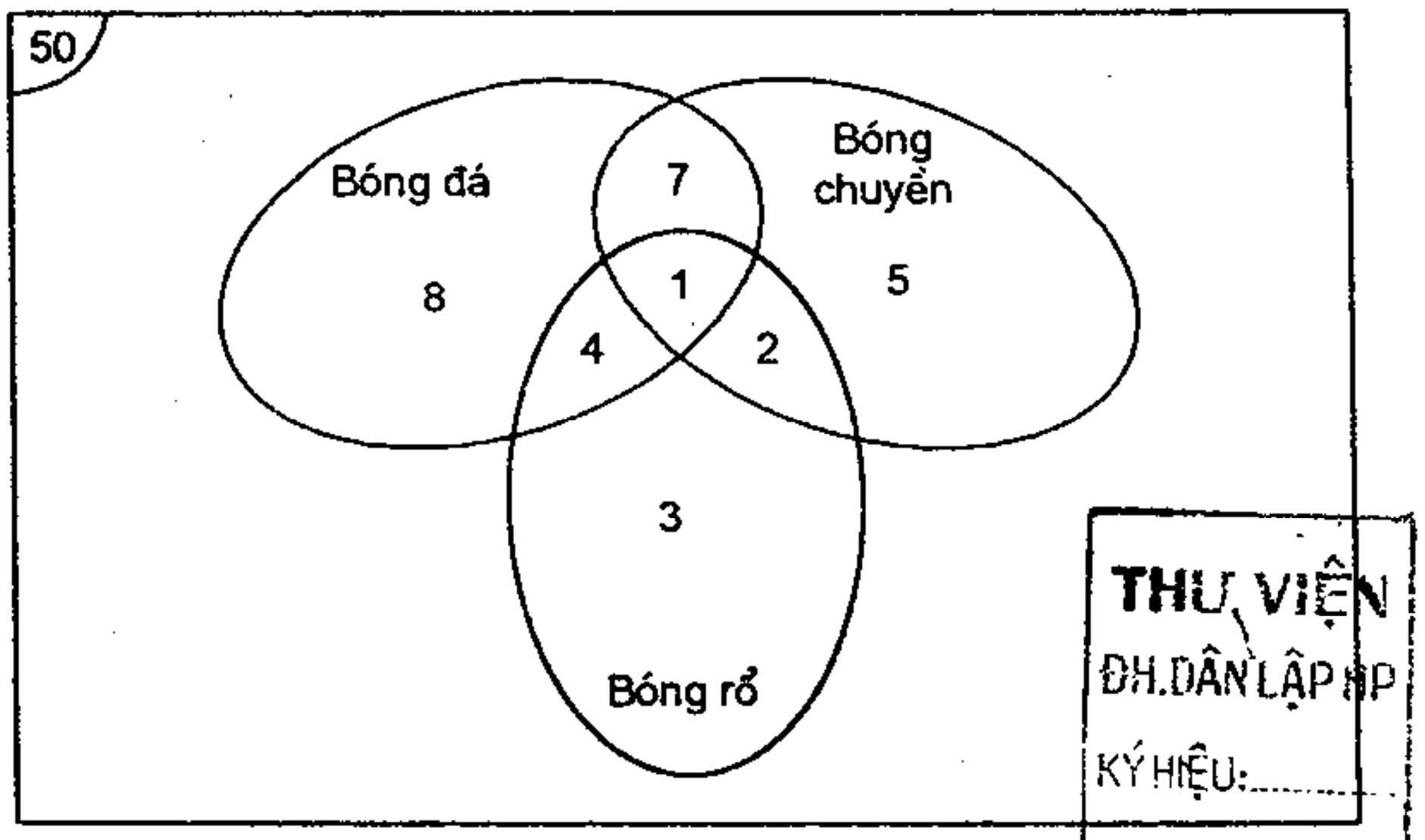
10 người chơi bóng rổ

8 người chơi bóng đá và bóng chuyền

5 người chơi bóng đá và bóng rổ

3 người chơi bóng chuyền và bóng rổ

1 người chơi bóng đá, bóng chuyền và bóng rổ.



***Hình 1.3.* Sơ đồ dạng tập hợp**

Lấy ngẫu nhiên 1 học sinh. Tìm xác suất để người đó chơi ít nhất 1 môn bóng.

Giải. Gọi A là biến cố "Lấy ngẫu nhiên một học sinh thì

người đó chơi ít nhất một môn bóng". Số kết cục đồng khả năng có thể mô tả dưới dạng tập hợp như trên hình 1.3.

Vậy trong n = 50 kết cục đồng khả năng thì số kết cục thuận lợi m = 8 + 5 + 3 + 7 + 4 + 2 + 1 = 30.

Vậy $$P(A) = \frac{m}{n} = \frac{30}{50} = 0,6.$$

3. Phương pháp dùng các công thức của giải tích tổ hợp

Nếu số kết cục của phép thử là rất lớn mà không thể suy đoán trực tiếp được thì có thể dùng các công thức của giải tích tổ hợp, chủ yếu là các công thức chỉnh hợp, chỉnh hợp lặp, hoán vị và tổ hợp để tính toán.

Thí dụ 5. Một người khi gọi điện thoại quên mất hai số cuối của số điện thoại và chỉ nhớ được rằng chúng khác nhau. Tìm xác suất để quay ngẫu nhiên một lần được đúng số cần gọi.

Giải. Gọi B là biến cố "Quay ngẫu nhiên một lần được đúng số cần gọi". Số kết cục đồng khả năng là tất cả các phương thức để lập nên một cặp hai số khác nhau từ 10 số tự nhiên đầu tiên. Nó bằng số chỉnh hợp chập 2 từ 10. Như vậy $n = A_{10}^2 = 10.9 = 90$. Còn số kết cục thuận lợi cho biến cố B xảy ra chỉ có một kết cục. Do đó theo định nghĩa cổ điển:

$$P(B) = \frac{m}{n} = \frac{1}{90}$$

Thí dụ 6. Trong bình có 6 quả cầu giống nhau được đánh số, lấy ngẫu nhiên lần lượt từng quả cầu. Tìm xác suất để số của quả cầu được lấy ra trùng với số thứ tự của lần lấy.

Giải. Gọi A là biến cố "Số của các quả cầu trùng với số thứ tự của lần lấy". Số kết cục đồng khả năng trong phép thử này là tất cả các phương thức để lần lượt lấy được 6 quả cầu ra khỏi bình. Nó bằng số hoán vị của 6 phần tử. Do đó $n = P_6 = 6! = 720$. Trong đó chỉ có một kết cục thuận lợi cho biến cố A xảy ra là lấy được các quả cầu theo trình tự các số 1, 2, 3, 4, 5, 6.

Như vậy: $$P(A) = \frac{m}{n} = \frac{1}{720}$$

Thí dụ 7. Một hộp có 10 sản phẩm, trong đó có 6 chính phẩm và 4 phế phẩm. Lấy ngẫu nhiên từ hộp đó 3 sản phẩm. Tìm xác suất để:

a) Cả 3 sản phẩm lấy ra đều là chính phẩm.

b) Trong ba sản phẩm lấy ra có đúng 2 chính phẩm.

Giải. a) Gọi A là biến cố "Lấy được 3 chính phẩm". Số kết cục đồng khả năng trong phép thử bằng số tổ hợp chập 3 từ 10 phần tử. Như vậy $n = C_{10}^3 = 120$. Số kết cục thuận lợi cho A xảy ra bằng số tổ hợp chập 3 (chính phẩm) từ 6 (chính phẩm) cho trước. Vậy $m = C_6^3 = 20$.

Do đó:

$$P(A) = \frac{m}{n} = \frac{20}{120} = \frac{1}{6}.$$

b) Gọi B là biến cố "Trong 3 sản phẩm lấy ra có đúng 2 chính phẩm". Số kết cục thuận lợi cho B xảy ra bằng số tổ hợp chập 2 (chính phẩm) từ 6 (chính phẩm) cho trước, bằng C_6^2. Ngoài ra sản phẩm thứ ba phải là phế phẩm. Ta có C_4^1 cách lấy được 1 phế phẩm. Như vậy số kết cục thuận lợi cho B xảy ra bằng $m = C_6^2 . C_4^1$.

Do đó: $$P(B) = \frac{m}{n} = \frac{C_6^2 . C_4^1}{C_{10}^3} = \frac{1}{2}$$

Thí dụ 8. Trong 3 tháng cuối năm biết rằng có 5 máy đã bị hỏng. Tìm xác suất để không có ngày nào có quá 1 máy bị hỏng.

Giải. Gọi A là biến cố "Không có ngày nào có quá 1 máy bị hỏng". Số kết cục đồng khả năng là số chỉnh hợp lặp chập 5 từ 92 phần tử ($n = \overline{A}_{92}^5 = 92^5$).

Số kết cục thuận lợi là số chỉnh hợp chập 5 từ 92 phần tử ($m = A_{92}^5 = 88.89.90.91.92$).

Vậy: $$P(A) = \frac{m}{n} = \frac{88.89.90.91.92}{92^5} = 0,8954.$$

3.5. Ưu điểm và hạn chế của định nghĩa cổ điển về xác suất

Định nghĩa cổ điển về xác suất có một ưu điểm cơ bản là để tìm xác suất của biến cố ta không cần phải tiến hành phép thử (phép thử chỉ tiến hành một cách giả định). Ngoài ra nếu đáp ứng đầy đủ các yêu cầu của định nghĩa thì nó cho phép ta tìm được một cách chính xác giá trị của xác suất.

Tuy nhiên định nghĩa cổ điển về xác suất cũng có những hạn chế đáng kể. Nó đòi hỏi là số kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử phải là hữu hạn. Trong thực tế có nhiều phép thử mà trong đó số kết cục có thể là vô hạn. Trong những trường hợp này định nghĩa cổ điển về xác suất không áp dụng được. Chỉ riêng điều đó đã hạn chế khả năng áp dụng của định nghĩa cổ điển. Thật ra hạn chế này có thể khắc phục được bằng cách mở rộng định nghĩa cổ điển.

Hạn chế lớn nhất của định nghĩa cổ điển là trong thực tế nhiều khi không thể biểu diễn kết quả của phép thử dưới dạng tập hợp các kết cục duy nhất và đồng khả năng. Thường thì tính đồng khả năng của các kết cục được suy ra từ tính đối xứng. Chẳng hạn khi tung một con xúc xắc ta giả thiết rằng nó đều đặn và đồng chất. Tuy nhiên những bài toán trong đó ta có thể đưa ra các giả thiết về tính đối xứng rất hiếm khi gặp trong thực tế.

Vì lý do đó mà ngoài định nghĩa cổ điển về xác suất, trong thực tế người ta còn sử dụng định nghĩa xác suất theo quan điểm thống kê sau đây.

§4. ĐỊNH NGHĨA THỐNG KÊ VỀ XÁC SUẤT

4.1. Định nghĩa

Tần suất xuất hiện biến cố trong n phép thử là tỷ số giữa số phép thử trong đó biến cố xuất hiện và tổng số phép thử được thực hiện.

Như vậy, nếu ký hiệu số phép thử là n, số lần xuất hiện biến cố A là k, tần suất xuất hiện biến cố A là f(A) thì:

$$f(A) = \frac{k}{n} \qquad (1.3)$$

Cùng với khái niệm xác suất, khái niệm tần suất là một trong những khái niệm cơ bản cửa lý thuyết xác suất.

Thí dụ 1. Khi kiểm tra ngẫu nhiên 80 sản phẩm do một

máy sản xuất, người ta phát hiện ra 3 phế phẩm. Gọi A là biến cố "Xuất hiện phế phẩm". Vậy tần suất xuất hiện phế phẩm bằng:

$$f(A) = \frac{3}{80}$$

Thí dụ 2. Bắn 50 phát đạn vào bia thấy có 47 phát trúng. Gọi A là biến cố "Bắn trúng bia". Tần suất của việc bắn trúng bia bằng:

$$f(A) = \frac{47}{50}$$

Người ta nhận thấy nếu tiến hành các thí nghiệm trong những điều kiện như nhau và số phép thử khá lớn thì tần suất thể hiện tính ổn định của nó khá rõ ràng. Tính chất này thể hiện ở chỗ là *nếu tiến hành một số khá lớn cùng một phép thử thì tần suất dao động rất ít xung quanh một giá trị nào đó*. Ta sẽ thấy rõ điều đó qua thí dụ sau.

Thí dụ 3. Để nghiên cứu khả năng xuất hiện mặt sấp khi tung một đồng xu, người ta tiến hành tung một đồng xu nhiều lần và thu được kết quả sau đây:

Người làm thí nghiệm	Số lần tung (n)	Số lần được mặt sấp (k)	Tần suất $f(A) = \frac{k}{n}$
Buffon	4040	2048	0,5069
Pearson	12000	6019	0,5016
Pearson	24000	12012	0,5005

Qua thí dụ trên ta thấy khi số phép thử tăng lên thì tần suất xuất hiện mặt sấp sẽ dao động ngày càng ít hơn xung quanh giá trị không đổi là 0,5. Điều đó cho phép hy vọng

rằng khi số phép thử tăng lên vô hạn, tần suất sẽ hội tụ về giá trị 0,5.

Tính ổn định của tần suất là cơ sở để đưa ra định nghĩa thống kê về xác suất.

4.2. Định nghĩa

Xác suất xuất hiện biến cố A trong một phép thử là một số p không đổi mà tần suất f xuất hiện biến cố đó trong n phép thử sẽ dao động rất ít xung quanh nó khi số phép thử tăng lên vô hạn.

Như vậy về mặt thực tế với số phép thử đủ lớn ta có thể lấy:

$$P(A) \approx f(A)$$

Cơ sở của cách lấy xấp xỉ này sẽ được trình bày ở chương V.

4.3. Ưu điểm và hạn chế của định nghĩa thống kê về xác suất

Định nghĩa thống kê về xác suất có ưu điểm lớn là nó không đòi hỏi những điều kiện áp dụng như đối với định nghĩa cổ điển. Nó hoàn toàn dựa trên các quan sát thực tế để làm cơ sở kết luận về xác suất xảy ra của một biến cố.

Tuy nhiên, định nghĩa thống kê về xác suất chỉ áp dụng được đối với các hiện tượng ngẫu nhiên mà tần suất của nó có tính ổn định. Hơn nữa, để xác định một cách tương đối chính xác giá trị của xác suất ta phải tiến hành trên thực tế một số đủ lớn các phép thử. Nói cách khác xác suất theo quan điểm thống kê là xác suất được tính sau khi phép thử đã thực hiện. Trong nhiều bài toán thực tế rất khó hoặc không thể

tiến hành nhiều phép thử để dựa vào đó mà tính xác suất của một biến cố.

Trong nhiều trường hợp nếu không cần thiết phải thực sự tiến hành phép thử thực tế, để khắc phục hạn chế trên người ta có thể mô phỏng kết quả của các phép thử bằng cách sử dụng bảng số ngẫu nhiên (Phụ lục 10). Để làm điều đó người ta chọn ngẫu nhiên một dòng của bảng số ngẫu nhiên và dùng các chữ số của dòng đó để thay thế cho kết quả của phép thử. Chẳng hạn, nếu tiến hành tung một con xúc xắc trong 10 lần thì có thể mô tả kết quả tung bằng cách chọn ngẫu nhiên một dòng của bảng số ngẫu nhiên. Giả sử ta chọn dòng thứ nhất và thu được dãy số sau:

$$1559\underbrace{9}_{\text{bỏ}}\ \underbrace{90}_{\text{bỏ}}\underbrace{68}_{\text{bỏ}}\ \underbrace{92}_{\text{bỏ}}\underbrace{90}_{\text{bỏ}}\ \underbrace{83}_{\text{bỏ}}\underbrace{03}_{\text{bỏ}}\ \underbrace{85}_{\text{bỏ}}\underbrace{08}_{\text{bỏ}}\ \underbrace{89}_{\text{bỏ}}54$$

Vậy kết quả của phép thử có thể mô phỏng như sau: Lần đầu được 1 điểm, lần 2 được 5 điểm, lần 3 được 5 điểm, lần 4 được 6 điểm... và dựa vào kết quả đó để xác định tần suất.

§5. MỘT SỐ ĐỊNH NGHĨA KHÁC VỀ XÁC SUẤT

Trong thực tế ngoài định nghĩa cổ điển và định nghĩa thống kê về xác suất người ta còn sử dụng một số định nghĩa sau về xác suất.

5.1. Định nghĩa hình học về xác suất

Định nghĩa hình học về xác suất có thể sử dụng khi xác suất để một điểm ngẫu nhiên rơi vào một phần nào đó của một miền cho trước tỷ lệ với độ đo của miền đó (độ dài, diện

tích, thể tích v.v...) và không phụ thuộc vào vị trí và dạng thức của miền đó.

Nếu độ đo hình học của toàn bộ miền cho trước là S, còn độ đo hình học của một phần A nào đó của nó là S_A thì xác suất để điểm ngẫu nhiên rơi vào phần A sẽ bằng:

$$p = \frac{S_A}{S}$$

trong đó S và S_A có thể có độ đo bất kỳ.

Như vậy, có thể xem định nghĩa hình học về xác suất là sự mở rộng tương ứng của định nghĩa cổ điển về xác suất.

5.2. Xác suất chủ quan

Xác suất chủ quan được định nghĩa như sự đánh giá chủ quan của một cá nhân nào đó về khả năng xảy ra của biến cố. Sự đánh giá này chủ yếu dựa vào những nhận xét cá nhân, thông tin ngoại lai, trực giác hoặc các kinh nghiệm tích lũy được của mỗi cá nhân liên quan đến hiện tượng được xem xét. Như vậy, với cùng một hiện tượng thì xác suất chủ quan của người này có thể khác biệt rất nhiều so với xác suất chủ quan của người khác, vì vậy, nó còn được gọi là xác suất của cá nhân.

Cách tiếp cận này chủ yếu được sử dụng khi không thể áp dụng các phương pháp tính xác suất một cách khách quan, chẳng hạn tình huống không thể quan niệm được hết các kết cục có thể có của một phép thử hay không thể lặp lại nhiều lần một phép thử để xác định tần suất xuất hiện biến cố.

5.3. Định nghĩa tiên đề về xác suất

Vào những năm 30 của thế kỷ 20, nhà toán học người

Nga là Kolmogorov đã xây dựng hệ tiên đề làm cơ sở cho việc định nghĩa một cách hoàn chỉnh khái niệm xác suất về mặt lý thuyết. Hệ tiên đề được xây dựng trên cơ sở khái niệm về không gian các biến cố sơ cấp E_1 E_2... E_n, thực tế là tập hợp mọi kết cục có thể có của một phép thử. Lúc đó mỗi biến cố A có thể quan niệm như một tập hợp con của không gian đó. Từ đó ta có các tiên đề sau:

Tiên đề 1: Với mọi biến cố A đều có $P(A) \geq 0$.

Tiên đề 2: Nếu E_1 E_2... E_n tạo nên không gian các biến cố sơ cấp thì:

$$P(E_1) + P(E_2) + ... + P(E_n) = 1$$

Tiên đề 3: Nếu các biến cố A_1 A_2... A_n... là các tập hợp con không giao nhau của các biến cố sơ cấp thì:

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Từ các tiên đề trên có thể xây dựng các định lý cơ bản của xác suất.

§6. NGUYÊN LÝ XÁC SUẤT LỚN VÀ XÁC SUẤT NHỎ

Trong nhiều bài toán thực tế ta thường gặp các biến cố có xác suất rất nhỏ, tức là gần bằng 0. Trong trường hợp đó liệu có thể cho rằng những biến cố có xác suất rất nhỏ sẽ không xảy ra khi thực hiện một phép thử? Tất nhiên là không thể kết luận như vậy, vì như trên đã nêu, thậm chí một biến cố có xác suất bằng không vẫn chưa chắc đã là biến cố không thể có, tức là vẫn có thể xảy ra.

Tuy nhiên qua nhiều lần quan sát người ta thấy rằng các biến cố có xác suất nhỏ gần như sẽ không xảy ra khi tiến hành một phép thử. Trên cơ sở đó có thể đưa ra *"Nguyên lý thực tế không thể có của các biến cố có xác suất nhỏ"* sau đây:

Nếu một biến cố có xác suất rất nhỏ thì thực tế có thể cho rằng trong một phép thử biến cố đó sẽ không xảy ra.

Hiển nhiên là việc quy định một mức xác suất được coi là rất nhỏ sẽ tùy thuộc vào từng bài toán cụ thể. Chẳng hạn, nếu xác suất để dù không mở khi sử dụng bằng 0,01 thì xác suất đó chưa thể coi là nhỏ và chắc chắn là không thể sử dụng loại dù đó. Song nếu xác suất để một chuyến tàu đường dài đến ga chậm bằng 0,01 thì thực tế lại có thể cho rằng tàu sẽ đến ga đúng giờ.

Một xác suất khá nhỏ mà với nó có thể cho rằng biến cố thực tế sẽ không xảy ra được gọi là *mức ý nghĩa*. Tùy thuộc vào từng bài toán thực tế, mức ý nghĩa này có thể được lấy trong khoảng từ 0,01 đến 0,05.

Tương tự như vậy ta có thể đưa ra *"Nguyên lý thực tế chắc chắn xảy ra của các biến cố có xác suất lớn"* như sau: *Nếu biến cố ngẫu nhiên có xác suất gần bằng 1 thì thực tế có thể cho rằng biến cố đó sẽ xảy ra trong một phép thử.* Hiển nhiên là, cũng như ở trên, việc qui định một mức xác suất đủ coi là lớn tùy thuộc vào từng bài toán cụ thể.

§7. ĐỊNH LÝ CỘNG XÁC SUẤT

Ở các mục trước chúng ta đã nghiên cứu các phương pháp tính trực tiếp xác suất của các biến cố bằng các định

nghĩa xác suất. Song những cách tính trực tiếp này không phải là cơ bản trong lý thuyết xác suất. Việc áp dụng chúng không phải lúc nào cũng tiện lợi và có thể dùng được.

Vì vậy, để xác định xác suất của các biến cố người ta thường không áp dụng các phương pháp tính trực tiếp mà áp dụng phương pháp gián tiếp, cho phép tính xác suất của một biến cố dựa vào xác suất đã biết của các biến cố khác có liên quan với nó thông qua các định lý xác suất, thường được gọi là định lý cộng và định lý nhân xác suất.

Định nghĩa 1. Biến cố C được gọi là tổng của hai biến cố A và B, ký hiệu C = A + B nếu C chỉ xảy ra khi có ít nhất một trong hai biến cố A và B xảy ra.

Thí dụ 1. Hai người cùng bắn vào một bia. Gọi A là biến cố "Người thứ nhất bắn trúng", B là biến cố "Người thứ hai bắn trúng", C là biến cố "Bia bị trúng đạn". Rõ ràng là biến cố C sẽ xảy ra khi có ít nhất một trong hai biến cố A và B xảy ra. Vậy C = A + B.

Thí dụ 2. Tung một con xúc xắc. Gọi A là biến cố "Xuất hiện mặt 6 chấm", B là biến cố "Xuất hiện mặt 5 chấm", C là biến cố "Được ít nhất 5 chấm". Biến cố C xảy ra khi hoặc A hoặc B xảy ra. Vậy C = A + B.

Định nghĩa 2. Biến cố A được gọi là tổng của n biến cố A_1 A_2... A_n nếu A xảy ra khi có ít nhất một trong n biến cố ấy xảy ra.

Ký hiệu $A = \sum_{i=1}^{n} A_i$.

Gắn liền với khái niệm tổng các biến cố là khái niệm về sự xung khắc của các biến cố.

Định nghĩa 3. Hai biến cố A và B gọi là xung khắc với nhau nếu chúng không thể đồng thời xảy ra trong một phép thử. Trường hợp ngược lại, nếu hai biến cố có thể cùng xảy ra trong một phép thử thì được gọi là không xung khắc.

Thí dụ 3. Trong thí dụ 1, khi thực hiện phép thử là cho hai người cùng bắn vào một bia và gọi A và B tương ứng là các biến cố người thứ nhất và thứ hai bắn trúng bia thì hiển nhiên A và B là không xung khắc. Mặt khác, trong thí dụ 2, khi tung một con xúc xắc và gọi A và B tương ứng là các biến cố được 6 chấm và 5 chấm thì A và B xung khắc với nhau.

Thí dụ 4. Một bình có 3 loại cầu là cầu trắng, cầu xanh và cầu đỏ. Lấy ngẫu nhiên từ bình đó một quả cầu. Gọi A là biến cố "Lấy được cầu trắng", B là biến cố "Lấy được cầu xanh", A và B là hai biến cố xung khắc với nhau.

Khi áp dụng khái niệm xung khắc cho nhóm gồm n biến cố, ta có khái niệm xung khắc từng đôi.

Định nghĩa 4. Nhóm n biến cố A_1 A_2... A_n được gọi là xung khắc từng đôi nếu bất kỳ hai biến cố nào trong nhóm này cũng xung khắc với nhau.

Chú ý rằng việc nhận xét tính chất xung khắc hay không xung khắc của các biến cố chủ yếu dựa vào trực giác.

Các khái niệm trên cho phép chúng ta phát biểu định lý cộng xác suất sau đây.

7.1. Định lý

Xác suất của tổng hai biến cố xung khắc bằng tổng xác suất của các biến cố đó.

Như vậy, nếu A và B là hai biến cố xung khắc với nhau thì

$$P(A + B) = P(A) + P(B) \qquad (1.4)$$

Chứng minh: Ta ký hiệu:

n - Số kết cục đồng khả năng có thể xảy ra khi phép thử được thực hiện.

m_1 - Số kết cục thuận lợi cho biến cố A xảy ra.

m_2 - Số kết cục thuận lợi cho biến cố B xảy ra.

Do A và B xung khắc nhau do đó không thể có các kết cục thuận lợi cho cả A và B cùng đồng thời xảy ra. Vậy số kết cục thuận lợi cho A hoặc cho B xảy ra bằng $m_1 + m_2$. Vì thế:

$$P(A + B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n}$$

Song ta lại có:

$$\frac{m_1}{n} = P(A) \; ; \; \frac{m_2}{n} = P(B)$$

Do đó:

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

Chú ý rằng (1.4) chỉ là điều kiện cần chứ không phải là điều kiện đủ để A và B xung khắc và mặc dù nó được chứng minh bằng định nghĩa cổ điển nhưng định lý đúng cho mọi trường hợp.

Từ định lý trên có thể suy ra một số hệ quả sau đây.

7.2. Hệ quả

Hệ quả 1. Xác suất của tổng các biến cố xung khắc từng đôi $A_1 A_2 ... A_n$ bằng tổng xác suất của các biến cố đó:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \qquad (1.5)$$

Chứng minh: Việc chứng minh được tiến hành theo phương pháp quy nạp toán học. Trong trường hợp có hai biến cố, hệ quả đúng như đã chứng minh trong định lý. Giả sử hệ quả cũng đúng với n – 1 biến cố, tức là:

$$P\left(\sum_{i=1}^{n-1} A_i\right) = \sum_{i=1}^{n-1} P(A_i)$$

Cần chứng minh rằng nó cũng đúng với n biến cố.

Thật vậy nếu ký hiệu $\sum_{i=1}^{n-1} A_i = B$ thì:

$$P\left(\sum_{i=1}^{n} A_i\right) = P(B + A_n) = P(B) + P(A_n)$$

Song theo giả thiết ở trên

$$P(B) = P\left(\sum_{i=1}^{n-1} A_i\right) = \sum_{i=1}^{n-1} P(A_i)$$

do đó $$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n-1} P(A_i) + P(A_n) = \sum_{i=1}^{n} P(A_i)$$

Chú ý rằng hệ quả trên có thể mở rộng cho một tổng vô hạn các biến cố.

Thí dụ 5. Xác suất để một xạ thủ bắn bia trúng điểm 10 là 0,1; trúng điểm 9 là 0,2; trúng điểm 8 là 0,25 và ít hơn 8 điểm là 0,45. Xạ thủ ấy bắn một viên đạn. Tìm xác suất để xạ thủ được ít nhất 9 điểm.

Giải. Gọi A_1 là biến cố "Xạ thủ bắn trúng điểm 10", A_2 là biến cố "Xạ thủ bắn trúng điểm 9", A là biến cố "Xạ thủ được ít nhất 9 điểm". Vậy $A = A_1 + A_2$.

Vì A_1 và A_2 xung khắc nhau do đó theo định lý cộng xác suất ta có:

$$P(A) = P(A_1 + A_2) = P(A_1) + P(A_2) = 0,1 + 0,2 = 0,3$$

Trước khi phát biểu hệ quả tiếp theo ta đưa ra khái niệm về nhóm đầy đủ các biến cố.

Định nghĩa 5. Các biến cố $A_1, A_2, ..., A_n$ được gọi là một nhóm đầy đủ các biến cố nếu trong kết quả của một phép thử sẽ xảy ra một và chỉ một trong các biến cố đó.

Nói cách khác các biến cố nói trện sẽ tạo nên một nhóm đầy đủ các biến cố nếu chúng xung khắc từng đôi với nhau và tổng của chúng là một biến cố chắc chắn.

Thí dụ 6. Gieo một con xúc xắc, gọi A_i ($i = \overline{1,6}$) là biến cố "Xuất hiện mặt i chấm" thì các biến cố $A_1, A_2, A_3, A_4, A_5, A_6$ tạo nên một nhóm đầy đủ các biến cố.

Đối với nhóm đầy đủ các biến cố ta có hệ quả sau.

Hệ quả 2. Nếu các biến cố A_1 A_2... A_n tạo nên một nhóm đầy đủ các biến cố thì tổng xác suất của chúng bằng 1.

$$\sum_{i=1}^{n} P(A_i) = 1 \qquad (1.6)$$

Chứng minh. Vì các biến cố $A_1, A_2, ..., A_n$ tạo nên một nhóm đầy đủ các biến cố do đó việc xảy ra của ít nhất một trong các biến cố đó là một biến cố chắc chắn. Vì vậy:

$$P\left(\sum_{i=1}^{n} A_i\right) = P(U) = 1$$

Mặt khác, các biến cố này xung khắc từng đôi với nhau, do đó theo hệ quả 1 ta có:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

Từ đó

$$\sum_{i=1}^{n} P(A_i) = 1$$

Trường hợp riêng của nhóm đầy đủ các biến cố là các biến cố đối lập.

Định nghĩa 6. Hai biến cố A và $\overline{A}$ gọi là đối lập với nhau nếu chúng tạo nên một nhóm đầy đủ các biến cố.

Thí dụ 7. Bắn một viên đạn vào bia. Gọi A là biến cố "Bắn trúng bia", $\overline{A}$ là biến cố "Bắn trượt bia". A và $\overline{A}$ là hai biến cố đối lập nhau.

Thí dụ 8. Một hòm có a chính phẩm và b phế phẩm. Lấy ngẫu nhiên từ hòm đó n sản phẩm. Gọi A là biến cố "Trong n sản phẩm lấy ra có ít nhất một chính phẩm", $\overline{A}$ là biến cố "Trong n sản phẩm lấy ra không có chính phẩm nào (toàn phế phẩm)" thì A và $\overline{A}$ là các biến cố đối lập nhau.

Đối với các biến cố đối lập ta có hệ quả sau đây.

Hệ quả 3. Tổng xác suất của hai biến cố đối lập nhau bằng 1.

$$P(A) + P(\overline{A}) = 1 \qquad (1.7)$$

Chứng minh. Các biến cố đối lập tạo nên một nhóm đầy đủ các biến cố mà theo hệ quả 2 thì tổng xác suất của các biến cố tạo nên một nhóm đầy đủ các biến cố bằng 1.

Trong thực tế có nhiều trường hợp xác định xác suất của biến cố đối lập $\overline{A}$ đơn giản hơn nhiều so với việc xác định xác suất của biến cố A. Lúc đó người ta thường xác định $P(\overline{A})$ trước, sau đó xác định P(A) theo công thức $P(A) = 1 - P(\overline{A})$.

Thí dụ 9. Xác suất để sản phẩm sản xuất ra là chính phẩm bằng 0,9. Tìm xác suất để sản phẩm sản xuất ra là phế phẩm.

Giải. Gọi A là biến cố "Sản phẩm là phế phẩm" và $\overline{A}$ là biến cố "Sản phẩm là chính phẩm". A và $\overline{A}$ đối lập nhau, do đó:

$$P(A) = 1 - P(\overline{A}) = 1 - 0,9 = 0,1$$

Thí dụ 10. Trong hòm có n sản phẩm, trong đó có m chính phẩm. Lấy ngẫu nhiên k sản phẩm. Tìm xác suất để trong đó có ít nhất một chính phẩm.

Giải.

Gọi A là biến cố "Trong k sản phẩm lấy ra có ít nhất một chính phẩm" thì biến cố đối lập $\overline{A}$ sẽ là "k sản phẩm lấy ra đều là phế phẩm".

Do đó $$P(A) = 1 - P(\overline{A})$$

Ta đi tìm $P(\overline{A})$. Tổng số kết cục duy nhất đồng khả năng là tổ hợp chập k từ n phần tử bằng C_n^k. Số phế phẩm có trong hòm là n – m. Do đó số kết cục thuận lợi cho biến cố $\overline{A}$ xảy ra là số tổ hợp chập k (k phế phẩm) từ n – m phế phẩm, bằng C_{n-m}^k. Vậy xác suất để cả k sản phẩm lấy ra đều là phế phẩm bằng:

$$P(\overline{A}) = \frac{C_{n-m}^k}{C_n^k}$$

Do đó $$P(A) = 1 - P(\overline{A}) = 1 - \frac{C_{n-m}^k}{C_n^k}$$

Thí dụ 11. Trong hòm có 10 chi tiết, trong đó có 2 chi tiết

hỏng. Tìm xác suất để khi lấy ngẫu nhiên ra 6 chi tiết thì có không quá một chi tiết hỏng.

Giải. Gọi A_0 là biến cố "Trong 6 chi tiết lấy ra không có chi tiết nào hỏng". Gọi A_1 là biến cố "Trong 6 chi tiết lấy ra có 1 chi tiết hỏng". Gọi A là biến cố "Trong 6 chi tiết lấy ra có không quá 1 chi tiết hỏng".

Vậy $A = A_0 + A_1$

Vì A_0 và A_1 xung khắc nhau do đó :

$$P(A) = P(A_0 + A_1) = P(A_0) + P(A_1)$$

Dùng định nghĩa cổ điển về xác suất ta tính được :

$$P(A_0) = \frac{C_8^6}{C_{10}^6} = \frac{2}{15}$$

$$P(A_1) = \frac{C_2^1.C_8^5}{C_{10}^6} = \frac{8}{15}$$

Vậy $$P(A) = \frac{2}{15} + \frac{8}{15} = \frac{2}{3}$$

§8. ĐỊNH LÝ NHÂN XÁC SUẤT

Bây giờ ta chuyển sang nghiên cứu trường hợp khi một biến cố có thể xem như tích của các biến cố khác.

Định nghĩa 1. Biến cố C được gọi là tích của hai biến cố A và B nếu C xảy ra khi và chỉ khi cả hai biến cố A và B cùng đồng thời xảy ra. Ký hiệu C = A.B.

Thí dụ 1. Một mạch điện gồm hai bóng đèn mắc song

song. Gọi A là biến cố "Bóng thứ nhất bị cháy khi điện quá tải", B là biến cố "Bóng thứ hai bị cháy khi điện quá tải", C là biến cố "Mạch điện bị ngắt khi điện quá tải". Rõ ràng là biến cố C chỉ xảy ra khi cả hai biến cố A và B cùng đồng thời xảy ra. Vậy C = A.B.

Thí dụ 2. Có hai hộp, mỗi hộp đều đựng một số cầu trắng và cầu đen. Lấy ngẫu nhiên từ mỗi hộp một quả cầu. Gọi A là biến cố "Lấy được cầu trắng ở hộp thứ nhất", B là biến cố "Lấy được cầu trắng ở hộp thứ hai", C là biến cố "Lấy được hai quả cầu trắng". Vậy C = A.B.

Định nghĩa 2. Biến cố A được gọi là tích của n biến cố A_1, A_2, ..., A_n *nếu A xảy ra khi và chỉ khi cả n biến cố nói trên cùng đồng thời xảy ra.*

Ký hiệu

$$A = \prod_{i=1}^{n} A_i$$

Gắn liền với khái niệm về tích các biến cố là khái niệm về sự độc lập và phụ thuộc của các biến cố đó.

Định nghĩa 3. Hai biến cố A và B gọi là độc lập với nhau nếu việc xảy ra hay không xảy ra của biến cố này không làm thay đổi xác suất xảy ra của biến cố kia và ngược lại. Trong trường hợp việc biến cố này xảy ra hay không xảy ra làm cho xác suất xảy ra của biến cố kia thay đổi thì hai biến cố đó gọi là phụ thuộc nhau.

Thí dụ 3. Trong bình có 3 cầu trắng và 2 cầu đen. Lấy ngẫu nhiên 1 quả cầu. Gọi A là biến cố "Lấy được cầu trắng". Hiển nhiên là $P(A) = \frac{3}{5}$. Quả cầu được bỏ trở lại bình và tiếp tục lấy 1 quả cầu. Gọi B là biến cố "Lần thứ hai cũng được

cầu trắng". Cũng như trước $P(B) = \frac{3}{5}$ và không phụ thuộc gì vào kết quả lấy của lần trước (biến cố A). Cũng như vậy xác suất lấy được cầu trắng lần thứ nhất (biến cố A) cũng không phụ thuộc gì vào kết quả lấy của lần thứ hai (biến cố B). Vậy hai biến cố A và B độc lập với nhau.

Thí dụ 4. Nếu ở thí dụ trên lần lượt lấy ra hai quả cầu theo phương thức không hoàn lại và gọi A là biến cố "Lần thứ nhất lấy được cầu trắng" thì $P(A) = \frac{3}{5}$. Song biến cố "Lần thứ hai lấy được cầu trắng" (biến cố B) sẽ phụ thuộc vào kết quả lấy của lần thứ nhất. Nếu lần thứ nhất lấy được cầu trắng (biến cố A xảy ra) thì $P(B) = \frac{1}{2}$, còn nếu lần thứ nhất lấy được cầu đen (biến cố A không xảy ra) thì $P(B) = \frac{3}{4}$. Vậy A và B phụ thuộc nhau.

Ta chú ý rằng tính chất độc lập của các biến cố có tính tương hỗ theo nghĩa là nếu A và B độc lập với nhau thì A và $\overline{B}$, $\overline{A}$ và B, $\overline{A}$ và $\overline{B}$ cũng độc lập với nhau.

Trong thực tế việc nhận xét tính độc lập hay phụ thuộc của các biến cố chủ yếu dựa vào trực giác.

Việc mở rộng khái niệm độc lập cho nhiều biến cố sẽ dẫn đến hai khái niệm khác nhau là sự độc lập từng đôi và sự độc lập toàn phần.

Định nghĩa 4. Các biến cố $A_1, A_2, ..., A_n$ gọi là độc lập từng đôi với nhau nếu mỗi cặp hai trong n biến cố đó độc lập với nhau.

Chẳng hạn ba biến cố A_1, A_2, A_3 sẽ độc lập từng đôi với

nhau nếu A_1 độc lập với A_2, A_1 độc lập với A_3 và A_2 độc lập với A_3.

Thí dụ 5. Tung một đồng xu 3 lần, gọi A_i ($i = \overline{1,3}$) là biến cố "Được mặt sấp ở lần tung thứ i". Rõ ràng là mỗi cặp hai trong ba biến cố đó độc lập với nhau.

Vậy A_1, A_2, A_3 độc lập từng đôi với nhau.

Định nghĩa 5. Các biến cố A_1, A_2, ..., A_n gọi là độc lập toàn phần với nhau nếu mỗi biến cố độc lập với một tổ hợp bất kỳ của các biến cố còn lại.

Chẳng hạn ba biến cố A_1, A_2, A_3 sẽ độc lập toàn phần với nhau nếu A_1 độc lập với A_2, A_1 độc lập với A_3, A_2 độc lập với A_3, A_1 độc lập với tích A_2A_3, A_2 độc lập với tích A_1A_3, A_3 độc lập với tích A_1A_2.

Chú ý rằng nếu các biến cố độc lập từng đôi với nhau thì từ đó chưa thể suy ra rằng chúng độc lập toàn phần với nhau. Xét theo nghĩa đó thì điều kiện độc lập toàn phần mạnh hơn điều kiện độc lập từng đôi.

Ta sẽ minh họa điều đó qua *thí dụ* sau đây. Giả sử trong bình có 4 quả cầu, 1 quả màu đỏ, 1 quả màu xanh, 1 quả màu vàng, 1 quả sơn cả 3 màu đó. Nếu gọi A là biến cố lấy ngẫu nhiên từ bình đó được cầu có màu đỏ thì $P(A) = \frac{1}{2}$. Tương tự như vậy, xác suất để lấy được quả cầu có màu xanh là $P(B) = \frac{1}{2}$, quả cầu có màu vàng là $P(C) = \frac{1}{2}$.

Bây giờ ta giả sử là quả cầu được lấy ra có màu xanh, tức là biến cố B đã xảy ra. Vậy lúc đó xác suất của biến cố A có

thay đổi không? Trong hai quả cầu có màu xanh có một quả được sơn cả màu đỏ, do đó vẫn như trước đây $P(A) = \frac{1}{2}$. Vậy A và B độc lập với nhau.

Tương tự có thể chứng tỏ rằng A và C, B và C cũng độc lập với nhau, do đó A, B và C độc lập từng đôi với nhau.

Song các biến cố trên có độc lập toàn phần với nhau không? Có thể chứng tỏ rằng không. Thật vậy, giả sử quả cầu lấy ra đã có hai màu xanh và vàng, tức là các biến cố B và C đã xảy ra. Vậy xác suất P(A) để quả cầu đó có màu đỏ là bao nhiêu? Vì trong các quả cầu chỉ có một quả được sơn cả ba màu do đó chắc chắn là quả cầu đó cũng có màu đỏ. Như vậy với điều kiện B và C đã xảy ra thì xác suất P(A) = 1. Do đó các biến cố A, B, C chỉ độc lập từng đôi chứ không độc lập toàn phần với nhau.

Khi giải nhiều bài toán thực tế nhiều lúc ta phải biểu diễn các biến cố phức hợp dưới dạng kết hợp của các biến cố đơn giản hơn bằng cách sử dụng phép nhân các biến cố.

Chẳng hạn một máy sản xuất ra ba sản phẩm. Ta xét các biến cố giản đơn sau:

A_1 - Sản phẩm thứ nhất là chính phẩm

$\overline{A}_1$ - Sản phẩm thứ nhất là phế phẩm

A_2 - Sản phẩm thứ hai là chính phẩm

$\overline{A}_2$ - Sản phẩm thứ hai là phế phẩm

A_3 - Sản phẩm thứ ba là chính phẩm

$\overline{A}_3$ - Sản phẩm thứ ba là phế phẩm

Lúc đó nếu gọi B là biến cố trong ba sản phẩm sản xuất

ra có đúng một chính phẩm thì B có thể biểu diễn qua các biến cố giản đơn như sau:

$$B = A_1\overline{A}_2\overline{A}_3 + \overline{A}_1A_2\overline{A}_3 + \overline{A}_1\overline{A}_2A_3$$

Gọi C là biến cố trong ba sản phẩm đó có ít nhất hai chính phẩm thì C được biểu diễn như sau:

$$C = A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3 + A_1A_2A_3$$

Những thủ thuật như vậy để biểu diễn các biến cố phức hợp thông qua các biến cố giản đơn thường được sử dụng trong lý thuyết xác suất.

Với những khái niệm đó ta có thể phát biểu định lý sau đây:

8.1. Định lý

Xác suất của tích hai biến cố độc lập bằng tích các xác suất thành phần:

$$P(A.B) = P(A).P(B) \qquad (1.8)$$

Chứng minh. Giả sử n_1 và n_2 tương ứng là số kết cục đồng khả năng cho biến cố A và B xảy ra; m_1 và m_2 tương ứng là số kết cục thuận lợi cho A và B xảy ra. Do A và B độc lập nên số kết cục đồng khả năng cho tích AB xảy ra sẽ là $n_1.n_2$ và số kết cục thuận lợi cho tích xảy ra là $m_1.m_2$. Vì vậy:

$$P(A.B) = \frac{m_1.m_2}{n_1.n_2} = \frac{m_1}{n_1}.\frac{m_2}{n_2} = P(A).P(B)$$

Chú ý rằng (1.8) vừa là điều kiện cần, vừa là điều kiện đủ để A và B độc lập.

Từ định lý trên có thể suy ra một số hệ quả sau:

Hệ quả 1. Nếu A và B độc lập thì:

$$P(A) = \frac{P(A.B)}{P(B)} \text{ và } P(B) = \frac{P(A.B)}{P(A)} \qquad (1.9)$$

khi $P(B) > 0$ và $P(A) > 0$.

Hệ quả 2. Xác suất của tích n biến cố độc lập toàn phần bằng tích các xác suất thành phần.

$$P\left(\prod_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i) \qquad (1.10)$$

Bạn đọc có thể tự chứng minh hệ quả trên bằng phương pháp quy nạp toán học.

Thí dụ 6. Có hai hộp đựng chi tiết. Hộp thứ nhất đựng 10 cái ốc, trong đó có 6 cái tốt. Hộp thứ hai đựng 15 cái vít, trong đó có 9 cái tốt. Lấy ngẫu nhiên từ mỗi hộp một chi tiết. Tìm xác suất để lấy được một bộ ốc vít tốt.

Giải Gọi A_1 là biến cố "Lấy được cái ốc tốt ở hộp thứ nhất", A_2 là biến cố "Lấy được cái vít tốt ở hộp thứ hai". Gọi A là biến cố "Lấy được một bộ ốc vít tốt".

Vậy: $A = A_1.A_2$

Vì các biến cố A_1 và A_2 độc lập với nhau, do đó:

$$P(A) = P(A_1.A_2) = P(A_1).P(A_2) = \frac{6}{10} \cdot \frac{9}{15} = \frac{9}{25}$$

Để xem xét trường hợp các biến cố phụ thuộc trước hết ta xét khái niệm xác suất có điều kiện.

Định nghĩa 6. Xác suất của biến cố A được tính với điều kiện biến cố B đã xảy ra gọi là xác suất có điều kiện của A và ký hiệu là P(A/B).

Thí dụ 7. Trong bình có 5 cầu trắng và 3 cầu đen. Lấy ngẫu nhiên lần lượt hai quả cầu. Tìm xác suất để lần thứ hai lấy được cầu trắng (biến cố B) nếu biết rằng lần thứ nhất đã lấy được cầu trắng (biến cố A).

Giải. Sau khi lấy lần thứ nhất (biến cố A đã xảy ra) thì trong bình chỉ còn lại 7 quả cầu, trong đó có 4 quả cầu trắng. Do đó xác suất có điều kiện của B bằng:

$$P(B/A) = \frac{4}{7}$$

Bây giờ ta có thể phát biểu định lý nhân xác suất đối với một tích các biến cố phụ thuộc.

8.2. Định lý

Xác suất của tích hai biến cố phụ thuộc A và B bằng tích xác suất của một trong hai biến cố đó với xác suất có điều kiện của biến cố còn lại.

$$P(A.B) = P(A).P(B/A) = P(B).P(A/B) \qquad (1.11)$$

Chứng minh. Giả sử n là số kết cục đồng khả năng có thể xảy ra trong phép thử, m_1 là số kết cục thuận lợi cho biến cố A xảy ra, m_2 là số kết cục thuận lợi cho biến cố B xảy ra. Vì ta không giả thiết A và B xung khắc do đó nói chung sẽ có k kết cục thuận lợi cho cả A và B cùng đồng thời xảy ra. Lúc đó:

$$P(A.B) = \frac{k}{n};\ P(A) = \frac{m_1}{n}$$

Ta đi tìm P(B/A). Với điều kiện biến cố A đã xảy ra thì số kết cục duy nhất đồng khả năng của phép thử đối với biến cố B là m_1 trong đó có k kết cục thuận lợi cho B xảy ra. Do đó:

$$P(B/A) = \frac{k}{m_1}$$

Như vậy: $P(A.B) = \frac{k}{n} = \frac{m_1}{n} \cdot \frac{k}{m_1} = P(A).P(B/A)$

Hiển nhiên là khi áp dụng định lý nhân xác suất thì không nhất thiết phải phân biệt xem biến cố nào trong hai biến cố A và B là thứ nhất và thứ hai. Do đó có thể chứng minh được dạng thức thứ hai là P(A.B) = P(B).P(A/B).

Từ định lý trên có thể suy ra các hệ quả sau đây:

Hệ quả 1. Nếu P(B) > 0 thì xác suất của biến cố A với điều kiện biến cố B đã xảy ra được tính bằng công thức:

$$P(A/B) = \frac{P(AB)}{P(B)} \qquad (1.12)$$

Còn nếu P(B) = 0 thì xác suất trên không xác định. Tương tự nếu P(A) > 0 thì ta có:

$$P(B/A) = \frac{P(AB)}{P(A)} \qquad (1.13)$$

Các công thức trên được suy ra trực tiếp từ định lý vừa xét.

Hệ quả 2. Xác suất của tích n biến cố phụ thuộc bằng tích xác suất của n biến cố đó, trong đó xác suất của mỗi biến cố tiếp sau đều được tính với điều kiện tất cả các biến cố xét trước đó đã xảy ra

$$P(A_1A_2 \ldots A_n) = P(A_1)P(A_2/A_1) \ldots P(A_n/A_1 \ldots A_{n-1}) \qquad (1.14)$$

Bạn đọc có thể tự chứng minh hệ quả này bằng phương pháp quy nạp toán học.

Hệ quả 3. Nếu A và B độc lập thì:

$$P(A/B) = P(A) \text{ và } P(B/A) = P(B) \tag{1.15}$$

Thật vậy từ định lý vừa xét ta có:

$$P(A/B) = \frac{P(AB)}{P(B)} = \frac{P(A).P(B)}{P(B)} = P(A)$$

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{P(A).P(B)}{P(A)} = P(B)$$

Thí dụ 8. **Trong hòm có 7 chính phẩm và 3 phế phẩm. Lấy ngẫu nhiên lần lượt hai sản phẩm. Tìm xác suất để cả hai sản phẩm lấy ra đều là chính phẩm.**

Giải. Gọi A_1 là biến cố "Sản phẩm thứ nhất lấy ra là chính phẩm", A_2 là biến cố "Sản phẩm thứ hai lấy ra là chính phẩm", A là biến cố "Cả hai sản phẩm lấy ra đều là chính phẩm".

Vậy: $A = A_1.A_2$

Theo định lý nhân xác suất đối với hai biến cố phụ thuộc ta có:

$$P(A) = P(A_1.A_2) = P(A_1).P(A_2/A_1) = \frac{7}{10}.\frac{6}{9} = \frac{7}{15}$$

Trong thực tế ta thường gặp các bài toán trong đó phải áp dụng cùng một lúc cả định lý cộng và định lý nhân xác suất, tức là các biến cố mà ta phải xác định xác suất được biểu diễn dưới dạng biến cố đối lập hoặc dưới dạng tổng của một số biến cố xung khắc từng đôi với nhau và mỗi biến cố như vậy lại là tích của một số biến cố khác.

Thí dụ 9. **Một thiết bị gồm 3 bộ phận. Trong khoảng thời gian T, việc các bộ phận đó bị hỏng là độc lập với nhau và với**

các xác suất tương ứng là 0,1; 0,2; 0,3. Cả thiết bị sẽ bị hỏng nếu có ít nhất một bộ phận hỏng. Tìm xác suất hoạt động tốt của thiết bị đó.

Giải. Gọi A_i ($i = \overline{1, 3}$) là biến cố "Bộ phận thứ i của thiết bị hoạt động tốt trong khoảng thời gian T". Gọi A là biến cố "Thiết bị hoạt động tốt trong khoảng thời gian T".

Vậy $$A = A_1.A_2.A_3$$

Vì A_1, A_2, A_3 độc lập toàn phần với nhau do đó:

$$P(A) = P(A_1)P(A_2)P(A_3)$$

Các biến cố "Bộ phận thứ nhất hoạt động tốt" và "Bộ phận thứ nhất bị hỏng" là đối lập với nhau, do đó:

$$P(A_1) = 1 - 0,1 = 0,9$$

Tương tự $$P(A_2) = 1 - 0,2 = 0,8$$

$$P(A_3) = 1 - 0,3 = 0,7$$

Vì vậy $$P(A) = 0,9.0,8.0,7 = 0,504$$

Thí dụ 10. Một xí nghiệp có 3 ô tô hoạt động độc lập. Xác suất để trong một ngày các ô tô bị hỏng tương ứng là 0,1; 0,2; 0,15. Tìm xác suất để trong một ngày có đúng 1 ô tô bị hỏng.

Giải. Gọi A_i là biến cố "Ô tô thứ i bị hỏng trong ngày" ($i = \overline{1, 3}$); A là biến cố "Trong ngày có đúng 1 ô tô bị hỏng".

Vậy $$A = A_1\overline{A}_2\overline{A}_3 + \overline{A}_1A_2\overline{A}_3 + \overline{A}_1\overline{A}_2A_3$$

Vì các nhóm biến cố $A_1A_2A_3$, $A_1A_2A_3$ và $A_1A_2A_3$ là xung khắc từng đôi và trong mỗi nhóm các biến cố lại độc lập toàn phần với nhau, do đó:

$$P(A) = P(A_1)P(\overline{A}_2)P(\overline{A}_3) + P(\overline{A}_1)P(A_2)P(\overline{A}_3) + P(\overline{A}_1)P(\overline{A}_2)P(A_3)$$

Vì $P(A_1) = 0,1; P(A_2) = 0,2; P(A_3) = 0,15$

do đó: $P(\overline{A}_1) = 0,9; P(\overline{A}_2) = 0,8; P(\overline{A}_3) = 0,85$

và ta có: $P(A) = 0,1.0,8.0,85 + 0,9.0,2.0,85 + 0,9.0,8.0,15 = 0,329$

Thí dụ 11. Với các điều kiện của thí dụ trước, tìm xác suất để trong một ngày có ít nhất một ô tô bị hỏng.

Giải. Gọi B là biến cố "Trong 1 ngày có ít nhất 1 ô tô bị hỏng". Áp dụng thủ thuật như trên có thể biểu diễn biến cố B như sau:

$$B = A_1\overline{A}_2\overline{A}_3 + \overline{A}_1A_2\overline{A}_3 + \overline{A}_1\overline{A}_2A_3 + A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3 + A_1A_2A_3$$

Xác suất của biến cố B có thể tìm bằng cách áp dụng định lý cộng và định lý nhân như đã làm ở thí dụ trước. Tuy nhiên cách giải như vậy rất dài dòng và phức tạp. Ở đây nên chuyển từ biến cố B sang biến cố đối lập với nó. Hiển nhiên $\overline{B}$ sẽ là biến cố cả 3 ô tô cùng hoạt động tốt:

$$\overline{B} = \overline{A}_1\overline{A}_2\overline{A}_3$$

Theo định lý nhân xác suất:

$$P(\overline{B}) = P(\overline{A}_1).P(\overline{A}_2).P(\overline{A}_3) = 0,9.0,8.0,85 = 0,612$$

Do đó: $P(B) = 1 - P(\overline{B}) = 1 - 0,612 = 0,388$

Thí dụ 12. Tung một đồng xu 6 lần. Tìm xác suất để số lần được mặt sấp nhiều hơn số lần được mặt ngửa.

Giải. Để tìm xác suất của biến cố được số mặt sấp nhiều hơn số mặt ngửa (biến cố A) ta có thể phân nó ra thành các trường hợp sau đây:

A_1 - Được 6 mặt sấp, không có lần nào được mặt ngửa;

A_2 - Được 5 mặt sấp, 1 mặt ngửa...

Song bài toán này có thể giải bằng cách đơn giản hơn. Ta liệt kê các trường hợp có thể xảy ra như sau:

A - Số mặt sấp nhiều hơn số mặt ngửa

B - Số mặt ngửa nhiều hơn số mặt sấp

C - Số mặt sấp và ngửa bằng nhau.

Các biến cố A, B và C tạo nên một nhóm đầy đủ các biến cố. Vì vậy:

$$P(A) + P(B) + P(C) = 1$$

Vì mặt sấp và mặt ngửa đối xứng nhau, do đó $P(A) = P(B)$

Từ đó:

$$2P(A) + P(C) = 1$$

$$P(A) = \frac{1 - P(C)}{2}$$

Ta đi tìm P(C) là xác suất để trong 6 lần tung có 3 lần được sấp, 3 lần được ngửa. Xác suất của mỗi phương thức xảy ra của biến cố C (chẳng hạn theo trình tự SSSNNN) cùng bằng $\left(\frac{1}{2}\right)^6$. Số những phương thức như vậy là $C_6^3 = 20$. Vì vậy:

$$P(C) = \frac{20}{64} = \frac{5}{16}$$

Từ đó:

$$P(A) = \frac{1 - \frac{5}{16}}{2} = \frac{11}{32}$$

§9. CÁC HỆ QUẢ CỦA ĐỊNH LÝ CỘNG VÀ ĐỊNH LÝ NHÂN XÁC SUẤT

9.1. Định lý

Xác suất của tổng hai biến cố không xung khắc bằng tổng xác suất các biến cố đó trừ đi xác suất của tích các biến cố đó.

$$P(A + B) = P(A) + P(B) - P(AB) \qquad (1.16)$$

Chứng minh. Giả sử n là số kết cục đồng khả năng có thể xảy ra trong phép thử, m_1 là số kết cục thuận lợi cho A xảy ra, m_2 là số kết cục thuận lợi cho B xảy ra. Vì A và B không xung khắc nhau, do đó sẽ có k kết cục nào đó thuận lợi cho tích AB xảy ra. Lúc đó số kết cục thuận lợi cho ít nhất một trong hai biến cố A và B xảy ra bằng $m_1 + m_2 - k$. Như vậy:

$$P(A + B) = \frac{m_1 + m_2 - k}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{k}{n}$$

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

Nếu các biến cố A và B độc lập thì công thức có dạng:

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

Còn đối với các biến cố phụ thuộc thì công thức có dạng:

$$P(A + B) = P(A) + P(B) - P(A)P(B/A)$$

Nếu A và B xung khắc nhau thì lúc đó biến cố tích AB sẽ là biến cố không thể có, do đó P(AB) = 0. Ta lại thu được công thức cộng xác suất đã xét ở phần trước.

Khi mở rộng định lý trên cho trường hợp tổng của n biến cố không xung khắc, ta thu được hệ quả sau đây.

9.2. Hệ quả

Hệ quả 1. Xác suất của tổng n biến cố không xung khắc được xác định bằng công thức:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) -$$

$$\ldots + (-1)^{n-1} P(A_1 A_2, \ldots A_n) \qquad (1.17)$$

Hệ quả trên có thể chứng minh bằng phương pháp qui nạp toán học. Cũng bằng phương pháp quy nạp toán học có thể chứng minh được hệ quả sau đây, cho phép biểu diễn xác suất của tích n biến cố thông qua xác suất của tổng các biến cố đó.

Hệ quả 2. Xác suất của tích n biến cố được xác định bằng công thức:

$$P\left(\prod_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i + A_j) +$$

$$+ \sum_{i<j<k} P(A_i + A_j + A_k) - \ldots$$

$$+ (-1)^{n-1} P(A_1 + A_2 + \ldots + A_n) \qquad (1.18)$$

Thực tế các hệ quả đã xét ở trên được sử dụng để biến đổi các biểu thức chứa xác suất của tổng và tích các biến cố. Tùy thuộc vào đặc điểm của từng bài toán, trong một số trường hợp chỉ nên dùng tổng các biến cố, trong một vài trường hợp khác lại chỉ nên dùng tích các biến cố. Các công thức trên được sử dụng để thực hiện các phép biến đổi qua lại đó.

Tuy nhiên trong thực tế giải các bài toán của lý thuyết xác suất việc vận dụng các hệ quả trên là khá phức tạp. Trong một số trường hợp có thể chuyển qua biến cố đối lập để việc giải quyết trở nên đơn giản hơn. Ta xét một trường hợp thường gặp trong thực tế sau đây.

9.3. Định lý

Xác suất của tổng n biến cố không xung khắc và độc lập toàn phần với nhau bằng một trừ đi tích xác suất của các biến cố đối lập với các biến cố đó.

$$P\left(\sum_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} P(\overline{A}_i) \qquad (1.19)$$

Chứng minh. Nếu ký hiệu $A = \sum_{i=1}^{n} A_i$ thì lúc đó $\overline{A} = \prod_{i=1}^{n} \overline{A}_i$, do đó:

$$P\left(\sum_{i=1}^{n} A_i\right) = P(A) = 1 - P(\overline{A}) = 1 - P\left(\prod_{i=1}^{n} \overline{A}_i\right)$$

Theo giả thiết các biến cố A_i độc lập toàn phần, do đó các biến cố $\overline{A}_i$ cũng độc lập toàn phần với nhau.

Vì thế:

$$P\left(\sum_{i=1}^{n} \overline{A}_i\right) = \prod_{i=1}^{n} P(\overline{A}_i)$$

Từ đó:
$$P\left(\sum_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} P(\overline{A}_i)$$

Trong trường hợp riêng nếu $P(A_1) = P(A_2) = \ldots = P(A_n) = p$ thì công thức trên có dạng sau đây:

$$P\left(\sum_{i=1}^{n} A_i\right) = 1 - (1-p)^n \qquad (1.20)$$

Thí dụ 1. Hai máy bay ném bom một mục tiêu, mỗi máy bay ném 1 quả với xác suất trúng mục tiêu tương ứng là 0,7 và 0,8. Tìm xác suất để mục tiêu bị trúng bom.

Giải. Gọi A_1 và A_2 tương ứng là biến cố quả bom thứ nhất và thứ hai trúng mục tiêu. Gọi A là biến cố mục tiêu bị trúng bom. Vậy $A = A_1 + A_2$.

Vì A_1 và A_2 không xung khắc nhau, do đó:

$$P(A) = P(A_1) + P(A_2) - P(A_1A_2)$$

A_1 và A_2 lại độc lập với nhau, do đó:

$$P(A) = P(A_1) + P(A_2) - P(A_1)P(A_2) =$$
$$= 0,7 + 0,8 - 0,7.0,8 = 0,94$$

Bài toán trên cũng có thể giải bằng cách sử dụng định lý đã xét ở trên. Do A_1 và A_2 là không xung khắc và độc lập nên

$$P(A) = 1 - P(\overline{A}_1)P(\overline{A}_2) = 1 - 0,3.0,2 = 0,94$$

Ta cũng thu được kết quả tương tự.

Thí dụ 2. Phải tung một con xúc xắc tối thiểu bao nhiêu lần để với xác suất không nhỏ hơn 0,5 có thể hy vọng rằng trong đó có ít nhất một lần được mặt sáu chấm.

Giải. Giả sử ta tung con xúc xắc n lần.

Gọi A_i là biến cố "Tung lần thứ i được mặt 6 chấm", $i = \overline{1,n}$. Gọi A là biến cố "Trong n lần tung có ít nhất một lần được mặt 6 chấm". Vậy $A = \sum_{i=1}^{n} A_i$.

Các biến cố A_i là không xung khắc và độc lập toàn phần với nhau, do đó:

$$P(A) = 1 - \prod_{i=1}^{n} P(\overline{A}_i)$$

Vì $P(A_1) = P(A_2) = ... = P(A_n) = \frac{1}{6}$ (xác suất để mỗi lần tung được mặt sáu chấm) do đó:

$$P(\overline{A}_1) = P(\overline{A}_2) = ... = P(\overline{A}_n) = \frac{5}{6}$$

nên ta có:

$$P(A) = 1 - \left(\frac{5}{6}\right)^n$$

Theo giả thiết xác suất của biến cố A không nhỏ hơn 0,5 do đó ta thu được bất phương trình sau:

$$1 - \left(\frac{5}{6}\right)^n \geq 0,5$$

Từ đó:
$$\left(\frac{5}{6}\right)^n \leq 0,5$$

Lấy lôgarit của cả hai vế ta có:

$$n \lg\left(\frac{5}{6}\right) \leq \lg 0,5$$

Chú ý rằng $\lg\left(\frac{5}{6}\right) \leq 0$, ta có:

$$n \geq \frac{\lg 0,5}{\lg\left(\frac{5}{6}\right)} = 3,7$$

Như vậy $n \geq 4$, tức là phải tung ít nhất 4 lần.

Thí dụ 3. Xác suất để động cơ thứ nhất của máy bay bị trúng đạn là 0,2; để động cơ thứ hai của máy bay bị trúng đạn là 0,3 còn xác suất trúng đạn của phi công là 0,1. Tìm xác suất để máy bay rơi, biết rằng máy bay rơi khi hoặc cả hai động cơ bị trúng đạn, hoặc phi công bị trúng đạn.

Giải. Gọi A_1 và A_2 tương ứng là biến cố "Động cơ thứ nhất và thứ hai của máy bay bị trúng đạn", gọi A_3 là biến cố "Phi công bị trúng đạn". Gọi A là biến cố "Máy bay rơi". Lúc đó:

$$A = A_1A_2 + A_3$$

Vì các biến cố A_1A_2 và A_3 không xung khắc với nhau, do đó:

$$P(A) = P(A_1A_2) + P(A_3) - P(A_1A_2A_3)$$

Mặt khác các biến cố $A_1A_2A_3$ độc lập toàn phần với nhau, do đó:

$$P(A) = P(A_1)P(A_2) + P(A_3) - P(A_1)P(A_2)P(A_3) =$$

$$= 0,2.0,3 + 0,1 - 0,2.0,3.0,1 = 0,154$$

Thí dụ 4. Một người viết n lá thư, bỏ ngẫu nhiên vào n phong bì có đề sẵn địa chỉ. Tìm xác suất để có ít nhất một lá thư bỏ đúng địa chỉ.

Giải. Gọi A_i là biến cố "Lá thư thứ i bỏ đúng địa chỉ", $i = \overline{1,n}$, A là biến cố "Trong n lá thư đó có ít nhất một lá thư bỏ đúng địa chỉ", $A = \sum_{i=1}^{n} A_i$.

Vì các biến cố A_i không xung khắc với nhau, do đó:

$$P(A) = \sum_i P(A_i) - \sum_{i<j} P(A_iA_j) + \sum_{i<j<k} P(A_iA_jA_k) - \ldots +$$

$$+ \ldots + (-1)^{n-1} P(A_1A_2 \ldots A_n)$$

Ta đi tìm các xác suất thành phần.

Vì trong n lá thư chỉ có một lá thư của phong bì thứ i, do đó xác suất để lá thư thứ i bỏ đúng địa chỉ bằng:

$$P(A_i) = \frac{1}{n} \quad (\forall i)$$

Tất cả có n lá thư, do đó:

$$\sum_i P(A_i) = n.\frac{1}{n} = 1$$

Ta đi tìm $P(A_iA_j)$ $\forall ij$. Vì các biến cố A_i và A_j phụ thuộc nhau do đó:

$$P(A_iA_j) = P(A_i)P(A_j/A_i)$$

Như ở trên đã xác định $P(A_i) = \frac{1}{n}$, $\forall i$. Với điều kiện biến cố A_i đã xảy ra thì xác suất bỏ đúng địa chỉ của lá thứ j sẽ bằng $\frac{1}{n-1}$, do đó:

$$P(A_iA_j) = \frac{1}{n}.\frac{1}{n-1}$$

Số phương thức để thành lập những nhóm hai biến cố như vậy từ n biến cố là C_n^2, do đó:

$$\sum_i \sum_j P(A_iA_j) = C_n^2 \frac{1}{n}.\frac{1}{n-1} = \frac{1}{2!}$$

Tương tự có thể tìm được:

$$\sum_{i<j<k} P(A_i A_j A_k) = \frac{1}{3!}$$

và $$P(A_1 A_2 \ldots A_n) = \frac{1}{n!}$$

Do đó:

$$P(A) = 1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{n-1} \frac{1}{n!} = \sum_{x=1}^{n} (-1)^{x-1} \frac{1}{x!}$$

Với n khá lớn $P(A) \approx 1 - e^{-1}$.

9.4. Công thức Bernoulli

Trong nhiều bài toán thực tế ta thường gặp trường hợp cùng một phép thử được lặp lại nhiều lần. Trong kết quả của mỗi phép thử có thể xảy ra hoặc không xảy ra một biến cố A nào đó và ta không quan tâm đến kết quả của từng phép thử mà quan tâm đến tổng số lần xảy ra của biến cố A trong cả dãy phép thử đó. Chẳng hạn nếu tiến hành sản xuất hàng loạt một loại chi tiết nào đó thì ta thường quan tâm đến tổng số chi tiết đạt tiêu chuẩn của cả quá trình sản xuất. Trong những bài toán như vậy cần phải biết cách xác định xác suất để biến cố A xảy ra một số lần nhất định trong kết quả của cả một dãy phép thử. Bài toán này sẽ được giải quyết khá dễ dàng nếu các phép thử là độc lập với nhau.

Các phép thử được gọi là độc lập với nhau nếu xác suất để xảy ra một biến cố nào đó trong từng phép thử sẽ không phụ thuộc vào việc biến cố đó có xảy ra ở các phép thử khác hay không. Chẳng hạn tung nhiều lần một đồng xu sẽ tạo nên các phép thử độc lập, lấy nhiều lần sản phẩm từ một lô sản phẩm theo phương thức có hoàn lại cũng tạo nên các phép thử độc lập v.v... Giả sử ta tiến hành n phép thử độc

lập. Trong mỗi phép thử chỉ có hai trường hợp: Hoặc biến cố A xảy ra, hoặc biến cố A không xảy ra, xác suất xảy ra của biến cố A trong mỗi phép thử đều bằng p và xác suất không xảy ra của biến cố A trong mỗi phép thử đều bằng $q = 1 - p$. Những bài toán thỏa mãn cả ba điều giả thiết trên được gọi là tuân theo lược đồ Bernoulli. Lúc đó xác suất để trong n phép thử độc lập nói trên, biến cố A xuất hiện đúng x lần, ký hiệu là $P_n(x)$, được tính bằng công thức Bernoulli sau đây:

$$P_n(x) = C_n^x p^x q^{n-x} \qquad (1.21)$$

trong đó $x = 0, 1, 2, ..., n$.

Chứng minh. Gọi A_i là biến cố "Xảy ra biến cố A trong phép thử thứ i", $i = \overline{1,n}$. Như vậy $\overline{A}_i$ là biến cố "Không xảy ra biến cố A trong phép thử thứ i", $i = \overline{1,n}$.

Gọi B là biến cố "Trong n phép thử biến cố A xảy ra đúng x lần". Biến cố này có thể xảy ra theo nhiều phương thức khác nhau, trong đó việc xảy ra của A đúng x lần và không xảy ra đúng n – x lần có thể diễn ra theo những trình tự khác nhau. Do đó:

$$B = A_1A_2...A_x\overline{A}_{x+1}...\overline{A}_n + ... + \overline{A}_1\overline{A}_2...\overline{A}_{n-x}A_{n-x+1}...A_n$$

Tổng số các tích biến cố như vậy trong biểu thức trên là C_n^x tức là số cách chọn ra x phép thử trong đó biến cố A xảy ra từ n phép thử. Đối với mỗi tích ta thấy biến cố A xảy ra x lần, còn $\overline{A}$ xảy ra n – x lần, do đó xác suất của mỗi biến cố tích đều bằng p^xq^{n-x}. Vì các tích biến cố đó là xung khắc từng đôi với nhau, do đó:

$$P_n(x) = P(B) = \underbrace{p^xq^{n-x} + ... + p^xq^{n-x}}_{C_n^x \text{ lần}} = C_n^x p^x q^{n-x}$$

Việc sử dụng công thức Bernoulli cho phép trong nhiều trường hợp giải bài toán ngắn gọn hơn nhiều so với việc sử dụng định lý cộng và định lý nhân xác suất.

Thí dụ 5. Trong phân xưởng có 5 máy hoạt động, xác suất để trong ca mỗi máy bị hỏng đều bằng 0,1. Tìm xác suất để trong ca đó đúng 2 máy hỏng.

Giải. Nếu coi sự hoạt động của mỗi máy là một phép thử, ta có 5 phép thử độc lập. Trong mỗi phép thử chỉ có hai trường hợp: Hoặc máy hỏng, hoặc máy chạy tốt. Xác suất hỏng của mỗi máy đều bằng 0,1. Như vậy, bài toán thỏa mãn lược đồ Bernoulli. Vì thế xác suất để trong ca có đúng 2 máy hỏng được tính bằng công thức Bernoulli như sau:

$$P_5(2) = C_5^2(0,1)^2(0,9)^3 = 0,0729$$

Hiển nhiên bài toán trên cũng có thể giải bằng cách sử dụng định lý cộng và định lý nhân xác suất, song sẽ dài dòng hơn nhiều.

Thí dụ 6. Bắn 6 viên đạn vào bia. Xác suất trúng đích của mỗi viên là 0,7. Tìm xác suất để có 3 viên trúng bia.

Giải. Bài toán thỏa mãn lược đồ Bernoulli. Do đó xác suất để trong 6 viên có 3 viên trúng bia bằng:

$$P_6(3) = C_6^3(0,7)^3(0,3)^3 = 0,18522$$

9.5. Công thức xác suất đầy đủ

Giả sử biến cố A có thể xảy ra đồng thời với một trong các biến cố $H_1, H_2, ..., H_n$. Nhóm $H_1, H_2, ..., H_n$ là nhóm đầy đủ các biến cố. Lúc đó xác suất của biến cố A được tính bằng công thức sau đây:

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i) \qquad (1.22)$$

Các biến cố H_1, H_2, ..., H_n thường được gọi là các giả thuyết.

Chứng minh. Vì các biến cố H_1, H_2, ..., H_n tạo nên một nhóm đầy đủ các biến cố nên biến cố A chỉ có thể xảy ra đồng thời với một trong các biến cố đó:

$$A = H_1A + H_2A + ... + H_nA$$

Vì các biến cố H_1, H_2, ..., H_n xung khắc từng đôi, do đó các tích biến cố H_1A, H_2A, ..., H_nA cũng xung khắc từng đôi. Do vậy:

$$P(A) = P(H_1A) + P(H_2A) + ... + P(H_nA)$$

Áp dụng định lý nhân xác suất đối với các tích H_iA ta có:

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i)$$

Thí dụ 7. Có 3 hộp giống nhau. Hộp thứ nhất đựng 10 sản phẩm, trong đó có 6 chính phẩm, hộp thứ hai đựng 15 sản phẩm, trong đó 10 chính phẩm, hộp thứ 3 đựng 20 sản phẩm, trong đó có 15 chính phẩm. Lấy ngẫu nhiên một hộp và từ đó lấy ngẫu nhiên một sản phẩm. Tìm xác suất để lấy được chính phẩm.

Giải. Gọi A là biến cố "Lấy được chính phẩm". Biến cố A có thể xảy ra đồng thời với một trong 3 biến cố sau đây tạo nên một nhóm đầy đủ các biến cố:

H_1 - Sản phẩm lấy ra thuộc hộp I.

H_2 - Sản phẩm lấy ra thuộc hộp II.

H_3 - Sản phẩm lấy ra thuộc hộp III.

Vì theo giả thiết của bài toán, các biến cố H_1, H_2 và H_3 là đồng khả năng, do đó:

$$P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$$

Xác suất có điều kiện của biến cố A khi các biến cố H_1, H_2 và H_3 xảy ra bằng

$$P(A/H_1) = \frac{6}{10}; P(A/H_2) = \frac{10}{15}; P(A/H_3) = \frac{15}{20}$$

Do đó:

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2) + P(H_3)P(A/H_3) =$$

$$= \frac{1}{3}\cdot\frac{6}{10} + \frac{1}{3}\cdot\frac{10}{15} + \frac{1}{3}\cdot\frac{15}{20} = \frac{124}{180} = \frac{31}{45}$$

Thí dụ 8. Có 2 hộp đựng sản phẩm. Hộp thứ nhất có 10 sản phẩm trong đó có 9 chính phẩm. Hộp thứ hai có 20 sản phẩm trong đó có 18 chính phẩm. Từ hộp thứ nhất lấy ngẫu nhiên một sản phẩm bỏ sang hộp thứ hai. Tìm xác suất để lấy ngẫu nhiên một sản phẩm từ hộp thứ hai được chính phẩm.

Giải. Gọi A là biến cố "Lấy được chính phẩm từ hộp thứ hai". Biến cố A có thể xảy ra đồng thời với một trong hai biến cố sau đây tạo nên một nhóm đầy đủ các biến cố:

H_1 - Sản phẩm bỏ từ hộp thứ nhất sang hộp thứ hai là chính phẩm.

H_2 - Sản phẩm bỏ từ hộp thứ nhất sang hộp thứ hai là phế phẩm.

Xác suất để từ hộp một bỏ sang hộp hai chính phẩm

bằng $P(H_1) = \frac{9}{10}$. Xác suất để từ hộp một bỏ sang hộp hai phế phẩm bằng $P(H_2) = \frac{1}{10}$.

Xác suất có điều kiện để từ hộp hai lấy được chính phẩm khi các giả thuyết H_1 và H_2 xảy ra là:

$$P(A/H_1) = \frac{19}{21};\ P(A/H_2) = \frac{18}{21} = \frac{6}{7}$$

Do đó:

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2)$$

$$= \frac{9}{10}\cdot\frac{19}{21} + \frac{1}{10}\cdot\frac{18}{21} = 0,9$$

9.6. Công thức Bayes

Giả sử biến cố A có thể xảy ra đồng thời với một trong n biến cố $H_1, H_2, ..., H_n$ tạo nên một nhóm đầy đủ các biến cố.

Lúc đó:

$$P(H_i/A) = \frac{P(H_i)P(A/H_i)}{\sum_{i=1}^{n} P(H_i)P(A/H_i)} \quad (i = \overline{1,n}) \qquad (1.23)$$

Chứng minh. Theo định lý nhân xác suất ta có:

$$P(AH_i) = P(A)P(H_i/A) = P(H_i)P(A/H_i) \quad (i = \overline{1,n})$$

Từ đó:

$$P(A)P(H_i/A) = P(H_i)P(A/H_i) \quad (i = \overline{1,n})$$

hay:

$$P(H_i/A) = \frac{P(H_i)P(A/H_i)}{P(A)} \quad (i = \overline{1,n})$$

Thay P(A) bằng công thức xác suất đầy đủ ta có:

$$P(H_i/A) = \frac{P(H_i)P(A/H_i)}{\sum_{i=1}^{n} P(H_i)P(A/H_i)} \quad (i = \overline{1,n})$$

Như trên đã nói, các biến cố H_1, H_2, ..., H_n thường được gọi là các *giả thuyết*. Các xác suất $P(H_1)$, $P(H_2)$, ..., $P(H_n)$ được xác định trước khi phép thử được tiến hành, do đó thường được gọi là các *xác suất tiên nghiệm*. Còn các xác suất $P(H_1/A)$, $P(H_2/A)$, ..., $P(H_n/A)$ được xác định sau khi phép thử đã tiến hành và biến cố A đã xảy ra, do đó được gọi là các *xác suất hậu nghiệm*. Như vậy, công thức Bayes cho phép đánh giá lại xác suất xảy ra các giả thuyết sau khi đã biết kết quả của phép thử là biến cố A đã xảy ra.

Thí dụ 9. Dây chuyền lắp ráp nhận được các chi tiết do hai máy sản xuất. Trung bình máy thứ nhất cung cấp 60% chi tiết, máy thứ hai cung cấp 40% chi tiết. Khoảng 90% chi tiết do máy thứ nhất sản xuất là đạt tiêu chuẩn, còn 85% chi tiết do máy thứ hai sản xuất là đạt tiêu chuẩn. Lấy ngẫu nhiên từ dây chuyền một sản phẩm, thấy nó đạt tiêu chuẩn. Tìm xác suất để sản phẩm đó do máy thứ nhất sản xuất.

Giải. Gọi A là biến cố "Chi tiết lấy từ dây chuyền đạt tiêu chuẩn". Biến cố A có thể xảy ra đồng thời với một trong hai biến cố sau đây tạo nên một nhóm đầy đủ các biến cố:

H_1 - Chi tiết do máy một sản xuất;

H_2 - Chi tiết do máy hai sản xuất.

Như vậy, xác suất để chi tiết do máy một sản xuất bằng:

$$P(H_1/A) = \frac{P(H_1)P(A/H_1)}{P(H_1)P(A/H_1) + P(H_2)P(A/H_2)}$$

Theo điều kiện bài toán:

$$P(H_1) = 0,6 \qquad P(A/H_1) = 0,9$$
$$P(H_2) = 0,4 \qquad P(A/H_2) = 0,85$$

Do đó:

$$P(H_1/A) = \frac{0,6.0,9}{0,6.0,9 + 0,4.0,85} = 0,614.$$

Như ta thấy, trước khi phép thử được thực hiện, xác suất của giả thuyết H_1 bằng 0,6. Còn sau khi đã biết kết quả của phép thử thì xác suất đó đã thay đổi và bằng 0,614.

Thí dụ 10. **Trước khi đưa sản phẩm ra thị trường người ta đã phỏng vấn ngẫu nhiên 200 khách hàng về sản phẩm đó và thấy có 34 người trả lời "sẽ mua", 96 người trả lời "Có thể sẽ mua" và 70 người trả lời "Không mua". Kinh nghiệm cho thấy tỷ lệ khách hàng thực sự sẽ mua sản phẩm tương ứng với những cách trả lời trên là 40%, 20% và 1%.**

a) Hãy đánh giá thị trường tiềm năng của sản phẩm đó.

b) Trong số khách hàng thực sự mua sản phẩm thì có bao nhiêu phần trăm trả lời "sẽ mua"?

Giải. **a) Thị trường tiềm năng của sản phẩm chính là tỷ lệ khách hàng thực sự sẽ mua sản phẩm đó. Vì vậy, gọi A là biến cố "lấy ngẫu nhiên một khách hàng thì người đó thực sự sẽ mua sản phẩm". Có 3 giả thuyết đối với người khách hàng đó:**

H_1 - Người đó trả lời "Sẽ mua".

H_2 - Người đó trả lời "Có thể mua".

H_3 - Người đó trả lời "Không mua".

Theo công thức xác suất đầy đủ:

$$P(A) = P(H_1).P(A/H_1) + P(H_2).P(A/H_2) + P(H_3).P(A/H_3) =$$

$$= \frac{34}{200} \cdot 0,4 + \frac{96}{200} \cdot 0,2 + \frac{70}{200} \cdot 0,01 = 0,1675$$

Vậy thị trường tiềm năng của sản phẩm đó là 16,75%

b) Theo công thức Bayes

$$P(H_1/A) = \frac{P(H_1).P(A/H_1)}{P(A)} = \frac{0,17.0,4}{0,1675}$$

$$= 0,40597 = 40,597\%$$

Thí dụ 11. Có 2 lô sản phẩm, lô thứ nhất có tỷ lệ chính phẩm là $\frac{3}{4}$, còn lô thứ hai có tỷ lệ chính phẩm là $\frac{2}{3}$. Lấy ngẫu nhiên một lô và từ đó lấy ngẫu nhiên một sản phẩm thấy nó là chính phẩm. Sản phẩm được bỏ trở lại và từ lô đó lấy tiếp một sản phẩm. Tìm xác suất để lần thứ hai cũng lấy được chính phẩm.

Giải. Gọi A là biến cố "Sản phẩm lấy lần đầu là chính phẩm". Biến cố A có thể xảy ra với một trong hai giả thuyết sau:

H_1 - Sản phẩm được lấy ra từ lô I;

H_2 - Sản phẩm được lấy ra từ lô II.

Theo công thức xác suất đầy đủ, ta có:

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2)$$

Theo điều kiện đầu bài

$$P(H_1) = P(H_2) = \frac{1}{2}$$

$$P(A/H_1) = \frac{3}{4}; P(A/H_2) = \frac{2}{3}$$

Do đó:

$$P(A) = \frac{1}{2}\cdot\frac{3}{4} + \frac{1}{2}\cdot\frac{2}{3} = \frac{17}{24}$$

Sau khi biến cố A đã xảy ra, xác suất của các biến cố H_1 và H_2 thay đổi theo công thức Bayes như sau:

$$P(H_1/A) = \frac{P(H_1).P(A/H_1)}{P(A)} = \frac{3}{8} : \frac{17}{24} = \frac{9}{17}$$

$$P(H_2/A) = \frac{P(H_2).P(A/H_2)}{P(A)} = \frac{1}{3} : \frac{17}{24} = \frac{8}{17}$$

Gọi B là biến cố "Sản phẩm lấy lần thứ hai là chính phẩm". B vẫn có thể xảy ra với một trong hai giả thuyết H_1 và H_2, do đó theo công thức xác suất đầy đủ:

$$P(B) = P(H_1/A)P(B/H_1A) + P(H_2/A)P(B/H_2A)$$

Vì sản phẩm thứ nhất được bỏ trở lại lô, do đó tỷ lệ chính phẩm ở các lô đó vẫn không thay đổi. Vì thế:

$$P(B/H_1A) = \frac{3}{4}; P(B/H_2A) = \frac{2}{3}$$

$$P(B) = \frac{9}{17}\cdot\frac{3}{4} + \frac{8}{17}\cdot\frac{2}{3} = \frac{145}{204} = 0,71$$

Các ký hiệu và công thức cơ bản

* A, B, C, A_1, A_2, ..., A_n - Biến cố ngẫu nhiên

* U - Biến cố chắc chắn

* V - Biến cố không thể có

* $\overline{A}$ - Biến cố đối lập với biến cố A

* P(A) - Xác suất của biến cố A

* Định nghĩa cổ điển về xác suất

$P(A) = \frac{m}{n}$, trong đó: m - Số kết cục thuận lợi

n - Số kết cục duy nhất đồng khả năng

* Định nghĩa thống kê về xác suất

$P(A) \approx f = \frac{k}{n}$, trong đó: k - Số lần xuất hiện biến cố

n - Số phép thử được thực hiện

* Định lý cộng xác suất

P(A + B) = P(A) + P(B) nếu A và B xung khắc

P(A + B) = P(A) + P(B) – P(AB) nếu A và B không xung khắc

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \text{ nếu } A_1, ..., A_n \text{ xung khắc từng đôi}$$

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) - ... + (-1)^{n-1}.P(A_1 ... A_n)$$

nếu A_1, ..., A_n không xung khắc.

thăm "không". Chứng minh rằng phương pháp này là công bằng.

16. Hãy cho biết hai ý nghĩa có thể khai thác từ giá trị xác suất của một biến cố. Cho ví dụ minh họa.

17. Điều kiện cần và đủ để hai biến cố A và B độc lập là gì?

18. Mục đích của việc sử dụng công thức Bayes là gì? Cho ví dụ minh họa.

Chương II
BIẾN NGẪU NHIÊN VÀ QUY LUẬT PHÂN PHỐI XÁC SUẤT

Ở chương I ta đã nghiên cứu các loại biến cố và phương pháp tính xác suất xảy ra của các biến cố đó. Nó cho phép ta chuyển sang nghiên cứu khái niệm trung tâm của lý thuyết xác suất, đó là khái niệm về biến ngẫu nhiên.

§1. ĐỊNH NGHĨA VÀ PHÂN LOẠI BIẾN NGẪU NHIÊN

1.1. Định nghĩa

Một biến số được gọi là ngẫu nhiên nếu trong kết quả của phép thử nó sẽ nhận một và chỉ một trong các giá trị có thể có của nó tùy thuộc vào sự tác động của các nhân tố ngẫu nhiên.

Các biến ngẫu nhiên được ký hiệu là X, Y, Z hoặc $X_1, X_2, ..., X_n$; $Y_1, Y_2, ..., Y_m, ...$, còn các giá trị có thể có của chúng được ký hiệu là $x_1, x_2, ..., x_n, x, y_1, y_2, ..., y_m, y...$

Ta chú ý rằng sở dĩ biến X nào đó gọi là ngẫu nhiên vì trước khi tiến hành phép thử ta chưa có thể nói một cách chắc chắn nó sẽ nhận giá trị bằng bao nhiêu, mà chỉ có thể dự đoán điều đó với một xác suất nhất định. Nói cách khác,

việc X nhận một giá trị nào đó ($X = x_1$) hoặc ($X = x_2$), ..., ($X = x_n$) về thực chất là các biến cố ngẫu nhiên. Hơn nữa vì trong kết quả của phép thử biến X nhất định sẽ nhận một và chỉ một trong các giá trị có thể có của nó, do đó các biến cố ($X = x_1$), ($X = x_2$), ..., ($X = x_n$) tạo nên một nhóm đầy đủ các biến cố.

Thí dụ 1. Tung một con xúc xắc. Gọi X là "Số chấm xuất hiện". X là biến ngẫu nhiên vì trong kết quả của phép thử nó sẽ nhận 1 trong 6 giá trị có thể có là 1, 2, 3, 4, 5, 6.

Thí dụ 2. Gọi X là "Số con trai trong 100 trẻ sắp được sinh ra tại một nhà hộ sinh". X cũng là một biến ngẫu nhiên.

Thí dụ 3. Gọi Y là "Khoảng cách từ điểm chạm của viên đạn đến tâm bia". Y là biến ngẫu nhiên.

1.2. Phân loại biến ngẫu nhiên

Biến ngẫu nhiên có thể là rời rạc hoặc liên tục.

Biến ngẫu nhiên gọi là rời rạc nếu các giá trị có thể có của nó lập nên một tập hợp hữu hạn hoặc đếm được.

Nói cách khác, biến ngẫu nhiên sẽ là rời rạc nếu ta có thể liệt kê được tất cả các giá trị có thể có của nó.

Thí dụ 4. Trong phép thử về tung con xúc xắc, nếu ta gọi X là "Số điểm thu được" thì X là biến ngẫu nhiên rời rạc vì các giá trị có thể có của nó là một tập hợp hữu hạn.

Thí dụ 5. Gọi Y là "Số người vào mua hàng tại một siêu thị trong một ngày". Y là một biến ngẫu nhiên rời rạc vì các giá trị có thể có của nó lập nên một tập hợp đếm được Y = 0, 1, 2,...

Thí dụ 6. Một phân xưởng có 5 máy hoạt động. Gọi X là

"Số máy hỏng trong một ca". X là biến ngẫu nhiên rời rạc với các giá trị có thể có là X = 0, 1, 2, 3, 4, 5.

Biến ngẫu nhiên gọi là liên tục nếu các giá trị có thể có của nó lấp đầy một khoảng trên trục số.

Đối với biến ngẫu nhiên *liên tục* ta không thể liệt kê được tất cả các giá trị có thể có của nó.

Thí dụ 7. Phép thử là bắn một phát súng vào bia. Nếu gọi X là "Khoảng cách từ điểm chạm của viên đạn đến tâm bia" thì X là biến ngẫu nhiên liên tục vì ta không thể kể ra được tất cả các giá trị có thể có của nó. Ta chỉ có thể nói rằng các giá trị có thể có của X nằm trong khoảng (a, b) nào đó.

Thí dụ 8. Gọi Y là "Sai số khi đo lường một đại lượng vật lý", Y là biến ngẫu nhiên liên tục.

Thí dụ 9. Gọi X là "Kích thước của chi tiết do một máy sản xuất ra", X là biến ngẫu nhiên liên tục.

Thí dụ 10. Gọi X là "Năng suất lúa vụ mùa của một tỉnh", X là biến ngẫu nhiên liên tục.

Có thể nói rằng gần như tất cả các đại lượng mà ta gặp trong thực tế đều là các biến ngẫu nhiên và chúng sẽ thuộc về một trong hai loại biến ngẫu nhiên đã kể ở trên.

§2. QUY LUẬT PHÂN PHỐI XÁC SUẤT CỦA BIẾN NGẪU NHIÊN

Ta có thể nghĩ rằng chỉ cần xác định các giá trị có thể có của một biến ngẫu nhiên là đủ để xác định biến ngẫu nhiên ấy. Tuy nhiên điều này chưa đủ. Trong thực tế có những đại

lượng rất khác nhau mà các giá trị có thể có của chúng lại giống nhau. Hơn nữa việc các biến ngẫu nhiên nhận một giá trị nào đó trong kết quả của phép thử chỉ là một biến cố ngẫu nhiên, do đó nếu chỉ mới biết được các giá trị có thể có của nó thì ta mới nắm được rất ít thông tin về biến ngẫu nhiên ấy. Vì vậy ta còn phải xác định các xác suất tương ứng với các giá trị có thể có của biến ngẫu nhiên để hoàn toàn xác định nó. Từ đó ta có định nghĩa sau đây.

2.1. Định nghĩa

Quy luật phân phối xác suất của biến ngẫu nhiên là sự tương ứng giữa các giá trị có thể có của nó và các xác suất tương ứng với các giá trị đó.

Trong thực tế người ta thường sử dụng ba phương pháp để mô tả quy luật phân phối xác suất của biến ngẫu nhiên là: Bảng phân phối xác suất, hàm phân bố xác suất và hàm mật độ xác suất. Ta sẽ lần lượt nghiên cứu các phương pháp đó.

2.2. Bảng phân phối xác suất

Bảng phân phối xác suất chỉ dùng để mô tả quy luật phân phối xác suất của các biến ngẫu nhiên rời rạc.

Giả sử biến ngẫu nhiên rời rạc X có thể nhận một trong các giá trị có thể có là $x_1, x_2, ..., x_n$ với các xác suất tương ứng là $p_1, p_2, ..., p_n$. Bảng phân phối xác suất của biến ngẫu nhiên rời rạc X có dạng:

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

Ta chú ý rằng để tạo nên một quy luật phân phối xác suất thì các xác suất p_i phải thỏa mãn điều kiện:

$$\begin{cases} 0 \le p_i \le 1 \forall i \\ \sum_{i=1}^{n} p_i = 1 \end{cases} \qquad (2.1)$$

Điều kiện thứ nhất là hiển nhiên theo tính chất của xác suất, còn điều kiện thứ hai suy ra từ định nghĩa của biến ngẫu nhiên. Do các biến cố $(X = x_1)$, $(X = x_2)$, ..., $(X = x_n)$ tạo nên một nhóm đầy đủ các biến cố nên tổng các xác suất của chúng bằng một.

Thí dụ 1. Tung một con xúc xắc. Gọi X là "Số chấm xuất hiện". Hãy tìm quy luật phân phối xác suất của X.

Giải. Vì X là biến ngẫu nhiên rời rạc với các giá trị có thể có X = 1, 2, 3, 4, 5, 6 với các xác suất tương ứng đều bằng 1/6, do đó bảng phân phối xác suất của X có dạng:

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Kiểm tra: $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$

Thí dụ 2. Trong hộp có 10 sản phẩm trong đó có 6 chính phẩm. Lấy ngẫu nhiên 2 sản phẩm. Tìm quy luật phân phối xác suất của số chính phẩm được lấy ra.

Giải. Gọi Y là "Số chính phẩm lấy ra trong 2 sản phẩm". Y là biến ngẫu nhiên rời rạc với các giá trị có thể có Y = 0, 1, 2.

Ta tìm các xác suất tương ứng.

Xác suất P(Y = 0) chính là xác suất để trong 2 sản phẩm lấy ra không có chính phẩm nào (được 2 phế phẩm). Theo định nghĩa cổ điển về xác suất ta có:

$$P(Y = 0) = \frac{C_4^2}{C_{10}^2} = \frac{6}{45} = \frac{2}{15}$$

Tương tự: $$P(Y = 1) = \frac{C_6^1.C_4^1}{C_{10}^2} = \frac{24}{45} = \frac{8}{15}$$

$$P(Y = 2) = \frac{C_6^2}{C_{10}^2} = \frac{15}{45} = \frac{5}{15}$$

Như vậy quy luật phân phối xác suất của X có dạng:

X	0	1	2
P	$\frac{2}{15}$	$\frac{8}{15}$	$\frac{5}{15}$

Kiểm tra: $\frac{2}{15} + \frac{8}{15} + \frac{5}{15} = 1$

Thí dụ 3. Xác suất để xạ thủ bắn trúng bia là 0,8. Xạ thủ được phát từng viên đạn để bắn cho đến khi trúng bia. Tìm quy luật phân phối xác suất của số viên đạn được phát.

Giải. Gọi X là "Số viên đạn xạ thủ được phát". X là biến ngẫu nhiên rời rạc với các giá trị có thể có X = 1, 2, ..., k, ...

Ta tìm các xác suất tương ứng.

Xác suất P(X = 1) là xác suất để số viên đạn được phát bằng 1. Muốn xảy ra biến cố đó thì ngay phát đạn đầu tiên xạ thủ phải bắn trúng bia. Do đó:

$$P(X = 1) = 0,8$$

Xác suất P(X = 2) là xác suất để người ấy được phát 2 viên đạn. Muốn vậy phải xảy ra đồng thời hai biến cố: Phát

thứ nhất bắn trượt và phát thứ hai bắn trúng. Theo định lý nhân xác suất ta có:

$$P(X = 2) = 0,2 \,.\, 0,8$$

Ta tìm xác suất tổng quát $P(X = k)$. Biến cố $(X = k)$ là tích của k biến cố: k–1 phát đầu bắn trượt và phát thứ k bắn trúng. Theo định lý nhân xác suất ta có:

$$P(X = k) = (0,2)^{k-1} \,.\, 0,8$$

Như vậy bảng phân phối xác suất của X có dạng:

X	1	2	...	k	...
P	0,8	0,2.0,8	...	$0,2^{k-1}.0,8$	...

Kiểm tra: Ta sẽ chứng tỏ rằng các xác suất vừa tìm được tạo nên một quy luật phân phối xác suất. Thật vậy, điều kiện thứ nhất hiển nhiên thỏa mãn. Còn đối với điều kiện thứ hai ta phải tìm tổng

$$0,8 + 0,2 \,.\, 0,8 + ... + (0,2)^{k-1} \,.\, 0,8 + ...$$

Song đó chính là tổng của một cấp số nhân lùi vô hạn với công bội $q = 0,2$, do đó:

$$\sum_{k=1}^{\infty}(0,2)^{k-1}0,8 = \frac{0,8}{1-0,2} = 1$$

Hạn chế của bảng phân phối xác suất là nó chỉ mô tả được quy luật phân phối xác suất của biến ngẫu nhiên rời rạc mà thôi.

2.3. Hàm phân bố xác suất

Khái niệm hàm phân bố xác suất áp dụng được đối với cả biến ngẫu nhiên rời rạc và liên tục. Giả sử X là biến ngẫu

nhiên bất kỳ, x là một số thực nào đó. Xét biến cố "Biến ngẫu nhiên X nhận giá trị nhỏ hơn x", ký hiệu (X < x). Hiển nhiên là x thay đổi thì xác suất P (X < x) cũng thay đổi theo. Như vậy, xác suất này là một hàm số của x.

1. Định nghĩa. *Hàm phân bố xác suất của biến ngẫu nhiên X, ký hiệu F(x), là xác suất để biến ngẫu nhiên X nhận giá trị nhỏ hơn x, với x là một số thực bất kỳ.*

$$F(x) = P(X < x) \tag{2.2}$$

Ta chú ý rằng đây chỉ là định nghĩa tổng quát của hàm phân bố xác suất. Đối với từng loại biến ngẫu nhiên hàm phân bố xác suất được tính theo những công thức riêng. Chẳng hạn nếu X là biến ngẫu nhiên rời rạc thì hàm phân bố xác suất được xác định bằng công thức:

$$F(x) = \sum_{x_i < x} P_i$$

Thí dụ 4. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

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

Hãy tìm hàm phân bố xác suất của X và vẽ đồ thị.

Giải. Nếu $x \leq 1$ biến cố $(X < x)$ là biến cố không thể có, do đó:

$$F(x) = 0.$$

Nếu $1 < x \leq 3$ biến cố $(X < x)$ chỉ xảy ra khi $(X = 1)$, do đó:

$$F(x) = 0,1.$$

Nếu $3 < x \leq 4$ biến cố $(X < x)$ sẽ xảy ra hoặc khi $(X = 1)$ hoặc khi $(X = 3)$ do đó:

$$F(x) = 0,1 + 0,5 = 0,6.$$

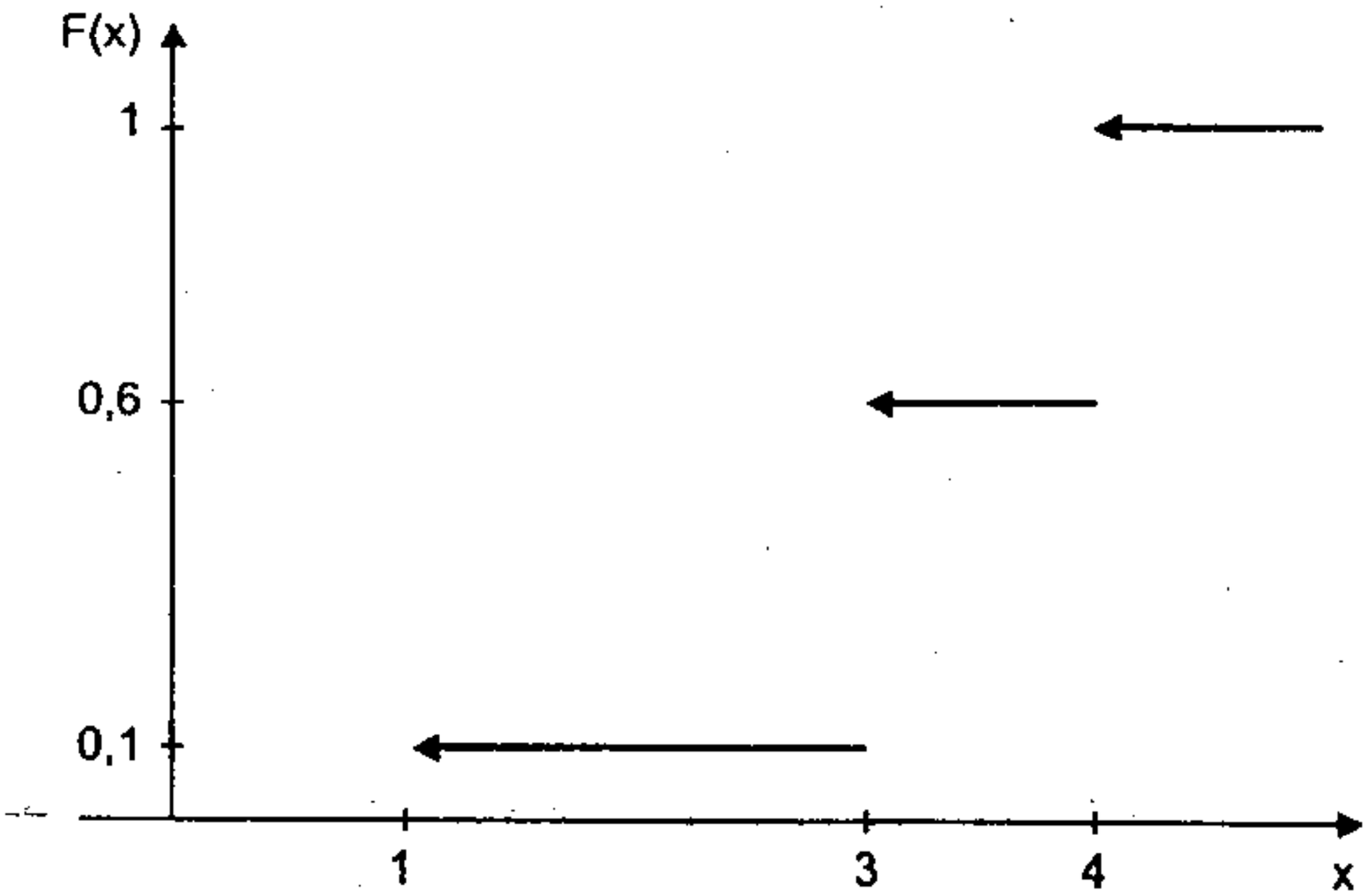
Nếu $x > 4$ biến cố $(X < x)$ sẽ xảy ra hoặc khi $(X = 1)$ hoặc khi $(X = 3)$ hoặc khi $(X = 4)$, do đó:

$$F(x) = 0,1 + 0,5 + 0,4 = 1$$

Vậy hàm phân bố xác suất của X có dạng:

$$F(x) = \begin{cases} 0 \text{ với } x \leq 1 \\ 0,1 \text{ với } 1 < x \leq 3 \\ 0,6 \text{ với } 3 < x \leq 4 \\ 1 \text{ với } x > 4 \end{cases}$$

Đồ thị của hàm F(x) có dạng (hình 2.1).



Hình 2.1. **Đồ thị hàm F(x)**

Như vậy đồ thị của hàm phân bố xác suất của biến ngẫu nhiên rời rạc có dạng bậc thang với số điểm gián đoạn chính bằng số giá trị có thể có của X.

2. Các tính chất của hàm phân bố xác suất

Tính chất 1. Hàm phân bố xác suất luôn nhận giá trị trong đoạn [0, 1]:

$$0 \le F(x) \le 1 \qquad (2.3)$$

Tính chất này trực tiếp suy ra từ định nghĩa của hàm phân bố xác suất, vì nó là một xác suất nên giá trị của nó luôn nằm trong đoạn [0; 1].

Tính chất 2. Hàm phân bố xác suất là hàm không giảm, tức là với $x_2 > x_1$ thì:

$$F(x_2) \ge F(x_1)$$

Chứng minh. Giả sử $x_2 > x_1$. Xét biến cố $(X < x_2)$. Biến cố này có thể phân tích thành tổng của hai biến cố xung khắc là $(X < x_1)$ và $(x_1 \le X < x_2)$. Theo định lý cộng xác suất ta có:

$$P(X < x_2) = P(X < x_1) + P(x_1 \le X < x_2)$$

Từ đó $$P(X < x_2) - P(X < x_1) = P(x_1 \le X < x_2)$$

hay $$F(x_2) - F(x_1) = P(x_1 \le X < x_2)$$

Song vế phải là một xác suất, nó luôn không âm, do đó ta có:

$$F(x_2) - F(x_1) \ge 0 \text{ từ đó } F(x_2) \ge F(x_1)$$

Từ tính chất thứ hai có thể suy ra một số hệ quả sau đây:

Hệ quả 1. Xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng [a, b) bằng hiệu số của hàm phân bố xác suất tại hai đầu khoảng đó:

$$P(a \le X < b) = F(b) - F(a) \qquad (2.4)$$

Hệ quả này suy ra trực tiếp từ quá trình chứng minh tính chất.

Hệ quả 2. Xác suất để biến ngẫu nhiên liện tục X nhận một giá trị xác định bằng 0:

$$P(X = x) = 0 \qquad (2.5)$$

Thật vậy, nếu tà đặt a = x và b = x + Δx ta có:

$$P(x \le X < x + \Delta x) = F(x + \Delta x) - F(x)$$

Lấy giới hạn của cả hai vế khi $\Delta x \to 0$

$$\lim_{\Delta x \to 0} P(x \le X < x + \Delta x) = \lim_{\Delta x \to 0} F(x + \Delta x) - F(x)$$

Vì X là biến ngẫu nhiên liên tục, do đó tại điểm x hàm phân bố xác suất cũng liên tục. Vì vậy $\lim_{\Delta x \to 0} F(x + \Delta x) = F(x)$

Từ đó: $P(X = x) = F(x) - F(x) = 0.$

Hệ quả 3. Đối với biến ngẫu nhiên liên tục X ta có các đẳng thức sau đây:

$$P(a \le x \le b) = P(a \le X < b) = P(a < X \le b)$$
$$= P(a < X < b) \qquad (2.6)$$

Chẳng hạn đẳng thức $P(a \le X < b) = P(a < X < b)$ có thể chứng minh như sau:

$$P(a \le X < b) = P(X = a) + P(a < X < b)$$
$$= P(a < X < b)$$

Như vậy việc xét xác suất để biến ngẫu nhiên liên tục X nhận một giá trị xác định là không có ý nghĩa, song việc tìm xác suất để nó nhận giá trị trong một khoảng, dù là rất nhỏ lại có ý nghĩa.

Tính chất 3. Ta có biểu thức giới hạn sau:

$$F(-\infty) = 0; \quad F(+\infty) = 1 \qquad (2.7)$$

Thật vậy:

$$F(-\infty) = P(X < -\infty) = P(V) = 0$$

$$F(+\infty) = P(X < +\infty) = P(U) = 1$$

Từ tính chất trên có thể suy ra hệ quả sau:

Hệ quả. Nếu biến ngẫu nhiên X chỉ nhận giá trị trong đoạn [a, b] thì với $x \le a$, $F(x) = 0$ và với $x > b$, $F(x) = 1$.

Thật vậy, với $x \le a$ biến cố $(X < x)$ là biến cố không thể có, do đó xác suất của nó bằng 0. Còn với $x > b$ thì biến cố $(X < x)$ là biến cố chắc chắn, do đó xác suất của nó bằng 1.

Chú ý rằng, nếu X là biến ngẫu nhiên rời rạc thì hàm phân bố xác suất chỉ liên tục về phía trái tại mỗi giá trị có thể có của nó, còn về phía phải thì nó bị gián đoạn.

3. Ý nghĩa của hàm phân bố xác suất

Từ định nghĩa của hàm phân bố xác suất $F(x) = P(X < x)$ ta thấy hàm phân bố xác suất phản ánh mức độ tập trung xác suất ở về phía bên trái một số thực x nào đó. Như đã biết toàn bộ xác suất của biến ngẫu nhiên bằng một, do đó giá trị của hàm phân bố xác suất tại mỗi điểm x cho biết có bao nhiêu phần của một đơn vị xác suất phân bố trong đoạn $(-\infty, x)$.

Thí dụ 5. Biến ngẫu nhiên X có hàm phân bố xác suất như sau:

$$F(x) = \begin{cases} 0 & \text{với } x \le -1 \\ \frac{3}{4}x + \frac{3}{4} & \text{với } -1 < x < \frac{1}{3} \\ 1 & \text{với } x > \frac{1}{3} \end{cases}$$

Tìm xác suất để trong kết quả của phép thử, X nhận giá trị trong khoảng $[0, \frac{1}{3})$.

Theo tính chất của hàm phân bố xác suất:

$$P(0 \le X < \frac{1}{3}) = F(\frac{1}{3}) - F(0)$$

Vì trong khoảng $[0, \frac{1}{3})$ giá trị của hàm phân bố xác suất bằng:

$$F(x) = \frac{3}{4}x + \frac{3}{4}$$

do đó: $$F\left(\frac{1}{3}\right) - F(0) = \left[\frac{3}{4} \cdot \frac{1}{3} + \frac{3}{4}\right] - \left[\frac{3}{4} \cdot 0 + \frac{3}{4}\right] = \frac{1}{4}$$

Như vậy: $$P\left(0 \le X < \frac{1}{3}\right) = \frac{1}{4}$$

2.4. Hàm mật độ xác suất

Đối với biến ngẫu nhiên liên tục X có thể dùng hàm phân bố xác suất để mô tả quy luật phân phối xác suất của nó. Tuy nhiên phương pháp này cũng có hạn chế. Hàm phân bố xác suất không thể đặc trưng được xác suất để biến ngẫu nhiên liên tục X nhận một giá trị xác định. Vì thế, đối với các biến ngẫu nhiên liên tục, người ta thường dùng hàm mật độ xác suất để mô tả quy luật phân phối xác suất của nó.

1. Định nghĩa. *Hàm mật độ xác suất của biến ngẫu nhiên liên tục X (ký hiệu là f(x)) là đạo hàm bậc nhất của hàm phân bố xác suất của biến ngẫu nhiên đó.*

$$f(x) = F'(x) \qquad (2.8)$$

Chú ý rằng, khái niệm hàm mật độ xác suất chỉ áp dụng được đối với các biến ngẫu nhiên liên tục mà không áp dụng được đối với biến ngẫu nhiên rời rạc vì muốn F'(x) tồn tại thì tối thiểu F(x) phải liên tục, do đó X phải là biến ngẫu nhiên liên tục.

2. Các tính chất của hàm mật độ xác suất

Tính chất 1. Hàm mật độ xác suất luôn không âm:

$$f(x) \geq 0 \ \forall x \tag{2.9}$$

Chứng minh. Hàm phân bố xác suất F(x) là một hàm không giảm, do đó đạo hàm của nó F'(x) = f(x) là một hàm không âm. Về mặt hình học điều đó có nghĩa là đồ thị của hàm f(x) không nằm thấp hơn trục 0x.

Tính chất 2. Xác suất để biến ngẫu nhiên liên tục X nhận giá trị trong khoảng (a, b) bằng tích phân xác định của hàm mật độ xác suất trong khoảng đó:

$$P(a < X < b) = \int_a^b f(x)dx \tag{2.10}$$

Chứng minh. Theo tính chất của hàm phân bố xác suất ta có:

$$P(a \leq X < b) = F(b) - F(a)$$

Theo công thức Newton - Laibnitz

$$F(b) - F(a) = \int_a^b F'(x)dx = \int_a^b f(x)dx$$

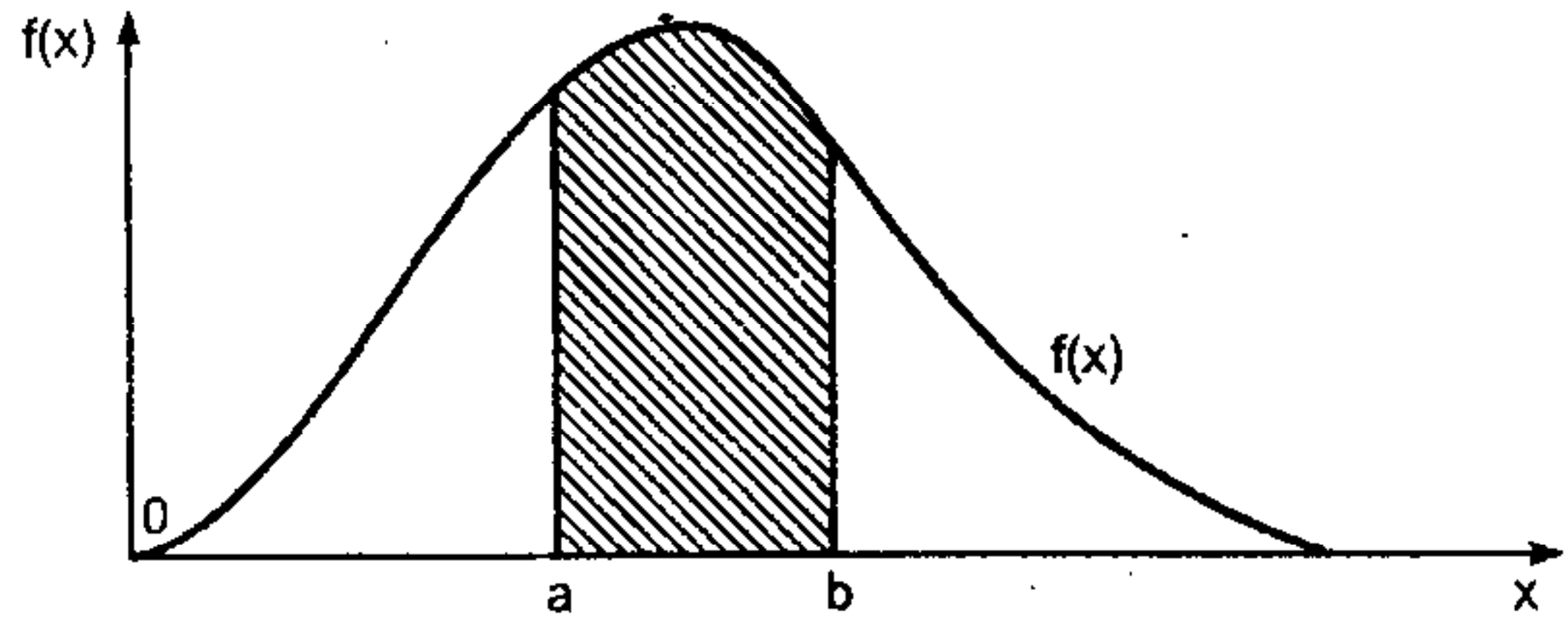
Như vậy: $P(a \leq X < b) = \int_a^b f(x)dx$

Song vì X là biến ngẫu nhiên liên tục nên

$$P(a \le X < b) = P(a < X < b)$$

từ đó ta có: $$P(a < X < b) = \int_a^b f(x)dx$$

Về mặt hình học, kết quả trên có thể minh họa như sau: Xác suất để biến ngẫu nhiên liên tục X nhận giá trị trong khoảng (a, b) bằng diện tích của hình giới hạn bởi trục 0x, đường cong f(x) và các đường thẳng x = a và x = b (xem hình 2.2).



Hình 2.2. Đồ thị hàm f(x)

Tính chất 3. Hàm phân bố xác suất F(x) của biến ngẫu nhiên liên tục X bằng tích phân suy rộng của hàm mật độ xác suất trong khoảng (–∞, x):

$$F(x) = \int_{-\infty}^{x} f(x)dx \qquad (2.11)$$

Chứng minh. Theo định nghĩa của hàm phân bố xác suất, ta có:

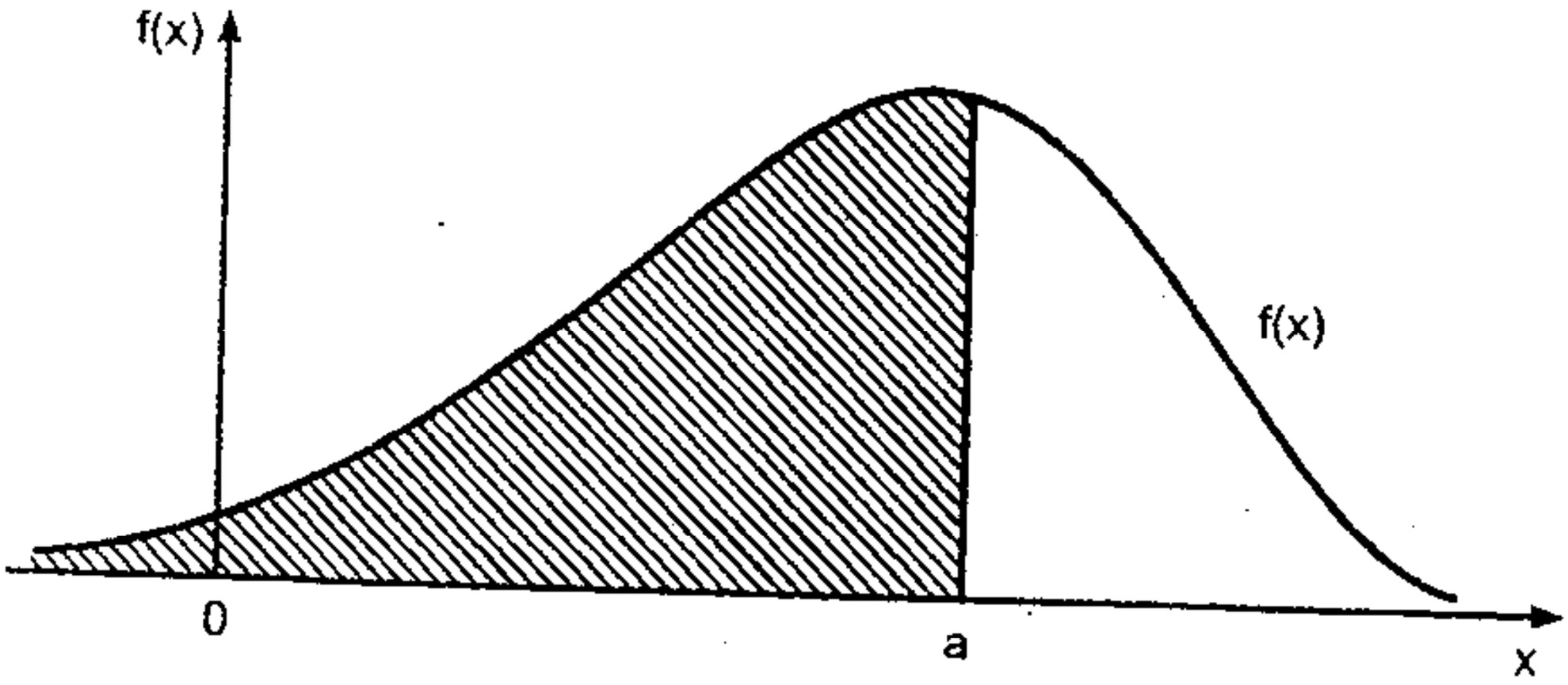
$$F(x) = P(X < x) = P(-\infty < X < x)$$

Theo tính chất 2, đặt $a = -\infty$ và $b = x$, ta có:

$$P(-\infty < X < x) = \int_{-\infty}^{x} f(x)dx$$

Công thức trên cho phép tìm hàm phân bố xác suất của biến ngẫu nhiên liên tục khi đã biết hàm mật độ xác suất của nó.

Về mặt hình học, công thức trên cho thấy giá trị của hàm phân bố xác suất F(x) tại điểm a bằng diện tích giới hạn bởi trục 0x, đường cong f(x) và đường thẳng x = a (xem hình 2.3).



Hình 2.3. Giá trị của hàm F(x)

Tính chất 4. Tích phân suy rộng trong khoảng $(-\infty, +\infty)$ của hàm mật độ xác suất bằng 1.

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \qquad (2.12)$$

Nếu theo tính chất 2 ta đặt $a = -\infty$ và $b = +\infty$, ta có:

$$P(-\infty < X < +\infty) = \int_{-\infty}^{+\infty} f(x)dx$$

Song biến cố $(-\infty < X < +\infty)$ là biến cố chắc chắn, do đó:

$$\int f(x)dx = P(U) = 1$$

Về mặt hình học, điều đó có nghĩa là toàn bộ diện tích giới hạn bởi đường cong f(x) và trục 0x bằng 1.

Ta chú ý rằng để một hàm số f(x) có thể là hàm mật độ xác suất của một biến ngẫu nhiên liên tục nào đó thì nó phải thỏa mãn hai tính chất cơ bản là tính chất 1 và tính chất 4, tức là:

$$\begin{cases} f(x) \geq 0 \quad \forall x \\ \int_{-\infty}^{+\infty} f(x)dx = 1 \end{cases} \qquad (2.13)$$

Thí dụ 6. Hàm phân bố xác suất của biến ngẫu nhiên liên tục X có dạng:

$$F(x) = \begin{cases} 0 & \text{với } x \leq 0 \\ ax^2 & \text{với } 0 < x \leq 1 \\ 1 & \text{với } x > 1 \end{cases}$$

a) Tìm hệ số a;

b) Tìm hàm mật độ xác suất f(x);

c) Tìm xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng (0,25; 0,75).

Giải. a) Vì hàm phân bố xác suất F(x) là liên tục, do đó tại $x = 1$. $ax^2 = 1$, từ đó $a = 1$.

b) Theo định nghĩa của hàm mật độ xác suất:

$$f(x) = F'(x) = \begin{cases} 0 & \text{với } x \le 0 \\ 2x & \text{với } 0 < x \le 1 \\ 0 & \text{với } x > 1 \end{cases}$$

c) Theo tính chất của hàm phân bố xác suất

$$P(0,25 < X < 0,75) = F(0,75) - F(0,25)$$

$$= (0,75)^2 - (0,25)^2 = 0,5$$

Thí dụ 7. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} a\cos x & \text{với } x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \\ 0 & \text{với } x \notin \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \end{cases}$$

a) Tìm hệ số a;

b) Tìm hàm phân bố xác suất F(x);

c) Tìm xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng $(0, \pi/4)$.

Giải. a) Theo tính chất của hàm mật độ xác suất:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\pi/2}^{\pi/2} a\cos x dx = 2a = 1$$

Từ đó: $a = \frac{1}{2}$

b) Để tìm hàm phân bố xác suất, ta sử dụng tính chất của hàm mật độ xác suất:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Với $x \le -\frac{\pi}{2}$: $F(x) = \int_{-\infty}^{x} 0dx = 0$

Với $-\frac{\pi}{2} < x \le \frac{\pi}{2}$

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{-\pi/2} 0dx + \int_{-\pi/2}^{x} \frac{1}{2}\cos x dx$$

$$= \frac{1}{2}(\sin x + 1)$$

Với $x > \frac{\pi}{2}$

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{-\pi/2} 0dx + \int_{-\pi/2}^{\pi/2} \frac{1}{2}\cos x dx + \int_{\pi/2}^{x} 0dx = 1$$

Vậy hàm phân bố xác suất của X có dạng:

$$F(x) = \begin{cases} 0 & \text{với } x \le -\frac{\pi}{2} \\ \frac{1}{2}(\sin x + 1) & \text{với } -\frac{\pi}{2} < x \le \frac{\pi}{2} \\ 1 & \text{với } x > \frac{\pi}{2} \end{cases}$$

c) Theo tính chất của hàm phân bố xác suất:

$$P\left(0 < x < \frac{\pi}{4}\right) = F\left(\frac{\pi}{4}\right) - F(0) =$$

$$= \frac{1}{2}\left(\sin\frac{\pi}{4} + 1\right) - \frac{1}{2}(\sin 0 + 1) = \frac{\sqrt{2}}{4}$$

Thí dụ 8. Hàm mật độ xác suất của biến ngẫu nhiên X có dạng:

$$f(x) = \frac{1}{\pi(1+x^2)} \qquad \forall x$$

Tìm xác suất để tiến hành 3 phép thử độc lập có 2 lần X nhận giá trị trong khoảng (–1; +1).

Giải. Trước hết ta tìm xác suất để tiến hành một phép thử biến ngẫu nhiên X nhận giá trị trong khoảng (–1; +1). Theo tính chất của hàm mật độ xác suất:

$$P(-1 < X < 1) = \int_{-1}^{1} \frac{dx}{\pi(1+x^2)} = \frac{1}{\pi}\text{arctgx}\Big|_{-1}^{1} = \frac{1}{2}$$

Theo điều kiện của bài toán, ta tiến hành ba phép thử độc lập, trong mỗi phép thử chỉ có hai khả năng: hoặc X nhận giá trị trong khoảng (–1, 1) (biến cố A), hoặc không nhận giá trị trong khoảng đó (biến cố $\overline{A}$). Xác suất để trong mỗi phép thử biến ngẫu nhiên X nhận giá trị trong khoảng (–1, 1), như đã xác định ở trên, đều bằng 0,5. Do đó theo công thức Bernoulli, xác suất để trong 3 phép thử có 2 lần X nhận giá trị trong khoảng đó bằng:

$$P_3(2) = C_3^2(0,5)^2(1-0,5) = 0,375$$

Thí dụ 9. Tuổi thọ của một loại sản phẩm là biến ngẫu nhiên liên tục có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \dfrac{a}{x^2} & \text{với } x \geq 400 \text{ (giờ)} \\ 0 & \text{với } x < 400 \text{ (giờ)} \end{cases}$$

a) Tìm a.

b) Tìm xác suất để lấy ngẫu nhiên 1 sản phẩm thì tuổi thọ của nó kéo dài ít nhất là 600 giờ.

Giải. a) Theo điều kiện của hàm mật độ xác suất

$$1 = \int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{400} 0dx + a\int_{400}^{+\infty} \frac{dx}{x^2} = -a.\frac{1}{x}\Big|_{400}^{+\infty} = \frac{a}{400}$$

$$\Rightarrow a = 400$$

$$b)\ P(X > 600) = \int_{600}^{+\infty} f(x)dx = 400\int_{600}^{+\infty} \frac{dx}{x^2} = -400.\frac{1}{x}\Big|_{600}^{+\infty} = \frac{2}{3}$$

3. Ý nghĩa của hàm mật độ xác suất

Hàm mật độ xác suất của biến ngẫu nhiên X tại mỗi điểm x cho biết mức độ tập trung xác suất tại điểm đó. Thật vậy, theo định nghĩa của hàm mật độ xác suất, ta có:

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} =$$

$$= \lim_{\Delta x \to 0} \frac{P(x \le X < x + \Delta x)}{\Delta x}$$

Như vậy hàm mật độ xác suất tại điểm x chính là giới hạn của xác suất để biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(x, x + \Delta x)$ chia cho độ dài của khoảng đó (khi $\Delta x \to 0$), tức là giới hạn của mật độ xác suất trung bình trên đoạn $(x; x + \Delta x)$ khi $\Delta x \to 0$.

§3. CÁC THAM SỐ ĐẶC TRƯNG CỦA BIẾN NGẪU NHIÊN

Như đã thấy ở trên, quy luật phân phối xác suất của biến ngẫu nhiên (dưới dạng bảng phân phối xác suất, hàm phân bố xác suất hay hàm mật độ xác suất) hoàn toàn xác định biến ngẫu nhiên.

Như vậy, khi ta đã xác định được quy luật phân phối xác suất của một biến ngẫu nhiên thì ta đã nắm được toàn bộ thông tin về biến ngẫu nhiên đó. Tuy nhiên, trong thực tế ta không chỉ cần đến những thông tin đó mà còn phải quan tâm đến những thông tin cô đọng phản ánh tổng hợp những đặc trưng quan trọng nhất của biến ngẫu nhiên được nghiên cứu. Những thông tin cô đọng phản ánh từng phần về biến ngẫu nhiên được gọi là *các tham số đặc trưng*.

Các tham số đặc trưng của biến ngẫu nhiên được chia thành ba loại sau:

- Các tham số đặc trưng cho xu hướng trung tâm của biến ngẫu nhiên như: Kỳ vọng toán, Trung vị, Mốt v.v...

- Các tham số đặc trưng cho độ phân tán của biến ngẫu nhiên như: Phương sai, Độ lệch chuẩn, Hệ số biến thiên v.v...

- Các tham số đặc trưng cho dạng phân phối xác suất.

Ở phần này chỉ hạn chế ở việc xem xét một số tham số đặc trưng quan trọng nhất. Sau này trong một số trường hợp cụ thể các tham số khác cũng sẽ được đề cập.

3.1. Kỳ vọng toán

Kỳ vọng toán có những định nghĩa riêng đối với biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục.

1. Định nghĩa. *Giả sử biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có $x_1, x_2, ..., x_n$ với các xác suất tương ứng $p_1, p_2, ..., p_n$. Kỳ vọng toán của biến ngẫu nhiên rời rạc X, ký hiệu E(X) là tổng các tích giữa các giá trị có thể có của biến ngẫu nhiên với các xác suất tương ứng:*

$$E(X) = \sum_{i=1}^{n} x_i p_i \qquad (2.14)$$

Nếu X là biến ngẫu nhiên liên tục với hàm mật độ xác suất f(x) thì kỳ vọng toán E(X) được xác định bằng biểu thức:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \qquad (2.15)$$

Thí dụ 1. Tìm kỳ vọng toán của biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

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

Giải. Theo định nghĩa kỳ vọng toán của biến ngẫu nhiên rời rạc ta có:

$$E(X) = 1.0,1 + 3.0,5 + 4.0,4 = 3,2$$

Thí dụ 2. Tìm kỳ vọng toán của biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \frac{3}{4}(x^2 + 2x) & \text{với } x \in (0.1) \\ 0 & \text{với } x \notin (0.1) \end{cases}$$

Giải. Theo định nghĩa kỳ vọng toán của biến ngẫu nhiên liên tục ta có:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \frac{3}{4}\int_0^1 x(x^2+2x)dx =$$

$$= \frac{3}{4}\int_0^1 (x^3+2x^2)dx = \frac{3}{4}\left(\frac{x^4}{4}+\frac{2x^3}{3}\right)\Big|_0^1 = \frac{11}{16}$$

Ta chú ý rằng kỳ vọng toán của biến ngẫu nhiên là một số xác định.

2. Các tính chất của kỳ vọng toán

Sau đây sẽ phát biểu và chứng minh một số tính chất cơ bản của kỳ vọng toán. Các tính chất được chứng minh đối với các biến ngẫu nhiên rời rạc. Việc chứng minh trong trường hợp các biến ngẫu nhiên là liên tục cũng tiến hành tương tự.

Tính chất 1. Kỳ vọng toán của một hằng số bằng chính hằng số đó. Như vậy, nếu C là hằng số thì E(C) = C.

Thật vậy, có thể coi C như một biến ngẫu nhiên rời rạc đặc biệt, với một giá trị có thể có bằng C và xác suất tương ứng bằng 1. Lúc đó theo định nghĩa của kỳ vọng toán ta có:

$$E(C) = C.1 = C$$

Tính chất 2. Kỳ vọng toán của tích giữa một hằng số và một biến ngẫu nhiên bằng tích giữa hằng số đó và kỳ vọng toán của biến ngẫu nhiên ấy.

$$E(CX) = C.E(X) \qquad (2.16)$$

Thật vậy, giả sử biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

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

Lúc đó tích CX sẽ là một biến ngẫu nhiên rời rạc mà các giá trị có thể có của nó bằng tích giữa hằng số C và các giá trị có thể có của X. Mặt khác, do đó là các biến cố tương đương nhau (có khả năng xảy ra như nhau) nên các xác suất tương ứng bằng nhau. Vậy bảng phân phối xác suất của biến ngẫu nhiên CX có dạng:

CX	Cx_1	Cx_2	...	Cx_n
P	p_1	p_2	...	p_n

Lúc đó theo định nghĩa kỳ vọng toán ta có:

$$E(CX) = C\,x_1p_1 + Cx_2p_2 + ... + Cx_np_n$$

$$= C(x_1p_1 + x_2p_2 + ...+ x_np_n) = C.E(X)$$

Trước khi phát biểu tính chất tiếp theo ta xét khái niệm độc lập của các biến ngẫu nhiên.

Hai biến ngẫu nhiên gọi là *độc lập* với nhau nếu quy luật phân phối xác suất của biến ngẫu nhiên này không phụ thuộc gì vào việc biến ngẫu nhiên kia nhận giá trị bằng bao nhiêu. Tương tự các biến ngẫu nhiên gọi là độc lập lẫn nhau, nếu các quy luật phân phối xác suất của một số bất kỳ các biến ngẫu nhiên nào đó không phụ thuộc vào việc các biến ngẫu nhiên còn lại nhận giá trị bằng bao nhiêu.

Tổng của hai biến ngẫu nhiên X và Y là biến ngẫu nhiên X + Y mà các giá trị có thể của nó là tổng của mỗi giá trị có thể có của X và mỗi giá trị có thể có của Y. Khi X và Y độc lập nhau thì các xác suất tương ứng sẽ bằng tích các xác suất

thành phần. Còn khi X và Y phụ thuộc nhau thì các xác suất tương ứng sẽ bằng xác suất của thành phần này nhân với xác suất có điều kiện của thành phần kia.

Tính chất sau đây đúng với cả các biến ngẫu nhiên độc lập và phụ thuộc.

Tính chất 3. Kỳ vọng toán của tổng hai biến ngẫu nhiên bằng tổng các kỳ vọng toán thành phần.

$$E(X + Y) = E(X) + E(Y) \qquad (2.17)$$

Thật vậy giả sử các biến ngẫu nhiên rời rạc X và Y có các quy luật phân phối xác suất như sau:

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

Y	y_1	y_2	...	y_m
P	q_1	q_2	...	q_m

Lúc đó ta có quy luật phân phối xác suất của tổng X + Y như sau:

X + Y	$x_1 + y_1$	$x_2 + y_2$	...	$x_n + y_m$
P	p_{11}	p_{12}	...	p_{nm}

trong đó ta ký hiệu p_{ij} là xác suất để tổng X + Y nhận giá trị bằng $x_i + y_j$.

Theo định nghĩa kỳ vọng toán ta có:

$$E(X + Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i + y_j)p_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m}x_i p_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}y_j p_{ij}$$

$$= \sum_{i=1}^{n}x_i \sum_{j=1}^{m}p_{ij} + \sum_{j=1}^{m}y_j \sum_{i=1}^{n}p_{ij}$$

Ta sẽ chứng minh rằng $\sum_{j=1}^{m}p_{ij} = p_i$

Thật vậy, biến cố $X = x_i$ sẽ xảy ra khi tổng X + Y nhận giá trị $x_i + y_1$ hoặc $x_i + y_2$ ... hoặc $x_i + y_m$. Do đó theo định lý cộng xác suất:

$$P(X = x_i) = P[(X + Y) = (x_i + y_1)] + ... + P[(X + Y) = (x_i + y_m)]$$

hay $$p_i = p_{i1} + p_{i2} + ... + p_{im} = \sum_{j=1}^{m} p_{ij} = p_i$$

Tương tự như vậy có thể chứng minh được rằng:

$$\sum_{i=1}^{n} P_{ij} = q_j$$

Từ đó ta có:

$$E(X + Y) = \sum_{i=1}^{n} x_i P_i + \sum_{j=1}^{m} y_j q_j = E(X) + E(Y)$$

Bằng phương pháp quy nạp toán học ta có thể chứng minh được hệ quả sau đây.

Hệ quả. Kỳ vọng toán của tổng n biến ngẫu nhiên X_1, X_2, ... X_n bằng tổng các kỳ vọng toán thành phần:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \qquad (2.18)$$

Tích của hai biến ngẫu nhiên độc lập X và Y là biến ngẫu nhiên XY mà các giá trị có thể có của nó là tích giữa mỗi giá trị có thể có của X và mỗi giá trị có thể có của Y. Các xác suất tương ứng là tích của các xác suất thành phần.

Tính chất 4. Kỳ vọng toán của tích hai biến ngẫu nhiên độc lập bằng tích các kỳ vọng toán thành phần

$$E(X.Y) = E(X).E(Y) \qquad (2.19)$$

Thật vậy, giả sử các biến ngẫu nhiên X và Y độc lập, có các quy luật phân phối xác suất như sau:

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

Y	y_1	y_2	...	y_m
P	q_1	q_2	...	q_m

lúc đó tích X.Y có quy luật phân phối xác suất như sau:

X.Y	x_1y_1	x_1y_2	...	x_ny_m
P	p_1q_1	p_1q_2	...	p_nq_m

Ta có:

$$E(X.Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_i q_j = \sum_{i=1}^{n} x_i p_i \sum_{j=1}^{m} y_j q_j = E(X).E(Y)$$

Bằng phương pháp quy nạp toán học có thể chứng minh được hệ quả sau đây:

Hệ quả. Kỳ vọng toán của tích n biến ngẫu nhiên X_1, X_2, ..., X_n độc lập lẫn nhau bằng tích các kỳ vọng toán thành phần.

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i) \qquad (2.20)$$

3. Bản chất và ý nghĩa của kỳ vọng toán

Giả sử đối với biến ngẫu nhiên X tiến hành n phép thử trong đó có n_1 lần X nhận giá trị x_1, n_2 lần X nhận giá trị x_2, ..., n_k lần X nhận giá trị $x_k \left(\sum_{i=1}^{k} n_i = n\right)$. Giá trị trung bình của biến ngẫu nhiên X trong n phép thử này là:

$$\bar{x} = \frac{x_1n_1 + x_2n_2 + ... + x_kn_k}{n} = x_1\frac{n_1}{n} + x_2\frac{n_2}{n} + ... + x_k\frac{n_k}{n}$$

Ta chú ý rằng $\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_k}{n}$ chính là tần suất xuất hiện các giá trị $x_1, x_2, \ldots, x_k$ trong n phép thử trên, do đó:

$$\overline{x} = x_1 f_1 + x_2 f_2 + \ldots + x_k f_k$$

Theo định nghĩa thống kê về xác suất khi $n \to \infty$ các tần suất sẽ hội tụ theo xác suất về các xác suất tương ứng, do đó với n đủ lớn ta có thể viết:

$$\overline{x} \approx x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = E(X)$$

Như vậy ta thu được kết quả:

$$E(X) \approx \overline{x} \qquad (2.21)$$

Vậy kỳ vọng toán của biến ngẫu nhiên gần bằng trung bình số học của các giá trị quan sát của biến ngẫu nhiên. Nó phản ánh giá trị trung tâm của phân phối xác suất của biến ngẫu nhiên.

Thí dụ 3. Tung con xúc xắc n lần. Tìm kỳ vọng toán của tổng số chấm thu được.

Giải. Gọi X_i $(i = \overline{1,n})$ là số chấm thu được ở lần tung thứ i và gọi X là tổng số chấm thu được trong n lần tung.

Như vậy $X = \sum_{i=1}^{n} X_i$. Theo tính chất của kỳ vọng toán

$$E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$$

Mỗi biến ngẫu nhiên X_i đều có bảng phân phối xác suất như sau:

X_i	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Do đó:

$$E(X_i) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}$$

Do đó:

$$E(X) = \frac{7}{2}n$$

Thí dụ 4. Thời gian xếp hàng chờ mua hàng của khách hàng là biến ngẫu nhiên liên tục T có hàm mật độ xác suất như sau (đơn vị: phút):

$$f(t) = \begin{cases} \frac{4}{81}t^3 & \text{với } t \in (0,3) \\ 0 & \text{với } t \notin (0,3) \end{cases}$$

Tìm thời gian xếp hàng trung bình của khách hàng.

Giải. Thời gian xếp hàng trung bình chính là kỳ vọng toán. Theo định nghĩa ta có:

$$E(T) = \int_0^3 t.f(t).dt = \frac{4}{81}\int_0^3 t^4 dt = \frac{4}{405}t^5\Big|_0^3 = 2,4 \text{ phút}$$

Thí dụ 5. Một dự án xây dựng được viện thiết kế C soạn thảo cho cả 2 bên A và B xét duyệt một cách độc lập. Xác suất để A và B chấp nhận dự án khi xét duyệt là 0,7 và 0,8. Nếu chấp nhận dự án thì A phải trả cho C là 4 triệu đồng, còn ngược lại thì phải trả 1 triệu. Với B, nếu chấp nhận dự án thì phải trả cho C là 10 triệu, ngược lại phải trả 3 triệu. Chi phí cho thiết kế là 10 triệu và thuế 10% doanh thu. Hỏi C có nên nhận thiết kế hay không?

Giải. Để quyết định xem có nên nhận thiết kê hay không thì C phải tính số lãi kỳ vọng mà C có thể nhận được. Nếu gọi

X là số lãi mà C có thể nhận được sau khi trừ mọi chi phí thì X có bảng phân phối xác suất như sau:

X	-6,4	-3,7	-0,1	2,6
P	0,06	0,14	0,24	0,56

Từ đó E(X) = 0,53 > 0. Vậy C vẫn có thể nhận thiết kế.

Thí dụ 6. Xác suất để một máy sản xuất ra phế phẩm bằng p. Máy sẽ được sửa chữa ngay sau khi làm ra phế phẩm.

Tìm số sản phẩm trung bình được sản xuất ra giữa hai lần sửa chữa.

Giải. Gọi X là số sản phẩm được sản xuất ra giữa hai lần sửa chữa. X có bảng phân phối xác suất như sau:

X	1	2	3	...	k	...
P	p	qp	q^2p	...	$q^{k-1}p$	...

trong đó q là xác suất để sản phẩm sản xuất là chính phẩm, q = 1 - p.

Số sản phẩm trung bình được sản xuất ra chính là kỳ vọng toán E(X) như đã phân tích trong phần bản chất của kỳ vọng toán.

Do đó:

$$E(X) = \sum_{k=1} kq^{k-1}p = p\sum_{k=1} kq^{k-1}$$

Song tổng $\sum_{k=1} kq^{k-1}$ có thể biểu diễn dưới dạng:

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{d}{dq}\sum_{k=0}^{\infty} q^k$$

Mà tổng $\sum_{k=0}^{\infty} q^k$ là tổng của cấp số nhân lùi vô hạn, do đó:

$$\frac{d}{dq}\sum_{k=0}^{\infty} q^k = \frac{d}{dq}\left(\frac{1}{1-q}\right) = \frac{1}{(1-q)^2} = \frac{1}{p^2}$$

do đó: $E(X) = p\frac{1}{p^2} = \frac{1}{p}$.

4. Ứng dụng thực tế của kỳ vọng toán

Khái niệm kỳ vọng toán lúc đầu xuất hiện trong các trò chơi may rủi để tính giá trị mà người chơi mong đợi sẽ nhận được. Hiện nay khái niệm này được áp dụng rộng rãi trong nhiều lĩnh vực kinh doanh và quản lý như một tiêu chuẩn để ra quyết định trong tình huống cần lựa chọn giữa nhiều chiến lược khác nhau. Tiêu chuẩn này thường được biểu diễn dưới dạng lợi nhuận kỳ vọng hay doanh số kỳ vọng để làm căn cứ lựa chọn chiến lược kinh doanh. Ta sẽ xét một bài toán như vậy làm ví dụ.

Giả sử một cửa hàng sách dự định nhập vào một số cuốn niên giám thống kê. Nhu cầu hàng năm về loại sách này được cho trong bảng phân phối xác suất sau đây:

Nhu cầu j (cuốn)	20	21	22	23	24	25
Xác suất P_j	0,3	0,25	0,18	0,14	0,1	0,03

Cửa hàng này mua vào với giá 7 USD/cuốn và bán ra với giá 10 USD/cuốn, song đến cuối năm thì phải bán hạ giá còn 4 USD/cuốn trước khi niên giám thống kê của năm tới được

xuất bản. Cửa hàng muốn xác định số lượng nhập vào sao cho lợi nhuận kỳ vọng là lớn nhất.

Bài toán được giải như sau: Gọi i là số lượng sách cần nhập và j là nhu cầu. Hiển nhiên là lợi nhuận sẽ phụ thuộc vào số lượng sách nhập và nhu cầu thực tế về loại sách đó. Từ đó có thể xây dựng một bảng liệt kê những kết quả khác nhau có thể thu được từ những chiến lược nhập hàng khác nhau. Ta gọi nó là *Bảng lợi nhuận có điều kiện*.

Gọi: R là giá bán 1 cuốn sách

C là giá mua

V là giá bán vào cuối năm.

Lúc đó lợi nhuận có điều kiện được xác định bằng biểu thức:

$$P_{ij} = \begin{cases} R.j - C.i + V(i-j) & \text{với } j \le i \\ R.i - C.i & \text{với } j > i \end{cases} \qquad (2.22)$$

Với các số liệu đã cho ta có:

$$P_{ij} = \begin{cases} 10.j - 7i + 4(i-j) & \text{với } j \le i \\ 10i - 7i = 3i & \text{với } j > i \end{cases}$$

Ta có bảng lợi nhuận có điều kiện sau đây (bảng 2.1)

Bảng 2.1

Nhu cầu	i_j \ p_j	0,3 20	0,25 21	0,18 22	0,14 23	0,1 24	0,03 25
Lượng hàng nhập	20	60	60	60	60	60	60
	21	57	63	63	63	63	63
	22	54	60	66	66	66	66
	23	51	57	63	69	69	69
	24	48	54	60	66	72	72
	25	45	51	57	63	69	75

Chiến lược của cửa hàng là phải chọn số lượng sách cần nhập i để cực đại lợi nhuận kỳ vọng. Với mỗi số lượng nhập i lợi nhuận kỳ vọng được tính bằng công thức:

$$PE_i = \sum_j P_j.P_{ij} \qquad (2.23)$$

Từ đó ta có các giá trị lợi nhuận kỳ vọng như sau tùy thuộc vào số lượng nhập.

Số lượng nhập (i)	Lợi nhuận kỳ vọng PE_i
20	60,00
21	61,20
22	60,90
23	59,52
24	57,30
25	54,48

Vậy chiến lược mang lại lợi nhuận kỳ vọng tối đa là nhập 21 cuốn sách.

3.2. Trung vị

Trung vị, ký hiệu là m_d là giá trị nằm ở chính giữa tập hợp các giá trị có thể có của biến ngẫu nhiên. Nói cách khác đó là giá trị chia phân phối của biến ngẫu nhiên thành hai phần bằng nhau.

Nếu X là biến ngẫu nhiên rời rạc thì giá trị X_i sẽ là trung vị m_d nếu thỏa mãn điều kiện:

$$F(X_i) \le 0,5 < F(X_{i+1}) \qquad (2.24)$$

Còn nếu X là biến ngẫu nhiên liên tục thì trung vị m_d là giá trị thỏa mãn điều kiện:

$$\int_{-\infty}^{m_d} f(x)dx = 0,5 \qquad (2.25)$$

3.3. Mốt

Mốt, ký hiệu là m_0, là giá trị của biến ngẫu nhiên tương ứng với:

1. Xác suất lớn nhất nếu là biến ngẫu nhiên rời rạc

2. Cực đại của hàm mật độ xác suất nếu là biến ngẫu nhiên liên tục.

Trong thực tế có thể gặp biến ngẫu nhiên không có giá trị Mốt hoặc ngược lại nhiều giá trị Mốt cùng một lúc.

Thí dụ 7. Tìm Trung vị và Mốt của biến ngẫu nhiên có bảng phân phối xác suất sau:

X	20	21	22	23	24	25
P	0,3	0,25	0,18	0,14	0,1	0,03

Giải. Để tìm trung vị trước hết ta xây dựng hàm phân bố xác suất của X.

$$F(x) = \begin{cases} 0 \text{ với } x \le 20 \\ 0,3 \text{ với } 20 < x \le 21 \\ 0,55 \text{ với } 21 < x \le 22 \\ 0,73 \text{ với } 22 < x \le 23 \\ 0,87 \text{ với } 23 < x \le 24 \\ 1 \text{ với } x > 25 \end{cases}$$

Từ đó $m_d = 21$.

Dễ thấy rằng $m_0 = 20$.

Thí dụ 8. Thu nhập của dân cư tại một vùng là biến ngẫu nhiên liên tục có hàm phân bố xác suất như sau:

$$F(x) = \begin{cases} 1 - \left(\frac{x_0}{x}\right)^{\alpha} & \text{với } x \geq x_0 \\ 0 & \text{với } x < x_0 \end{cases} \qquad (\alpha > 0)$$

Hãy tìm một mức thu nhập sao cho một nửa số dân của vùng đó có thu nhập cao hơn mức nói trên.

Giải. Mức thu nhập cần tìm chính là m_d.

Ta có:

$$f(x) = \begin{cases} \alpha x_0^{\alpha} x^{-\alpha-1} & \text{với } x \geq x_0 \\ 0 & \text{với } x < x_0 \end{cases}$$

Từ đó:

$$\int_{x_0}^{m_d} f(x)dx = 0{,}5 \rightarrow m_d = x_0 . 2^{1/\alpha}$$

3.4. Phương sai

Trong thực tế nhiều khi chỉ xác định kỳ vọng toán của biến ngẫu nhiên thì chưa đủ để xác định biến ngẫu nhiên đó. Ta còn phải xác định mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó nữa. Chẳng hạn, khi nghiên cứu biến ngẫu nhiên là năng suất lúa của một địa phương nào đó, thì năng suất lúa trung bình (kỳ vọng toán) mới chỉ phản ánh được một khía cạnh của đại lượng đó mà thôi. Mức độ biến động về năng suất của các thửa ruộng khác nhau xung quanh giá trị trung bình cũng là một khía cạnh quan trọng cần nghiên cứu.

Ta có thể nghĩ rằng để đặc trưng cho mức độ phân tán thì đơn giản nhất là tìm tất cả các sai lệch của các giá trị của biến ngẫu nhiên so với kỳ vọng toán của nó và lấy trung bình số học của các sai lệch đó. Song cách làm này không mang lại kết quả vì có thể dễ dàng chứng minh được rằng với mọi biến ngẫu nhiên thì $E[X - E(X)] = 0$. Sở dĩ có điều đó vì các sai lệch dương và các sai lệch âm xung quanh giá trị kỳ vọng toán bao giờ cũng bù trừ cho nhau, do đó giá trị trung bình của các sai lệch sẽ bằng không. Để khắc phục điều đó, người ta không tính trực tiếp trung bình của các sai lệch mà tính trung bình của các giá trị tuyệt đối hoặc bình phương của các sai lệch. Song đơn giản hơn là tìm trung bình của bình phương các sai lệch.

Từ đó ta có khái niệm phương sai.

1. Định nghĩa. *Phương sai của biến ngẫu nhiên X, ký hiệu V(X) là kỳ vọng toán của bình phương sai lệch của biến ngẫu nhiên so với kỳ vọng toán của nó.*

$$V(X) = E[X - E(X)]^2 \qquad (2.26)$$

Như vậy, nếu X là biến ngẫu nhiên rời rạc thì phương sai sẽ được xác định bằng công thức:

$$V(X) = \sum_{i=1}^{n}[x_i - E(X)]^2 p_i \qquad (2.27)$$

còn nếu X là biến ngẫu nhiên liên tục thì phương sai được xác định bằng công thức:

$$V(X) = \int_{x}^{+\infty}[x - E(X)]^2 f(x)dx \qquad (2.28)$$

Trong thực tế việc tính phương sai bằng các công thức

định nghĩa trên có thể gặp khó khăn. Người ta thường tính phương sai bằng công thức sau đây:

$$V(X) = E(X^2) - [E(X)]^2 \qquad (2.29)$$

Thật vậy, theo định nghĩa của phương sai ta có:

$$V(X) = E[X - E(X)]^2 = E\{X^2 - 2XE(X) + [E(X)]^2\} =$$

$$= E(X^2) + E[-2XE(X)] + E[E(X)]^2$$

Song ta biết rằng kỳ vọng toán E(X) là một số xác định, do đó theo tính chất của kỳ vọng toán, ta có:

$$V(X) = E(X^2) - 2E(X).E(X) + [E(X)]^2 = E(X^2) - [E(X)]^2$$

Như vậy, đối với biến ngẫu nhiên rời rạc, ta có thể tính phương sai bằng công thức:

$$V(X) = \sum_{i=1}^{n} x_i^2 p_i - E[(X)]^2 \qquad (2.30)$$

còn đối với biến ngẫu nhiên liên tục:

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - [E(x)]^2 \qquad (2.31)$$

Thí dụ 9. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

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

Tìm phương sai V(X).

Trước hết ta tìm kỳ vọng toán E(X):

$$E(X) = 1.0,1 + 3.0,5 + 4.0,4 = 3,2$$

Ta tìm kỳ vọng toán $E(X^2)$:

$$E(X^2) = 1^2.0,1 + 3^2.0,5 + 4^2.0,4 = 11$$

Như vậy: $V(X) = 11 - (3,2)^2 = 0,76$.

Thí dụ 10. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} 2x & \text{với } x \in (0,1) \\ 0 & \text{với } x \notin (0,1) \end{cases}$$

Tìm phương sai V(X).

Ta tìm kỳ vọng toán E(X)

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^1 x.2x.dx = \frac{2x^3}{3}\Big|_0^1 = \frac{2}{3}$$

Tìm kỳ vọng toán $E(X^2)$:

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx = \int_0^1 x^2.2x.dx = \frac{x^4}{2}\Big|_0^1 = \frac{1}{2}$$

Vậy:

$$V(X) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$

Ta thấy rằng phương sai của biến ngẫu nhiên là một giá trị xác định không âm.

2. *Các tính chất của phương sai*

Tính chất 1. Phương sai của một hằng số bằng không:

$$V(C) = 0 \qquad (2.32)$$

Thật vậy, theo định nghĩa của phương sai:

$$V(C) = E[C - E(C)]^2 = E[C - C]^2 = E(0) = 0$$

Tính chất 2. Phương sai của tích giữa một hằng số và một biến ngẫu nhiên bằng tích giữa bình phương hằng số đó và phương sai của biến ngẫu nhiên ấy.

$$V(CX) = C^2V(X) \tag{2.33}$$

Thật vậy, theo định nghĩa của phương sai

$$V(CX) = E[CX - E(CX)]^2 = E[CX - CE(X)]^2 =$$
$$= EC^2[X - E(X)]^2 = C^2E[X - E(X)]^2 = C^2V(X)$$

Tính chất 3. Phương sai của tổng hai biến ngẫu nhiên độc lập bằng tổng các phương sai thành phần:

$$V(X + Y) = V(X) + V(Y) \tag{2.34}$$

Thật vậy, theo công thức tính phương sai:

$$V(X + Y) = [E(X + Y)^2] - [E(X + Y)]^2 =$$
$$= E[X^2 + 2XY + Y^2] - [E(X) + E(Y)]^2 =$$
$$= E(X^2) + 2E(X).E(Y) + E(Y^2) - [E(X)]^2$$
$$- 2E(X).E(Y) - [E(Y)]^2$$
$$= E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 =$$
$$= V(X) + V(Y)$$

Bằng phương pháp quy nạp có thể chứng minh được hệ quả sau.

Hệ quả 1. Phương sai của tổng n biến ngẫu nhiên độc lập với nhau $X_1, X_2, ..., X_n$ bằng tổng các phương sai thành phần:

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) \tag{2.35}$$

Ngoài ra từ các tính chất nêu trên có thể chứng minh được các hệ quả sau đây:

Hệ quả 2. Phương sai của tổng một hằng số với một biến ngẫu nhiên bằng phương sai của chính biến ngẫu nhiên đó:

$$V(C + X) = V(X) \qquad (2.36)$$

Hệ quả 3. Phương sai của hiệu hai biến ngẫu nhiên độc lập bằng tổng các phương sai thành phần:

$$V(X - Y) = V(X) + V(Y) \qquad (2.37)$$

Một số tính chất khác của phương sai sẽ được đề cập tiếp ở chương IV.

3. Bản chất và ý nghĩa của phương sai

Xuất phát từ định nghĩa của phương sai, ta thấy phương sai chính là trung bình số học của bình phương các sai lệch giữa các giá trị có thể có của biến ngẫu nhiên so với giá trị trung bình của các giá trị đó. Do đó nó phản ánh mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó là kỳ vọng toán.

Thí dụ 1. Tung n con xúc xắc. Tìm phương sai của tổng số điểm thu được.

Giải. Gọi X_i ($i = \overline{1,n}$) là số điểm thu được ở con xúc xắc thứ i. Gọi X là tổng số điểm thu được ở cả n con xúc xắc. Vậy:

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

Vì các X_i ($1 = \overline{1,n}$) độc lập với nhau, do đó theo tính chất của phương sai ta có:

$$V(X) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$$

Ta tìm $V(X_i)$ theo công thức:

$$V(X_i) = E(X_i^2) - [E(X_i)]^2$$

Mỗi biến ngẫu nhiên X_i ($i = \overline{1,n}$) đều có bảng phân phối xác suất như sau:

X_i	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Do đó: $E(X_i) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}$

và $E(X_i^2) = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}$

Do đó: $V(X_i) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$ $(i = \overline{1,n})$

Vì vậy: $V(X) = \sum_{i=1}^{n} V(X_i) = \frac{35}{12}n$

Thí dụ 12. Xác suất để một máy sản xuất ra phế phẩm bằng p. Máy sẽ được sửa chữa ngay sau khi làm ra phế phẩm. Tìm phương sai của số sản phẩm được sản xuất ra giữa 2 lần sửa chữa.

Giải. Ở phần trình bày về kỳ vọng toán ta đã xây dựng được bảng phân phối xác suất của X (số sản phẩm sản xuất ra giữa hai lần sửa chữa) như sau:

X	1	2	3	...	k	...
P	p	qp	q^2p	...	$q^{k-1}p$	...

và đã tìm được $E(X) = \frac{1}{p}$

Ta tìm $E(X^2)$

$$E(X^2) = \sum_{k=1}^{x} k^2 q^{k-1} p = p \sum_{k=1}^{x} k^2 q^{k-1}$$

Ở trên ta đã chứng minh được rằng:

$$\sum_{k=1}^{x} k q^{k-1} = \frac{1}{(1-q)^2}$$

Nhân cả hai vế với q và sau đó lấy đạo hàm theo q, ta có:

$$\frac{d}{dq} \sum_{k=1}^{\infty} k q^k = \sum_{k=1}^{x} k^2 q^{k-1}$$

và $$\left(\frac{q}{(1-q)^2}\right)' = \frac{(1+q)}{(1-q)^3} = \frac{2-p}{p^3}$$

Như vậy: $$\sum_{k=1}^{x} k^2 q^{k-1} = \frac{2-p}{p^3}$$

do đó: $$E(X^2) = \frac{2-p}{p^2}$$

và $$V(X) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

4. Ứng dụng thực tế của phương sai

Cùng với kỳ vọng toán, phương sai có những ứng dụng to lớn trong nhiều lĩnh vực thực tiễn. Nếu như trong kỹ thuật phương sai đặc trưng cho mức độ phân tán của các chi tiết gia công hay sai số của thiết bị thì trong quản lý và kinh doanh nó đặc trưng cho mức độ rủi ro của các quyết định. Ta sẽ minh họa điều đó qua ví dụ sau: Một nhà đầu tư đang cân nhắc giữa việc đầu tư vào hai dự án A và B trong hai lĩnh vực độc lập nhau. Khả năng thu hồi vốn sau 2 năm (tính bằng %)

của hai dự án là các biến ngẫu nhiên có bảng phân phối xác suất như sau:

Dự án A

X_A	65	67	68	69	70	71	73
P	0,04	0,12	0,16	0,28	0,24	0,08	0,08

Dự án B

X_B	66	68	69	70	71
P	0,12	0,28	0,32	0,20	0,08

Từ các bảng phân phối xác suất trên tìm được:

$$E(X_A) = 69,16\%;\ V(X_A) = 3,0944;$$

$$E(X_B) = 68,72\%;\ V(X_B) = 1,8016;$$

Như vậy nếu cần chọn phương án đầu tư sao cho tỷ lệ thu hồi vốn kỳ vọng cao hơn thì nên chọn dự án A, song nếu cần chọn phương án đầu tư sao cho độ rủi ro của tỷ lệ thu hồi vốn thấp hơn tức là khả năng thu hồi vốn ổn định hơn thì lại nên chọn dự án B.

3.5. Độ lệch chuẩn

Độ lệch chuẩn của biến ngẫu nhiên X, ký hiệu σ_x, là căn bậc hai của phương sai:

$$\sigma_x = \sqrt{V(X)} \qquad (2.38)$$

Ta thấy rằng đơn vị đo của phương sai bằng bình phương đơn vị đo của biến ngẫu nhiên. Vì vây khi cần phải đánh giá mức độ phân tán của biến ngẫu nhiên theo đơn vị đo của nó người ta thường tính độ lệch chuẩn chứ không phải là phương

sai vì độ lệch chuẩn có cùng đơn vị đo với biến ngẫu nhiên cần nghiên cứu.

3.6. Hệ số biến thiên

Để đo lường mức độ quan trọng tương đối của sự phân tán của một phân phối người ta thường dùng *hệ số biến thiên* ký hiệu là CV và được xác định bằng biểu thức

$$CV = \left|\frac{\sigma_x}{E(X)}\right| \times 100(\%) \text{ nếu } E(X) \neq 0 \qquad (2.39)$$

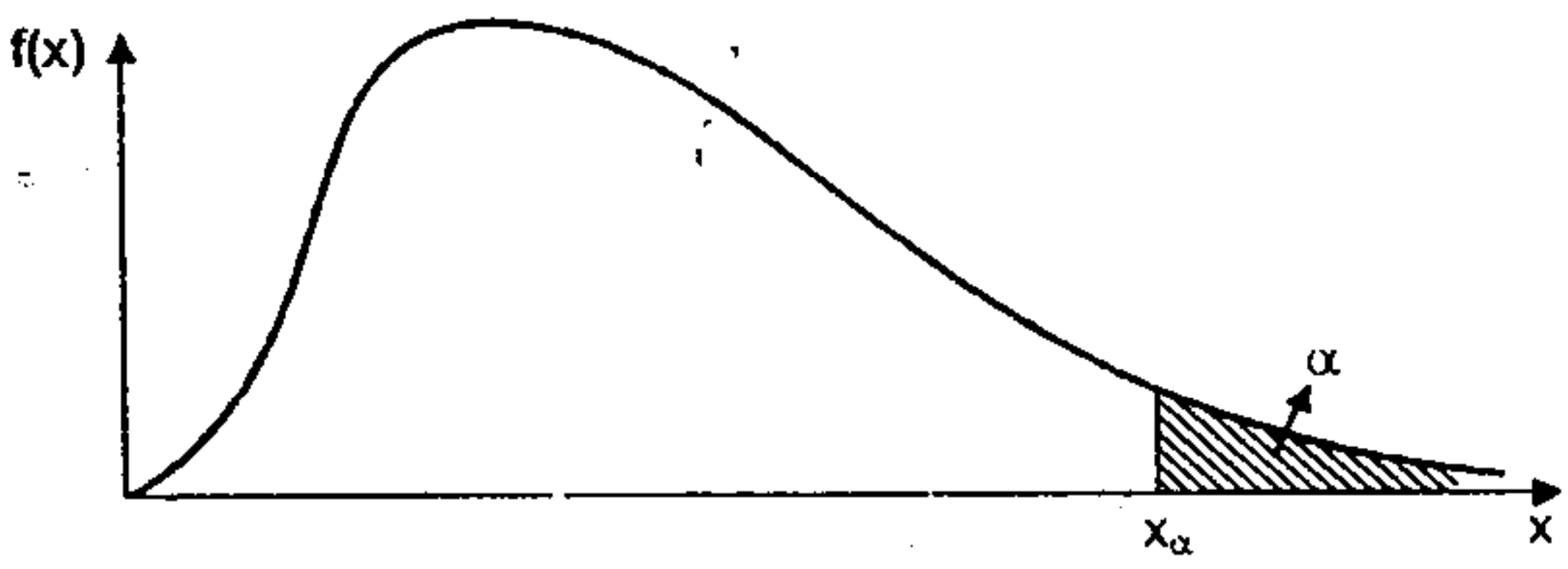
Hệ số biến thiên thường được dùng để đo mức độ thuần nhất của một phân phối. Giá trị của nó càng nhỏ thì mức độ thuần nhất càng lớn. Ngoài ra nó còn dùng để so sánh mức độ phân tán của hai phân phối mà kỳ vọng toán và độ lệch chuẩn của chúng không nhất thiết phải như nhau.

3.7. Giá trị tới hạn

Đối với biến ngẫu nhiên liên tục X trong một số trường hợp người ta còn tìm một loại giá trị gọi là *giá trị tới hạn*. Giá trị tới hạn mức α của biến ngẫu nhiên X, ký hiệu là x_α, là giá trị của X thỏa mãn điều kiện

$$P(X > x_\alpha) = \alpha \qquad (2.40)$$

Trên đồ thị giá trị tới hạn x_α có thể mô tả như sau:



Như vậy giá trị tới hạn x_α là giá trị sao cho diện tích giới hạn bởi trục hoành, đường cong hàm mật độ xác suất và đường thẳng $x = x_\alpha$ bằng α.

3.8. Một vài tham số đặc trưng dạng phân phối xác suất

1. Hệ số bất đối xứng

Mức độ đối xứng của một phân phối có thể quan sát qua đồ thị của nó song để đo lường mức độ bất đối xứng người ta dùng *hệ số bất đối xứng* được xác định bằng biểu thức

$$\alpha_3 = \frac{\mu_3}{\sigma^3} \qquad (2.41)$$

trong đó $\mu_3 = E[X - E(X)]^3$ và σ^3 là lập phương của độ lệch chuẩn.

Tùy thuộc vào giá trị của α_3 mà có thể đưa ra các kết luận sau:

- Nếu $\alpha_3 < 0$ thì phân phối là bất đối xứng và đồ thị sẽ xuôi về bên trái nhiều hơn.

- Nếu $\alpha_3 = 0$ thì phân phối là đối xứng.

- Nếu $\alpha_3 > 0$ thì phân phối là bất đối xứng và đồ thị sẽ xuôi về bên phải nhiều hơn.

Qua so sánh giá trị của trung vị và kỳ vọng toán cũng có thể biết được dấu của hệ số bất đối xứng như sau:

- Nếu $m_d > E(X)$ thì $\alpha_3 < 0$.
- Nếu $m_d = E(X)$ thì $\alpha_3 = 0$.
- Nếu $m_d < E(X)$ thì $\alpha_3 > 0$.

2. Hệ số nhọn

Hệ số nhọn cho phép nhận xét về dạng của một phân phối và bổ sung thêm thông tin về phương sai. Phương sai của biến ngẫu nhiên có thể được xem là nhỏ, lớn hay trung bình. Lúc đó đồ thị của phân phối hoặc sẽ rất tập trung, ít tập trung hay tập trung ở mức bình thường. Trường hợp cuối này thường là mốc để so sánh.

Hệ số nhọn được xác định bằng công thức sau:

$$\alpha_4 = \frac{\mu_4}{\sigma^4} \qquad (2.42)$$

trong đó $\mu_4 = E[X - E(X)]^4$ và σ^4 là bình phương của phương sai.

Khi phân phối xác suất được tập trung ở mức bình thường thì $\alpha_4 = 3$. Nếu phân phối tập trung ở mức độ cao hơn thì $\alpha_4 > 3$ còn ngược lại nếu phân phối tập trung ở mức thấp hơn thì $\alpha_4 < 3$.

Các ký hiệu và công thức cơ bản

- X, Y, Z, X_1, X_2, ..., X_n - Biến ngẫu nhiên
- Bảng phân phối xác suất của biến ngẫu nhiên rời rạc X

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

Điều kiện cơ bản:

$$\begin{cases} p_i \geq 0 & \forall i \\ \sum_{i=1}^{n} p_i = 1 \end{cases}$$

• Hàm phân bố xác suất $F(x) = P(X < x)$

Với biến ngẫu nhiên X rời rạc: $F(x) = \sum_{x_i < x} P_i$

Với biến ngẫu nhiên X liên tục $F(x) = \int_{-\infty}^{x} f(x)dx$

• Hàm mật độ xác suất $f(x) = F'(x)$

Điều kiện cơ bản: $\begin{cases} f(x) \geq 0 & \forall x \\ \int_{-\infty}^{+\infty} f(x)dx = 1 \end{cases}$

• Các công thức tính xác suất

- Nếu X là biến ngẫu nhiên rời rạc

$$P(a \leq X < b) = F(b) - F(a) = \sum_{a < x_i < b} P(X = x_i)$$

- Nếu X là biến ngẫu nhiên liên tục

$$P(a < X < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

• Các tham số đặc trưng cơ bản

+ Kỳ vọng toán:

$$E(X) = \sum_{i=1}^{n} x_i p_i \quad (X \text{ rời rạc})$$

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \quad (X \text{ liên tục})$$

+ Phương sai:

$$V(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

$$V(X) = \sum_{i=1}^{n} x_i^2 p_i - \left(\sum_{i=1}^{n} x_i p_i \right)^2 \quad (X \text{ rời rạc})$$

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - \left(\int_{-\infty}^{+\infty} xf(x)dx \right)^2 \quad (X \text{ liên tục})$$

+ Độ lệch chuẩn: $\sigma_X = \sqrt{V(X)}$

+ Trung vị: m_d

+ Mốt: m_0

+ Hệ số biến thiên:

$$CV_X = \left| \frac{\sigma_X}{E(X)} \right| \cdot 100(\%)$$

+ Mômen trung tâm bậc k:

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

+ Hệ số bất đối xứng: $\alpha_3 = \frac{\mu_3}{\sigma_X^3}$

+ Hệ số nhọn : $\alpha_4 = \frac{\mu_4}{\sigma_X^4}$

+ Giá trị tới hạn mức α: x_α thỏa mãn $P(X > x_\alpha) = \alpha$

Câu hỏi ôn tập

1. Tung một con xúc xắc. Tìm bảng phân phối xác suất của số lần xuất hiện mặt 6 chấm.

2. Trong một hộp có 6 quả cầu trắng, 4 quả cầu đỏ. Lấy ngẫu nhiên 1 quả cầu. Tìm quy luật phân phối xác suất của số cầu đỏ được lấy ra.

3. Một người được phát 3 viên đạn để bắn lần lượt vào

bia cho đến khi trúng với xác suất bắn trúng mỗi viên là 0,8. Tìm quy luật phân phối xác suất của số viên đạn được bắn ra.

4. Cho hàm số

$$f(x) = \begin{cases} 0 \text{ với } x \notin (10,20) \\ \dfrac{1}{20} \text{ với } x \in (10,20) \end{cases}$$

f(x) có phải là hàm mật độ xác suất không?

5. Chứng tỏ rằng hàm số

$$F(x) = \begin{cases} 0 & \text{với } x \le 1 \\ 2x - ?? & \text{với } 1 < x \le 2 \\ 1 & \text{với } x > 2 \end{cases}$$

không phải là hàm phân bố xác suất.

6. Hãy cho biết các mệnh đề sau đây là đúng hay sai? Tại sao?

a. Kỳ vọng toán của tổng một số hữu hạn các biến ngẫu nhiên bằng tổng các kỳ vọng toán thành phần.

b. Kỳ vọng toán của tích một số hữu hạn các biến ngẫu nhiên bằng tích các kỳ vọng toán thành phần.

c. Phương sai của hiệu hai biến ngẫu nhiên bằng hiệu các phương sai thành phần.

7. Cho X và Y là hai biến ngẫu nhiên độc lập có các bảng phân phối xác suất như sau:

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

Y	1	4
P	0,2	0,8

a. Tìm các bảng phân phối xác suất của X + Y và X.Y.

b. Tìm E(X + Y), E(X.Y), V(X + Y) và V(X.Y) bằng tất cả các phương pháp có thể.

8. Cho hai biến ngẫu nhiên X và Y độc lập. Tính V(Z) biết:

a. $Z = 2X + 3Y$

b. $Z = -3X$

9. Biến ngẫu nhiên X có kỳ vọng toán $E(X) = a$ và phương sai $V(X) = b^2$. Tìm kỳ vọng toán và phương sai của biến ngẫu nhiên

$$U = \frac{X - a}{b}$$

10. Cho X là biến ngẫu nhiên. Chứng minh rằng:

$$E[X - E(X)] = 0$$

11. Biến ngẫu nhiên X chỉ nhận hai giá trị có thể có là c và –c với xác suất như nhau. Tìm V(X).

12. Chứng minh rằng:

a. $E(Y) = a.E(X) + b$ và $V(Y) = a^2.V(x)$ nếu $Y = aX + b$

b. $E(Y) = \sum_{i=1}^{n} a_i E(X_i) + b$ và $V(Y) = \sum_{i=1}^{n} a_i^2 V(X_i)$ nếu

$Y = \sum_{i=1}^{n} a_i X_i + b$ X_i là các biến ngẫu nhiên độc lập.

13. Chứng minh rằng với mọi biến ngẫu nhiên X thì

$$E(X^2) \geq [E(X)]^2$$

14. Chứng minh rằng kỳ vọng toán của biến ngẫu nhiên rời rạc nằm trong khoảng giá trị bé nhất và giá trị lớn nhất của biến ngẫu nhiên đó.

15. Biến ngẫu nhiên rời rạc X chỉ nhận hai giá trị có thể

có là x_1 và x_2 với các xác suất tương ứng bằng nhau. Chứng minh rằng

$$V(X) = \frac{(x_2 - x_1)^2}{4}$$

16. Chứng minh rằng phương sai của số lần xuất hiện biến cố trong một phép thử không vượt quá giá trị 1/4.

17. Cho hàm số $f(x) = \frac{a}{1+x^2} \quad -\infty < x < +\infty$

Hãy tìm mọi giá trị của a để hàm số trên không thể là hàm mật độ xác suất của biến ngẫu nhiên.

18. Cho hàm số

$$f(x) = \begin{cases} 0 & \text{với } x \le -1 \\ x+1 & \text{với } -1 < x \le 0 \\ -x+1 & \text{với } 0 < x \le 1 \\ 0 & \text{với } x > 1 \end{cases}$$

a. Chứng minh rằng f(x) là hàm mật độ xác suất.

b. Tìm hàm phân bố xác suất F(x) tương ứng.

c. Tìm E(X) và V(X).

d. Với $k > 0$ chứng tỏ rằng

$$P(|X| > k) \le \frac{1}{6k^2}$$

e. Tìm m_d, m_0, $x_{0,125}$. Vẽ đồ thị của f(x) và mô tả các giá trị nói trên bằng hình vẽ.

19. Trong kinh tế và kinh doanh, ý nghĩa của kỳ vọng toán và phương sai là gì?

20. Tại sao trong thực tế dùng độ lệch chuẩn lại có ý nghĩa hơn dùng phương sai?

Chương III

MỘT SỐ QUY LUẬT PHÂN PHỐI XÁC SUẤT THÔNG DỤNG

Ở chương này ta sẽ nghiên cứu một số quy luật phân phối xác suất thông dụng nhất với các biến ngẫu nhiên rời rạc và liên tục. Điều đó làm cho việc phân loại các biến ngẫu nhiên trong thực tế theo các quy luật phân phối xác suất được dễ dàng hơn.

Để làm rõ những đặc điểm cơ bản của mỗi quy luật phân phối xác suất ta sẽ xuất phát từ các thí dụ có tính điển hình cho mỗi quy luật để làm cơ sở xây dựng những lược đồ khác nhau, từ đó đi đến các quy luật phân phối xác suất tương ứng với mỗi lược đồ.

Giả sử trong bình có N quả cầu trong đó có M quả cầu trắng và N – M quả cầu đen. Mỗi phép thử là việc lấy ngẫu nhiên từ bình ra một quả cầu. Theo những cách lấy khác nhau sẽ dẫn đến những lược đồ khác nhau và các quy luật phân phối xác suất khác nhau.

§1. QUY LUẬT KHÔNG - MỘT - A(P)

Giả sử từ bình lấy ngẫu nhiên một quả cầu. Như vậy trong phép thử này chỉ có 2 biến cố có thể xảy ra: Hoặc lấy được cầu trắng (biến cố A) hoặc lấy được cầu đen (biến cố A không xảy ra tức biến cố $\overline{A}$ xảy ra). Xác suất để A xảy ra (lấy

được cầu trắng) bằng $p = \frac{M}{N}$. Như vậy, xác suất để $\overline{A}$ xảy ra (lấy được cầu đen) bằng $q = \frac{N-M}{N} = 1 - \frac{M}{N} = 1 - p$.

Một cách tổng quát, giả sử ta tiến hành một phép thử, trong đó biến cố A có thể xảy ra với xác suất bằng p. Gọi X là số lần xuất hiện biến cố A trong phép thử đó. Như vậy X là biến ngẫu nhiên rời rạc với hai giá trị có thể có bằng 0 (nếu biến cố không xuất hiện) hoặc bằng 1 (nếu biến cố xuất hiện). Hiển nhiên là xác suất để biến ngẫu nhiên X nhận một trong hai giá trị có thể có nói trên có thể biểu thị bằng công thức:

$$P_x = p^x q^{1-x} \quad \text{với } x = 0; 1 \qquad (3.1)$$

trong đó $q = 1 - p$.

1.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong hai giá trị có thể có X = 0; 1 với các xác suất tương ứng được tính bằng công thức (3.1) gọi là phân phối theo quy luật không - một với tham số là p.

Quy luật không - một được ký hiệu là A(p).

Như vậy, bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật không - một có dạng:

X	0	1	
P	q	p	(q = 1 – p)

1.2. Các tham số đặc trưng của quy luật không - một

Theo bảng phân phối xác suất của X ta có:

$$E(X) = 0.q + 1.p = p$$

Như vậy $E(X) = p$.

Để tìm phương sai, trước hết ta tìm $E(X^2)$:

$$E(X^2) = 0^2.q + 1^2.p = p$$

Từ đó: $$V(X) = p - p^2 = p(1 - p) = pq$$

Như vậy:

$$V(X) = pq$$

Suy ra độ lệch chuẩn

$$\sigma_x = \sqrt{V(X)} = \sqrt{pq}$$

Trong thực tế quy luật không - một thường được dùng để đặc trưng cho các dấu hiệu nghiên cứu định tính có hai phạm trù luân phiên. Chẳng hạn, khi muốn nghiên cứu giới tính của khách hàng ta có thể đặc trưng cho giới tính bằng biến ngẫu nhiên với 2 giá trị bằng 0 (Nam) và bằng 1 (Nữ). Lúc đó, xác suất p sẽ đặc trưng cho tỷ lệ khách hàng nữ trong tập hợp khách hàng. Nếu dấu hiệu định tính có nhiều hơn hai phạm trù thì có thể dùng nhiều biến ngẫu nhiên phân phối "không - một" cùng một lúc.

Về mặt lý thuyết quy luật không - một có thể được dùng làm cơ sở để tìm quy luật phân phối xác suất của các biến ngẫu nhiên khác.

§2. QUY LUẬT NHỊ THỨC - B(n, p)

Ta quay trở lại thí dụ về lô cầu.

Giả sử từ lô cầu gồm M cầu trắng và N – M cầu đen, lấy lần lượt ra n quả cầu theo phương thức hoàn lại. Nếu lấy

theo phương thức này thì n phép thử nói trên sẽ độc lập với nhau, vì việc lấy được cầu trắng hoặc cầu đen trong mỗi lần lấy không ảnh hưởng đến khả năng lấy được cầu trắng hoặc đen trong các lần lấy khác. Trong mỗi lần lấy chỉ có hai trường hợp đối lập xảy ra: Hoặc lấy được cầu trắng (biến cố A), hoặc lấy được cầu đen (biến cố $\overline{A}$). Xác suất lấy được cầu trắng mỗi lần đều bằng $p = \frac{M}{N}$ và xác suất lấy được cầu đen mỗi lần cũng đều bằng $\frac{N-M}{N} = 1 - \frac{M}{N} = 1 - p = q$. Như ở chương I đã trình bày, những điều kiện đó sẽ dẫn đến một lược đồ gọi là lược đồ Bernoulli.

Một cách tổng quát, giả sử ta có một lược đồ Bernoulli, tức là tiến hành n phép thử độc lập, trong mỗi phép thử chỉ có hai trường hợp, hoặc biến cố A xuất hiện, hoặc A không xuất hiện, xác suất xuất hiện biến cố A trong mỗi phép thử đều bằng p như vậy xác suất A không xuất hiện trong mỗi phép thử đều bằng $q = 1 - p$.

Gọi X là "Số lần xuất hiện biến cố A trong n phép thử độc lập" nói trên thì X là biến ngẫu nhiên rời rạc với các giá trị có thể có $X = 0, 1, ..., n$. Như đã chứng minh ở Chương I, xác suất để X nhận các giá trị tương ứng được tính bằng công thức Bernoulli:

$$P_x = C_n^x p^x q^{n-x} \quad x = 0, 1, ..., n \qquad (3.2)$$

2.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có X = 0, 1, ..., n với các xác suất tương ứng được tính bằng công thức (3.2) gọi là phân phối theo quy luật nhị thức với các tham số là n và p.

Quy luật nhị thức được kí hiệu là B(n, p).

Như vậy, bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật nhị thức có dạng:

X	0	1	...	x	...	n
P	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	...	$C_n^x p^x q^{n-x}$	...	$C_n^n p^n q^0$

Trong thực tế đôi khi ta phải tính xác suất để biến ngẫu nhiên X phân phối theo quy luật nhị thức nhận giá trị trong một khoảng [x, x + h] trong đó h là một số nguyên dương ($h \le n - x$). Lúc đó ta có thể tính xác suất này theo công thức:

$$P(x \le X \le x + h) = p_x + p_{x+1} + \ldots + p_{x+h} \qquad (3.3)$$

trong đó mỗi xác suất thành phần được tính bằng công thức (3.2). Thật vậy, biến cố ($x \le X \le x + h$) có thể tách ra thành tổng của h + 1 biến cố xung khắc từng đôi là (X = x), (X = x + 1), ..., (X = x + h), do đó áp dụng định lý cộng xác suất với các biến cố đó ta có:

$$P(x \le X \le x + h) = P(X = x) + P(X = x + 1) + \ldots +$$
$$+ \ldots + P(X = x + h) = p_x + p_{x+1} + \ldots + p_{x+h}$$

Giữa các xác suất P_x và P_{x-1} có mối liên hệ truy chứng sau đây:

$$P_x = \frac{p(n - x + 1)}{qx} P_{x-1} \qquad (3.4)$$

Thật vậy, xét tỷ số:

$$\frac{P_x}{P_{x-1}} = \frac{\dfrac{n!}{x!(n-x)!} p^x q^{n-x}}{\dfrac{n!}{(x-1)!(n-x+1)!} p^{x-1} q^{n-x+1}}$$

Sau khi giản ước ta thu được (3.4). Quan hệ truy chứng này có thể làm cho việc tính toán theo công thức (3.3) được dễ dàng hơn.

Các xác suất theo công thức (3.2) được tính sẵn thành bảng (Phụ lục 1).

Thí dụ 1. Một phân xưởng có 5 máy hoạt động độc lập. Xác suất để trong một ngày mỗi máy bị hỏng đều bằng 0,1. Tìm xác suất để:

a. Trong một ngày có 2 máy hỏng.

b. Trong một ngày có không quá 2 máy hỏng.

Giải. Nếu coi sự hoạt động của mỗi máy là một phép thử, ta có 5 phép thử độc lập. Trong mỗi phép thử chỉ có hai trường hợp: Hoặc máy hỏng, hoặc máy tốt. Xác suất hỏng của mỗi máy đều bằng 0,1. Như vậy, ta có một lược đồ Bernoulli. Gọi X là số máy hỏng trong ngày, X phân phối theo quy luật nhị thức với các tham số $n = 5$ và $p = 0,1$.

Do đó xác suất để trong một ngày có 2 máy hỏng chính là xác suất để $X = 2$.

Theo công thức (3.2) ta có:

$$P_2 = C_5^2 (0,1)^2 . (0,9)^3 = 0,0729$$

Xác suất để trong một ngày có không quá 2 máy hỏng là xác suất để X nhận giá trị trong khoảng [0; 2]. Theo công thức (3.3) ta có:

$$P(0 \le X \le 2) = p_0 + p_1 + p_2$$

$$P_0 = C_5^0 (0,1)^0 . (0,9)^5 = 0,59049$$

$$P_1 = C_5^1 (0,1)^1 . (0,9)^4 = 0,32805$$

Vậy: $P(0 \le X \le 2) = 0,59049 + 0,32805 + 0,0729 = 0,99144$

2.2. Các tham số đặc trưng của quy luật nhị thức

Ta sẽ chứng minh rằng nếu biến ngẫu nhiên X phân phối theo quy luật nhị thức thì kỳ vọng toán:

$$E(X) = np \qquad (3.5)$$

và phương sai:

$$V(X) = npq \qquad (3.6)$$

Thật vậy, gọi X_i ($i = \overline{1,n}$) là số lần xuất hiện biến cố A trong phép thử thứ i. Lúc đó, do các phép thử tiến hành độc lập, các biến ngẫu nhiên X_i độc lập với nhau và mỗi X_i đều phân phối theo quy luật không - một với tham số là p. Như vậy, số lần xuất hiện biến cố A trong n phép thử X bằng:

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

Theo tính chất của kỳ vọng toán và phương sai ta có:

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

và

Vì X_i ($i = \overline{1,n}$) cũng phân phối theo quy luật không - một với tham số là p, do đó:

$$E(X_i) = p \qquad i = \overline{1,n}$$

và $$V(X_i) = pq \qquad i = \overline{1,n}$$

Từ đó:

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

và $$V(X) = \sum_{i=1}^{n} V(X_i) = npq$$

Như vậy, độ lệch chuẩn $\sigma_x = \sqrt{V(X)} = \sqrt{npq}$.

Thí dụ 2. Một nhân viên chào hàng mỗi ngày đi chào hàng ở 10 nơi với xác suất bán được hàng mỗi nơi là 0,2. Vậy nếu một năm người đó đi chào hàng 300 ngày thì trung bình sẽ có khoảng bao nhiêu ngày người đó bán được hàng.

Giải. Dễ thấy xét trong 1 ngày nào đó ta có một lược đồ Bernoulli và gọi X là số lần bán được hàng trong ngày của người đó thì X tuân theo quy luật nhị thức với n = 10 và p = 0,2. Vậy xác suất để người đó bán được hàng trong một ngày bằng:

$$p = P(X \geq 1) = 1 - P(X = 0) = 1 - 0,8^{10} = 0,8926$$

Tương tự nếu xét trong một năm ta cũng có một lược đồ Bernoulli và gọi Y là số ngày người ấy bán được hàng trong một năm thì Y tuân theo quy luật nhị thức với n = 300 và p = 0,8926. Vậy số ngày trung bình trong một năm mà người ấy bán được hàng chính là kỳ vọng toán:

$$E(X) = np = 300.0,8926 = 267,78 \text{ ngày}$$

Ngoài kỳ vọng toán, phương sai và độ lệch chuẩn, trong quy luật nhị thức tham số Mốt cũng hay được dùng.

Nếu X phân phối theo quy luật nhị thức thì Mốt m_0 có thể tìm trực tiếp từ bảng phân phối xác suất bằng cách tìm trong số các giá trị có thể có của X giá trị tương ứng với xác suất lớn nhất. Tuy nhiên, có thể tìm Mốt mà không cần phải xây dựng bảng phân phối xác suất. Nó được xác định bằng công thức sau:

$$np - q \le m_0 \le np + p \qquad (3.7)$$

Thật vậy, vì mốt là giá trị có xác suất lớn nhất trong phân phối, do đó xác suất tương ứng với giá trị m_0 phải không nhỏ hơn xác suất tương ứng với các giá trị cạnh nó là $m_0 - 1$ và $m_0 + 1$. Do đó ta có các bất đẳng thức sau:

$$P_{m_0} \ge P_{m_0-1}$$

và

$$P_{m_0} \ge P_{m_0+1}$$

Thay các biểu thức của xác suất vào các bất đẳng thức trên theo công thức Bernoulli:

$$C_n^{m_0} p^{m_0} q^{n-m_0} \ge C_n^{m_0-1} p^{m_0-1} q^{n-m_0+1}$$

hay:

$$\frac{n!}{m_0!(n-m_0)!} \cdot p^{m_0} q^{n-m_0} \ge \frac{n!}{(m_0-1)!(n-m_0+1)!} \cdot p^{m_0-1} q^{n-m_0+1}$$

Từ đó sau một vài phép biến đổi đơn giản, ta có:

$$m_0 \le np + p$$

Tương tự ta có:

$$C_n^{m_0} p^{m_0} q^{n-m_0} \ge C_n^{m_0+1} p^{m_0+1} q^{n-m_0-1}$$

hay:

$$\frac{n!}{m_0!(n-m_0)!} \cdot p^{m_0} q^{n-m_0} \ge \frac{n!}{(m_0+1)!(n-m_0-1)!} \cdot p^{m_0+1} q^{n-m_0-1}$$

Từ đó ta có: $m_0 \ge np - q$

Kết hợp hai kết quả vừa thu được ta có công thức xác định mốt m_0.

Ta chú ý rằng, vì trong quy luật nhị thức mốt phải là một giá trị nguyên, do đó có thể xảy ra hai trường hợp: Nếu

np + p là một số nguyên thì np – q cũng là một số nguyên, lúc đó mốt sẽ cùng một lúc nhận hai giá trị $m_0 = np + p$ và $m_0 = np - q$. Còn nếu np + p là một số thập phân thì mốt sẽ là giá trị nguyên nằm trong khoảng hai số thập phân là np + p và np – q.

Thí dụ 3. Xác suất để mỗi con lợn khi tiêm phòng bằng một loại văcxin được miễn dịch là 0,9. Có 50 con lợn được tiêm phòng. Tìm số lợn được miễn dịch có khả năng nhiều nhất.

Giải. Bài toán thỏa mãn lược đồ Bernoulli, do đó nếu gọi X là số lợn được miễn dịch thì X là biến ngẫu nhiên phân phối theo quy luật nhị thức.

Vậy số lợn được miễn dịch có khả năng xảy ra nhiều nhất chính là giá trị mốt. Theo công thức mốt ta có:

$$np - q \le m_0 \le np + p$$

với $n = 50,\ p = 0,9,\ q = 1 - 0,9 = 0,1$

$$50.0,9 - 0,1 \le m_0 \le 50.0,9 + 0,9$$

$$44,9 \le m_0 \le 45,9$$

Vậy $m_0 = 45$ tức là số lợn miễn dịch có khả năng nhiều nhất là 45 con.

Như đã trình bày ở trên, mối liên hệ giữa quy luật nhị thức và quy luật không - một được thể hiện như sau: Nếu $X_1\ X_2 \ldots X_n$ là các biến ngẫu nhiên độc lập lẫn nhau và cùng phân phối theo quy luật không - một với tham số là p thì tổng của các biến ngẫu nhiên đó sẽ là biến ngẫu nhiên tuân theo quy luật nhị thức với tham số là n và p.

Mặt khác, nếu X_1 và X_2 là các biến ngẫu nhiên độc lập và

cùng phân phối nhị thức với các tham số tương ứng là n_1, p và n_2, p thì tổng $X = X_1 + X_2$ cũng sẽ phân phối nhị thức với các tham số là $n_1 + n_2$ và p (xem thêm mục 9 chương IV).

2.3. Quy luật phân phối xác suất của tần suất

Trong thực tế nhiều khi người ta quan tâm đến tỷ lệ xuất hiện biến cố A trong lược đồ Bernoulli hơn là bản thân số lần xuất hiện biến cố đó. Để làm điều đó có thể biến đổi biến ngẫu nhiên X thành tần suất xuất hiện biến cố A trong n phép thử độc lập qua phép chia:

$$f = \frac{X}{n}$$

Chú ý rằng việc chia biến ngẫu nhiên cho một hằng số không làm thay đổi phân phối xác suất của biến ngẫu nhiên đó mà chỉ dẫn đến sự thay đổi của các tham số đặc trưng mà thôi. Vì vậy, tần suất f vẫn phân phối theo quy luật nhị thức với tham số là n và p. Lúc đó, bảng phân phối xác suất của f có dạng:

f	0	$\frac{1}{n}$	...	$\frac{x}{n}$	...	1
P	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	...	$C_n^x p^x q^{n-x}$	...	$C_n^n p^n q^0$

Lúc đó, các tham số đặc trưng của biến ngẫu nhiên f như sau:

$$E(f) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np}{n} = p$$

Vậy: $E(f) = p$ (3.8)

Và $$V(f) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{npq}{n^2} = \frac{pq}{n} \quad (3.9)$$

Từ đó:
$$\sigma_f = \frac{\sqrt{pq}}{\sqrt{n}} \qquad (3.10)$$

Quy luật phân phối xác suất của tần suất thường được gọi là quy luật nhị thức theo tỷ lệ.

§3. QUY LUẬT POISSON - P(λ)

Giả sử tiến hành n phép thử độc lập, trong mỗi phép thử xác suất để biến cố A xảy ra đều bằng p và không xảy ra đều bằng q = 1 – p. Lúc đó, nếu gọi X là số lần xuất hiện biến cố A trong n phép thử đó thì X phân phối theo quy luật nhị thức và xác suất để X nhận một trong các giá trị có thể có của nó được tính bằng công thức Bernoulli. Tuy nhiên, nếu số phép thử n quá lớn mà xác suất p lại quá nhỏ thì việc tính toán sẽ gặp nhiều khó khăn. Vì vậy, trong trường hợp này (n lớn, p nhỏ) người ta sử dụng công thức xấp xỉ Poisson.

Như vậy, trong một số rất lớn phép thử độc lập mà xác suất xuất hiện biến cố A trong mỗi phép thử lại rất nhỏ ta phải tìm xác suất để biến cố A xuất hiện đúng x lần.

Giả sử tích np luôn luôn bằng một giá trị không đổi np = λ, lúc đó công thức Bernoulli có thể viết như sau:

$$p_x = \frac{n(n-1)(n-2)...[n-(x-1)]}{x!} p^x q^{n-x}$$

Vì np = λ nên $p = \frac{\lambda}{n}$ do đó:

$$p_x = \frac{n(n-1)(n-2)...[n-(x-1)]}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{x-1}{n}\right)\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Vì n lớn, do đó thay cho p_x ta tìm $\lim_{n\to\infty} p_x$ .. Lúc đó, mỗi thừa số $\left(1-\frac{i}{n}\right)$ với $i = \overline{1, x-1}$ đều tiến tới 1, còn giới hạn

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-x} = e^{-\lambda}$$

do đó:

$$\lim_{n\to\infty} p_x = \frac{\lambda^x}{n!}e^{-\lambda}$$

Như vậy là trong trường hợp số phép thử n rất lớn, xác suất p rất nhỏ và tích $np = \lambda$ không đổi, các xác suất p_x của công thức Bernoulli có thể thay thế bằng công thức xấp xỉ Poisson sau đây:

$$P_x = \frac{\lambda^x}{x!}e^{-\lambda} \quad x = 0, 1, 2, ... \qquad (3.11)$$

Trong thực tế, công thức Poisson có thể dùng thay cho công thức Bernoulli nếu thỏa mãn điều kiện $n \geq 20$ và $p \leq 0,1$.

Một cách tổng quát quy luật Poisson được định nghĩa như sau:

3.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có X = 0, 1, ... với các xác suất tương ứng được tính bằng công thức (3.11) gọi là phân phối theo quy luật Poisson với tham số là λ.

Quy luật Poisson được ký hiệu là $P(\lambda)$.

Như vậy, bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật Poisson có dạng:

X	0	1	...	x	...
P	$e^{-\lambda}.\frac{\lambda^0}{0!}$	$e^{-\lambda}.\frac{\lambda^1}{1!}$	...	$e^{-\lambda}.\frac{\lambda^x}{x!}$	...

Nếu phải tìm xác suất để trong n phép thử biến ngẫu nhiên X phân phối theo quy luật Poisson nhận giá trị trong khoảng [x, x + h] trong đó h là số nguyên dương tùy ý thì xác suất này được tính bằng công thức:

$$P(x \le X \le x + h) = P_x + P_{x+1} + ... + P_{x+h} \qquad (3.12)$$

trong đó mỗi xác suất thành phần được tính bằng công thức (3.11).

Giữa các xác suất P_x và P_{x-1} có mối quan hệ truy chứng sau:

$$P_x = \frac{\lambda}{x} P_{x-1} \qquad (3.13)$$

Thật vậy xét tỷ số:

$$\frac{P_x}{P_{x-1}} = \frac{\frac{e^{-\lambda}\lambda^x}{x!}}{e^{-\lambda}\frac{\lambda^{x-1}}{(x-1)!}} = \frac{\lambda}{x}$$

Từ đó suy ra (3.13).

Các xác suất P_x theo công thức (3.11) được tính sẵn thành bảng (Phụ lục 2).

Thí dụ 1. Một máy dệt có 5000 ống sợi, xác suất để trong

một phút một ống sợi bị đứt bằng 0,0002. Tìm xác suất để trong một phút có không quá 2 ống sợi bị đứt.

Giải. Bài toán thỏa mãn lược đồ Bernoulli song vì n = 5000 rất lớn, p = 0,0002 quá nhỏ và tích np = 5000 × 0,0002 = 1 ≈ npq không đổi, do đó nếu gọi X là số ống sợi bị đứt trong một phút thì X là biến ngẫu nhiên rời rạc và có thể coi như phân phối theo quy luật Poisson. Xác suất để số ống sợi bị đứt trong một phút không quá 2 chính là xác suất để X nhận giá trị trong khoảng [0, 2]. Theo công thức (3.12) ta có:

$$P(0 \le X \le 2) = P_0 + P_1 + P_2$$

$$P_0 = e^{-\lambda}\frac{\lambda^0}{0!} = (2,71)^{-1}\frac{1^0}{0!} = (2,71)^{-1}$$

$$P_1 = \frac{1}{1}P_0 = (2,71)^{-1}$$

$$P_2 = \frac{1}{2}P_1 = \frac{1}{2}(2,71)^{-1}$$

Do đó:

$$P(0 \le X \le 2) = (2,71)^{-1}\left[1+1+\frac{1}{2}\right] = \frac{2,5}{1,71} = 0,9225$$

3.2. Các tham số đặc trưng của quy luật Poisson

Giả sử X phân phối theo quy luật Poisson. Ta sẽ chứng minh rằng:

$$E(X) = \lambda \qquad (3.14)$$

Thật vậy, theo định nghĩa của kỳ vọng toán, ta có:

$$E(X) = \sum_{x=0}^{\infty} xP_x = \sum_{x=0}^{\infty} xe^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}$$

Song ta lại có $\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = e^{\lambda}$ do đó $E(X) = e^{-\lambda}.\ e^{\lambda}.\lambda = \lambda$

Bằng cách tính tương tự có thể tìm được $E(X^2) = \lambda^2 + \lambda$.

Do đó: $$V(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Vậy: $$V(X) = \lambda \qquad (3.15)$$

Như vậy là trong quy luật Poisson cả kỳ vọng toán và phương sai đều bằng λ. Đó là tính chất đặc biệt của quy luật Poisson.

Bằng cách so sánh các xác suất như đã làm trong quy luật nhị thức có thể chứng minh rằng nếu X phân phối theo quy luật Poisson thì mốt được xác định bằng công thức:

$$\lambda - 1 \leq m_0 \leq \lambda$$

Ở đây cũng có thể xảy ra hai trường hợp: Nếu λ là một số nguyên thì mốt sẽ cùng một lúc nhận hai giá trị nguyên là $m_0 = \lambda - 1$ và $m_0 = \lambda$. Còn nếu λ là một số thập phân thì mốt sẽ là giá trị nguyên nằm trong khoảng của hai số thập phân là λ và $\lambda - 1$.

Thí dụ 2. Xác suất để trong khi vận chuyển mỗi chai rượu bị vỡ là 0,001. Người ta tiến hành vận chuyển 2000 chai rượu đến cửa hàng.

a. Tìm số chai vỡ trung bình khi vận chuyển;

b. Tìm số chai vỡ có khả năng nhiều nhất khi vận chuyển.

Giải. Bài toán thỏa mãn lược đồ Bernoulli. Vì n = 2000 khá lớn, p = 0,001 khá nhỏ và tích np = 2000 × 0,001 = 2 không đổi do đó nếu gọi X là số chai rượu bị vỡ khi vận chuyển thì X là biến ngẫu nhiên phân phối theo quy luật

Poisson. Số chai vỡ trung bình chính là kỳ vọng toán của X. Ta có:

$$E(X) = \lambda = 2 \text{ chai}$$

Số chai vỡ có khả năng xảy ra nhiều nhất là giá trị mốt $m_0 \Rightarrow \lambda - 1 \leq m_0 \leq \lambda$. Vì $\lambda = 2$ do đó mốt sẽ nhận hai giá trị là $m_0 = 2$ và $m_0 = 1$. Như vậy số chai vỡ có khả năng nhiều nhất là 1 và 2 chai.

Chú ý rằng nếu X_1 và X_2 là các biến ngẫu nhiên độc lập, X_1 phân phối $P(\lambda_1)$ còn X_2 phân phối $P(\lambda_2)$ thì lúc đó tổng của chúng là biến ngẫu nhiên $X = X_1 + X_2$ cũng sẽ phân phối Poisson với tham số là $\lambda_1 + \lambda_2$ (xem thêm mục 9 chương IV).

Quy luật Poisson có ứng dụng rộng rãi trong nhiều lĩnh vực thực tế như kiểm tra chất lượng sản phẩm, lý thuyết phục vụ công cộng, lý thuyết quản lý dự trữ v.v... Trong lý thuyết phục vụ công cộng, người ta đề cập đến các hệ thống phục vụ dòng các yêu cầu đến như dòng người vào cửa hàng mậu dịch, dòng các con tàu đến cảng chờ bốc xếp, dòng xe ô tô vào một xưởng sửa chữa, dòng khách vào một cửa hàng cắt, sấy tóc v.v...

Trong nhiều trường hợp dòng các yêu cầu đó thường là dòng tối giản tức là thỏa mãn các điều kiện dừng, không hậu quả và đơn nhất. Trong những dòng tối giản như vậy số yêu cầu đến hệ thống trong một khoảng thời gian nào đó thường phân phối theo quy luật Poisson, do đó ta có thể đánh giá được các tính chất của những dòng yêu cầu này, từ đó đưa ra những phương thức tổ chức hệ thống để đạt hiệu quả kinh tế và kỹ thuật cao nhất.

§4. QUY LUẬT SIÊU BỘI - M(N, n)

Trong thực tế ta thường gặp trường hợp tiến hành n phép thử và trong mỗi phép thử cũng chỉ có hai trường hợp: hoặc A xảy ra, hoặc A không xảy ra. Tuy nhiên, các phép thử này không tiến hành một cách độc lập với nhau. Do đó, xác suất để biến cố A xảy ra hoặc không xảy ra trong mỗi phép thử sẽ không bằng nhau nữa mà thay đổi từ phép thử này qua phép thử khác. Trong những lược đồ như vậy số lần xuất hiện biến cố A trong n phép thử sẽ không phân phối theo quy luật nhị thức hoặc quy luật Poisson nữa, vì vậy không thể dùng công thức Bernoulli hoặc công thức Poisson để tìm xác suất để biến cố A xuất hiện x lần trong n phép thử nói trên. Một trong những lược đồ như vậy dẫn đến quy luật siêu bội.

Ta trở lại thí dụ về lô cầu. Giả sử trong bình có N quả cầu trong đó có M quả cầu trắng và N – M quả cầu đen. Lấy ngẫu nhiên lần lượt ra n quả cầu theo phương thức không hoàn lại. Lúc đó các phép thử sẽ không độc lập với nhau nữa và xác suất để lấy được cầu trắng ở mỗi lần lấy sẽ thay đổi. Theo định nghĩa cổ điển về xác suất, xác suất để trong n quả cầu lấy ra có x quả cầu trắng được tính bằng công thức:

$$P_x = \frac{C_M^x \cdot C_{N-M}^{n-x}}{C_N^n} \qquad (3.16)$$

trong đó x có thể bằng 0, 1, ..., n. Từ đó ta có định nghĩa sau đây:

4.1. Định nghĩa

Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có X = 0, 1, 2, ..., n với các xác suất tương ứng được tính bằng công thức (3.16) gọi là phân phối theo quy luật siêu bội với các tham số là N và n.

Quy luật siêu bội được ký hiệu là M(N, n).

Bảng phân phối xác suất của biến ngẫu nhiên X phân phối theo quy luật siêu bội có dạng:

X	0	1	...	x	...	n
P	$\frac{C_M^0 . C_{N-M}^n}{C_N^n}$	$\frac{C_M^1 . C_{N-M}^{n-1}}{C_N^n}$	...	$\frac{C_M^x . C_{N-M}^{n-x}}{C_N^n}$	...	$\frac{C_M^n . C_{N-M}^0}{C_N^n}$

Thí dụ. Trong cửa hàng có bán 100 bóng đèn trong đó có lẫn 5 bóng hỏng mà không kiểm tra thì không thể xác định được. Một người khách chọn ngẫu nhiên 2 bóng. Tìm xác suất để người đó mua được cả 2 bóng đều tốt.

Giải. Gọi X là số bóng tốt mà người đó có thể mua được. X phân phối theo quy luật siêu bội với N = 100, M = 95 và n = 2. Xác suất để mua được cả 2 bóng tốt là xác suất để X = 2. Theo công thức (3.16) ta có:

$$P_2 = \frac{C_{95}^2 \cdot C_5^0}{C_{100}^2} = \frac{95.94}{100.99} \approx 0,9$$

4.2. Các tham số đặc trưng của quy luật siêu bội

Có thể chứng minh được rằng nếu X phân phối theo quy luật siêu bội thì kỳ vọng toán:

$$E(X) = n\frac{M}{N} = np \qquad (3.17)$$

và phương sai:

$$V(X) = n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1} = npq \cdot \frac{N-n}{N-1} \qquad (3.18)$$

Giá trị $\frac{N-n}{N-1}$ được gọi là *hệ số hiệu chỉnh.*

Trong thực tế quy luật siêu bội được áp dụng để tính xác suất xuất hiện x lần biến cố khi lấy một cách ngẫu nhiên n đơn vị từ một tập hợp nào đó theo phương thức *không hoàn lại.* Chẳng hạn, để kiểm tra chất lượng của một lô sản phẩm được sản xuất ra người ta thường lấy ngẫu nhiên từ lô đó n sản phẩm theo phương thức không hoàn lại và đánh giá xác suất để trong đó có x phế phẩm hoặc chính phẩm.

Ta chú ý rằng khi số phép thử n là rất bé so với số N thì phân phối siêu bội thực tế không khác biệt so với phân phối nhị thức vì:

$$\lim_{\frac{n}{N} \to 0} \frac{C_M^x \cdot C_{N-M}^{n-x}}{C_N^n} = C_n^x \cdot \left(\frac{M}{N}\right)^x \cdot \left(1 - \frac{M}{N}\right)^{n-x}$$

Điều đó có nghĩa là khi tỷ số $\frac{n}{N}$ là rất nhỏ (tức là $\frac{n}{N} < 0,1$) thì phương pháp lấy không hoàn lại và có hoàn lại gần như không khác nhau. Do đó có thể lấy theo phương thức không hoàn lại song vẫn có thể tính toán như trong trường hợp lấy có hoàn lại cho đơn giản.

Trên đây ta xét một số quy luật phân phối xác suất thông dụng nhất của các *biến ngẫu nhiên rời rạc*. Sau đây ta sẽ xét một số quy luật phân phối xác suất cơ bản của các *biến ngẫu nhiên liên tục* vì nhiều đại lượng cần nghiên cứu trong

thực tế chính là các biến ngẫu nhiên liên tục. Do đó, việc hiểu biết các quy luật phân phối xác suất của chúng cho phép tiến hành phân tích một cách sâu sắc, cụ thể và chính xác hơn các hiện tượng này.

§5. QUY LUẬT PHÂN PHỐI ĐỀU - U(a, b)

Phân phối đều là quy luật xác suất đơn giản nhất trong các quy luật phân phối xác suất của biến ngẫu nhiên liên tục. Nếu biến ngẫu nhiên X có thể nhận bất kỳ giá trị nào trên khoảng (a, b) với a và b là các số thực và ứng với mỗi giá trị là một mật độ xác suất như nhau thì biến X sẽ có phân phối đều.

Như vậy, trong khoảng (a, b) hàm mật độ xác suất của biến ngẫu nhiên phải bằng một giá trị xác định tức là $f(x) = c$ trong khoảng (a, b). Từ đó, theo tính chất của hàm mật độ xác suất ta có:

$$\int_a^b f(x)dx = \int_a^b cdx = 1$$

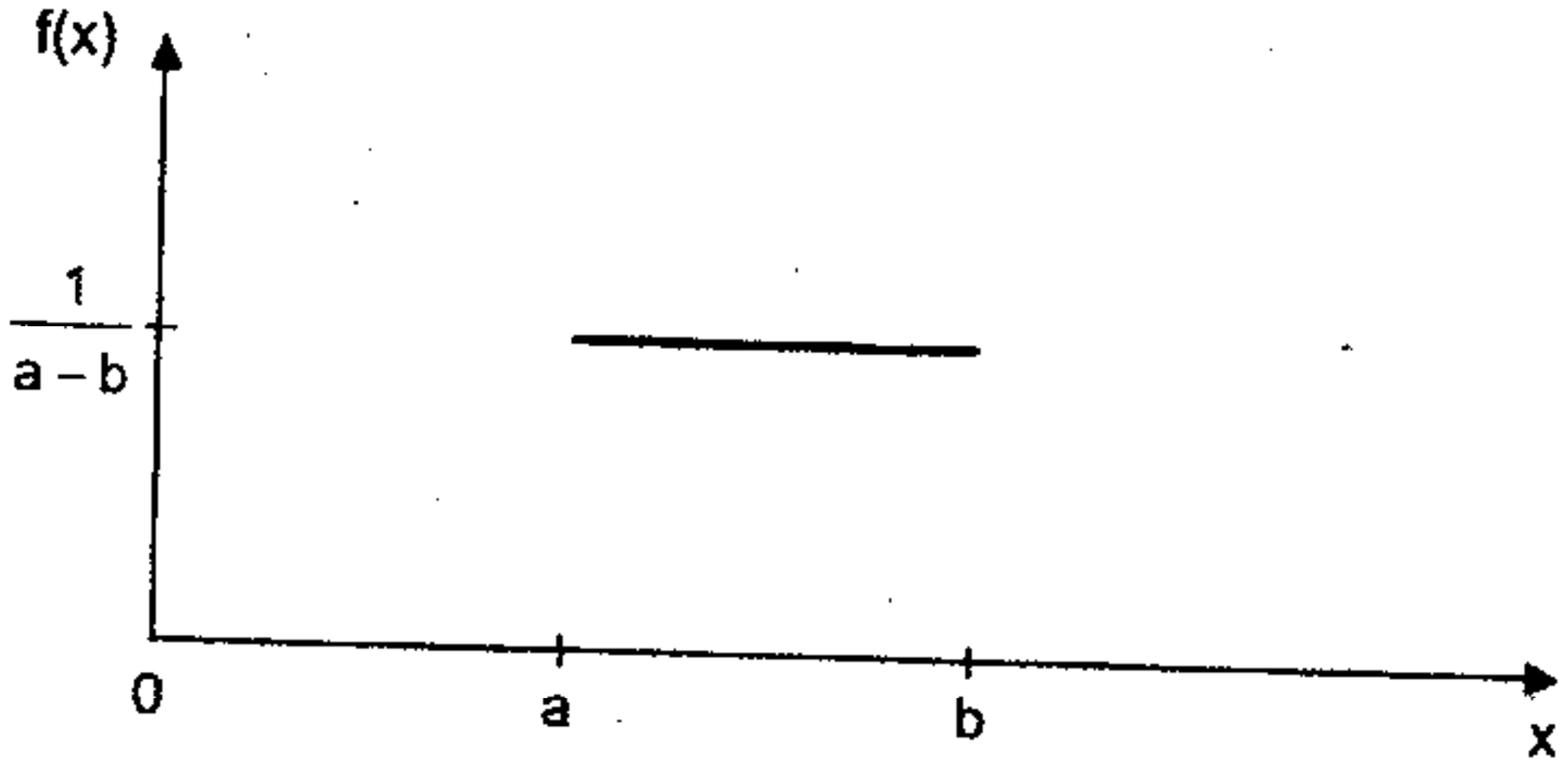
Từ đó: $cb - ca = 1$ suy ra $c = \frac{1}{b-a}$. Như vậy, ta có định nghĩa sau:

5.1. Định nghĩa

Biến ngẫu nhiên liên tục X gọi là phân phối theo quy luật đều trong khoảng (a, b) nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{với } x \in (a,b) \\ 0 & \text{với } x \notin (a,b) \end{cases} \qquad (3.19)$$

Đồ thị hàm f(x) có dạng như ở hình 3.1.



Hình 3.1. **Đồ thị hàm f(x) của quy luật phân phối đều**

5.2. Các tham số đặc trưng của quy luật phân phối đều

Giả sử X phân phối đều, lúc đó theo định nghĩa kỳ vọng toán của biến ngẫu nhiên liên tục, ta có:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_a^b x\frac{1}{b-a}dx =$$

$$= \frac{1}{b-a}\cdot\frac{x^2}{2}\Big|_b^a = \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} = \frac{a+b}{2}$$

Như vậy: $$E(X) = \frac{a+b}{2}. \qquad (3.20)$$

Để tìm phương sai ta tìm $E(X^2)$:

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^3}{3}\Big|_a^b =$$

$$= \frac{1}{b-a} \cdot \frac{b^3 - a^3}{3} = \frac{b^2 + ab + a^2}{3}$$

Do đó:

$$V(X) = E(X^2) - [E(X)]^2 = \frac{(b-a)^2}{12}$$

Như vậy:
$$V(X) = \frac{(b-a)^2}{12} \qquad (3.21)$$

Quy luật phân phối đều có ứng dụng rộng trong thống kê toán. Nó có ý nghĩa to lớn trong các phương pháp phi tham số. Khái niệm phân phối đều đôi khi còn được sử dụng trong lý thuyết các ước lượng thống kê. Trong một số lý thuyết kết luận thống kê người ta thường xuất phát từ quy tắc sau đây: *Nếu ta không biết gì về giá trị của tham số cần ước lượng thì mỗi giá trị có thể có của tham số đó là đồng khả năng*. Điều đó dẫn đến việc quan niệm tham số cần ước lượng như một biến ngẫu nhiên tuân theo quy luật phân phối đều.

Thí dụ: Khi thâm nhập vào một thị trường mới, doanh nghiệp không thể khẳng định được một cách chắc chắn doanh số hàng tháng có thể đạt được sẽ là bao nhiêu mà chỉ dự kiến được rằng doanh số tối thiểu sẽ là 20 triệu đồng/tháng và tối đa là 40 triệu đồng/tháng. Tìm xác suất để doanh nghiệp đạt được doanh số tối thiểu là 35 triệu đồng/tháng.

Giải. Gọi X là doanh số hàng tháng mà doanh nghiệp có thể đạt được ở thị trường đó. Do không có thông tin gì hơn

nên có thể xem X là biến ngẫu nhiên liên tục phân phối đều trên khoảng (20; 40).

Vậy X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \dfrac{1}{40-20} = 0{,}05 & \text{với } x \in (20;\ 40) \\ 0 & \text{với } x \notin (20;\ 40) \end{cases}$$

Từ đó xác suất để doanh nghiệp đạt được doanh số tối thiểu là 35 triệu đồng/tháng được tìm theo tính chất của hàm mật độ xác suất như sau:

$$P(X > 35) = \int_{35}^{+\infty} f(x)dx = \int_{35}^{40} 0{,}05dx = 0{,}05x\Big|_{35}^{40} = 0{,}25$$

§6. QUY LUẬT PHÂN PHỐI LŨY THỪA - E(λ)

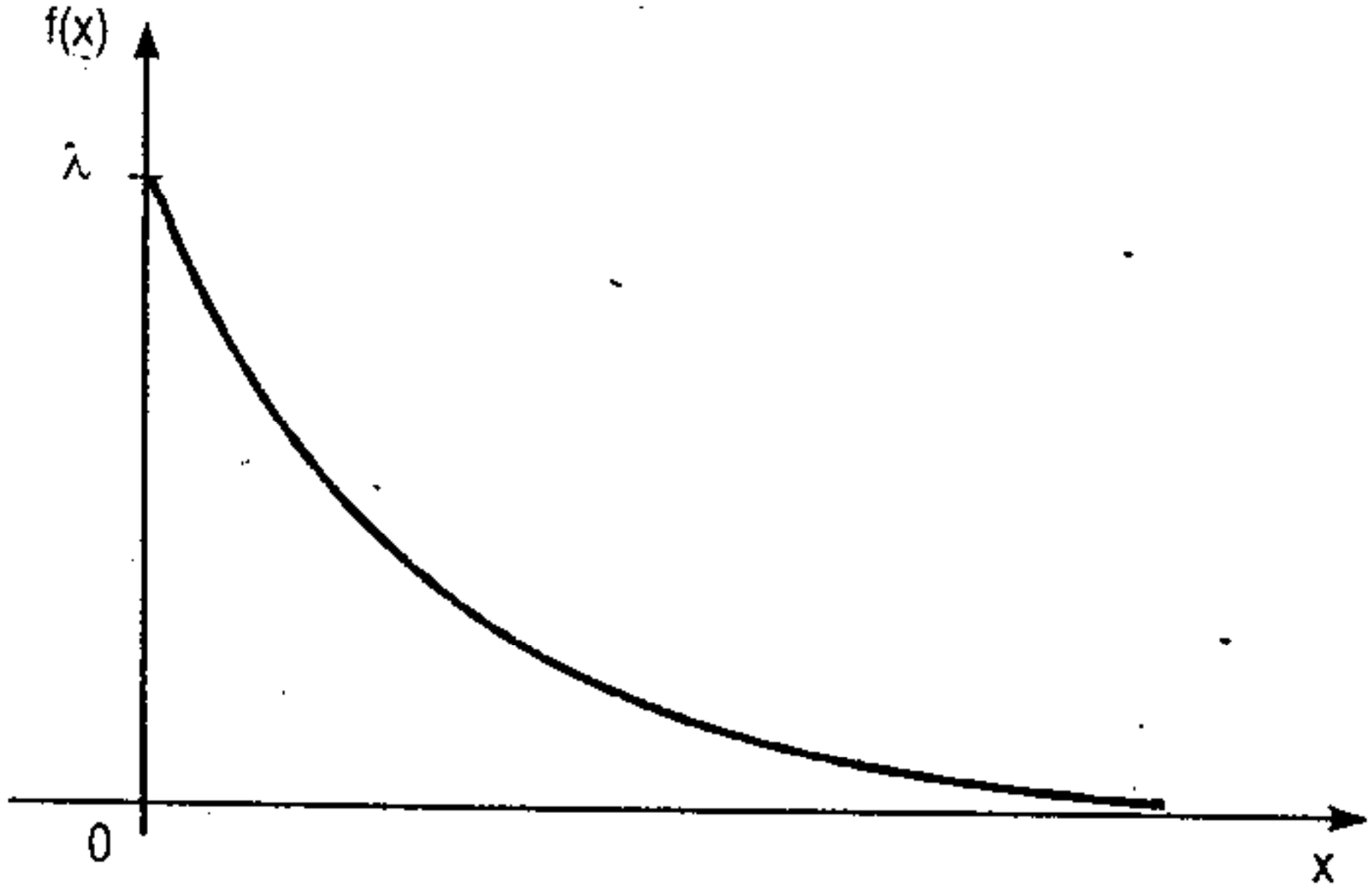
6.1. Định nghĩa

Biến ngẫu nhiên liên tục X gọi là phân phối theo quy luật lũy thừa (quy luật mũ) nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \begin{cases} 0 & \text{với } x < 0 \\ \lambda e^{-\lambda x} & \text{với } x \geq 0 \end{cases} \qquad (3.22)$$

Trong đó λ là một hằng số dương.

Đồ thị của hàm f(x) có dạng như ở hình 3.2.



Hình 3.2. **Đồ thị hàm f(x) của quy luật lũy thừa**

Ta tìm hàm phân bố xác suất của quy luật lũy thừa. Theo tính chất của hàm mật độ xác suất, ta có:

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} \lambda e^{-\lambda x}dx = 1 - e^{-\lambda x} \qquad (3.23)$$

6.2. Các tham số đặc trưng của quy luật lũy thừa

Giả sử biến ngẫu nhiên X phân phối theo quy luật lũy thừa. Lúc đó kỳ vọng toán

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{0}^{+\infty} x\lambda e^{-\lambda x}dx$$

Lấy tích phân từng phần, ta được:

$$E(X) = \frac{1}{\lambda} \qquad (3.24)$$

Còn phương sai:

$$V(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - [E(X)]^2 = \int_{0}^{\infty} \lambda x^2 e^{-\lambda x} dx - \frac{1}{\lambda^2}$$

Lấy tích phân từng phần ta được:

$$\lambda \int_{0}^{\infty} x^2 e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

Do đó:
$$V(X) = \frac{1}{\lambda^2} \qquad (3.25)$$

Suy ra độ lệch chuẩn:

$$\sigma_x = \sqrt{V(X)} = \frac{1}{\lambda} \qquad (3.26)$$

Như vậy trong quy luật lũy thừa kỳ vọng toán và độ lệch chuẩn đều bằng $\frac{1}{\lambda}$. Đây chính là tính chất đặc biệt của quy luật lũy thừa. Nó có thể được sử dụng để kiểm tra xem một biến ngẫu nhiên mà ta nghiên cứu trong thực tế có phân phối theo quy luật lũy thừa hay không.

Ta tìm xác suất để biến ngẫu nhiên X phân phối theo quy luật lũy thừa nhận giá trị trong khoảng (a, b). Theo tính chất của hàm phân bố xác suất:

$$P(a \le X < b) = F(b) - F(a) = [1 - e^{-b\lambda}] - [1 - e^{-a\lambda}]$$

Vậy:

$$P(a < X < b) = e^{-a\lambda} - e^{-b\lambda} \qquad (3.27)$$

Giá trị của hàm e^{-x} được tính sẵn thành bảng (Phụ lục 3).

Thí dụ. Biến ngẫu nhiên liên tục X phân phối theo quy luật lũy thừa với hàm mật độ xác suất.

$$f(x) = \begin{cases} 2e^{-2x} & \text{với } x \geq 0 \\ 0 & \text{với } x < 0 \end{cases}$$

a. Viết hàm F(x)

b. Tìm xác suất để trong kết quả của phép thử X nhận giá trị trong khoảng [0,3; 1].

c. Tìm kỳ vọng toán và phương sai của X.

Giải. a. Hàm phân bố xác suất có dạng:

$$F(x) = \begin{cases} 1 - e^{-2x} & \text{với } x \geq 0 \\ 0 & \text{với } x < 0 \end{cases}$$

b. Theo công thức (3.27) ta có:

$$P(0,3 < X < 1) = e^{-2.0,3} - e^{-2.1} = e^{-0,6} - e^{-2} =$$

$$= 0,54881 - 0,13534 \approx 0,41$$

c. Theo công thức tính kỳ vọng toán và phương sai:

$$E(X) = \frac{1}{\lambda} = \frac{1}{2} = 0,5$$

$$V(X) = \frac{1}{\lambda^2} = \frac{1}{4} = 0,25$$

$$\sigma_X = \frac{1}{\lambda} = \frac{1}{2} = 0,5$$

Quy luật phân phối lũy thừa có ứng dụng trong nhiều lĩnh vực khác nhau. Người ta chứng minh được rằng thời gian giữa hai lần xuất hiện yêu cầu của một dòng yêu cầu tối giản trong các hệ thống phục vụ công cộng phân phối theo quy luật lũy thừa. Trong các hệ thống kỹ thuật, thời gian làm việc liên tục của máy móc thiết bị giữa hai lần sửa chữa cũng thường phân phối theo quy luật lũy thừa.

Khi áp dụng quy luật lũy thừa để giải quyết các bài toán nảy sinh trong thực tế, ngoài ưu điểm là đơn giản (nó chỉ phụ thuộc vào một tham số là λ) nó còn có một tính chất rất quan trọng sau đây: *Xác suất hoạt động liên tục của thiết bị trong khoảng thời gian t không phụ thuộc vào quãng thời gian hoạt động trước đó mà chỉ phụ thuộc vào độ dài của khoảng thời gian t mà thôi.*

Thật vậy, gọi A là biến cố thiết bị hoạt động tốt trong khoảng thời gian $(0, t_0)$, gọi B là biến cố thiết bị hoạt động tốt trong khoảng thời gian $(t_0, t_0 + t)$. Lúc đó tích AB là biến cố thiết bị hoạt động tốt trong khoảng thời gian $(0, t_0 + t)$ có độ dài $t_0 + t$, lúc đó P(A) là xác suất để thời gian hoạt động tốt T của thiết bị không nhỏ hơn t_0 và P(B) là xác suất để thời gian hoạt động tốt của thiết bị không nhỏ hơn t.

$$P(A) = P(T \geq t_0) = 1 - P(T < t_0) = 1 - F(t_0)$$

$$P(B) = P(T \geq t) = 1 - P(T < t) = 1 - F(t)$$

Vì T phân phối theo quy luật lũy thừa do đó theo công thức (3.23) ta có:

$$P(A) = 1 - [1 - e^{-\lambda t_0}] = e^{-\lambda t_0}$$

$$P(B) = 1 - [1 - e^{-\lambda t}] = e^{-\lambda t}$$

$$P(AB) = P(T \geq t_0 + t) = 1 - P(T < t_0 + t) =$$

$$= 1 - F(t_0 + t) = 1 - [1 - e^{-\lambda(t_0 + t)}]$$

Vậy $P(AB) = e^{-\lambda(t_0 + t)}$

Bây giờ ta tìm xác suất để thiết bị sẽ hoạt động tốt trong khoảng $(t_0, t_0 + t)$ với điều kiện nó đã hoạt động tốt trong khoảng $(0, t_0)$. Từ định lý nhân xác suất ta có:

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{e^{-\lambda t_0} \cdot e^{-\lambda t}}{e^{-\lambda t_0}} = e^{-\lambda t}$$

Như vậy, P(B/A) = P(B) tức là A và B độc lập với nhau và trong biểu thức của P(B) chỉ có tham số t chứ không có t_0. Điều đó chứng tỏ xác suất để thiết bị hoạt động tốt trong khoảng thời gian t sẽ không phụ thuộc gì vào khoảng thời gian hoạt động tốt đã diễn ra trước đó, mà chỉ phụ thuộc vào độ dài của khoảng thời gian đang xét mà thôi.

Người ta đã chứng minh được rằng chỉ có quy luật phân phối lũy thừa mới có tính chất trên. Đó cũng là tiêu chuẩn để nhận biết quy luật này trong thực tế.

§7. QUY LUẬT PHÂN PHỐI CHUẨN - $N(\mu, \sigma^2)$

7.1. Định nghĩa

Biến ngẫu nhiên liên tục X nhận các giá trị trong khoảng $(-\infty, +\infty)$ gọi là phân phối theo quy luật chuẩn với các tham số μ và σ^2, nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (3.28)$$

Nếu tiến hành khảo sát hàm số trên và vẽ đồ thị của nó ta sẽ thu được các kết luận sau đây:

a. Hàm số xác định trên toàn trục Ox.

b. Với mọi giá trị của x hàm số luôn luôn dương, như vậy, đồ thị của nó luôn nằm cao hơn trục Ox.

c. Khi $x \to \pm\infty$ thì $f(x) \to 0$ tức là trục Ox là đường tiệm cận ngang.

d. Ta tìm đạo hàm bậc nhất

$$f'(x) = -\frac{x-\mu}{\sigma^3\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Dễ dàng thấy rằng $f'(x) = 0$ khi $x = \mu$; $f'(x) > 0$ khi $x < \mu$, $f'(x) < 0$ khi $x > \mu$.

Như vậy khi $x = \mu$ hàm số có cực đại bằng $\frac{1}{\sigma\sqrt{2\pi}}$

e. Hiệu $x - \mu$ trong biểu thức của hàm f(x) nằm trong dạng bình phương, tức là hàm số đối xứng qua đường thẳng $x = \mu$.

g. Ta tìm điểm uốn của hàm. Đạo hàm bậc hai

$$f''(x) = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$$

Dễ dàng thấy rằng khi $x = \mu + \sigma$ và $x = \mu - \sigma$ đạo hàm bậc hai bằng không và đi qua hai điểm đó nó đổi dấu (tại cả hai điểm đó hàm số đều bằng $\frac{1}{\sigma\sqrt{2\pi e}}$).

Như vậy các điểm:

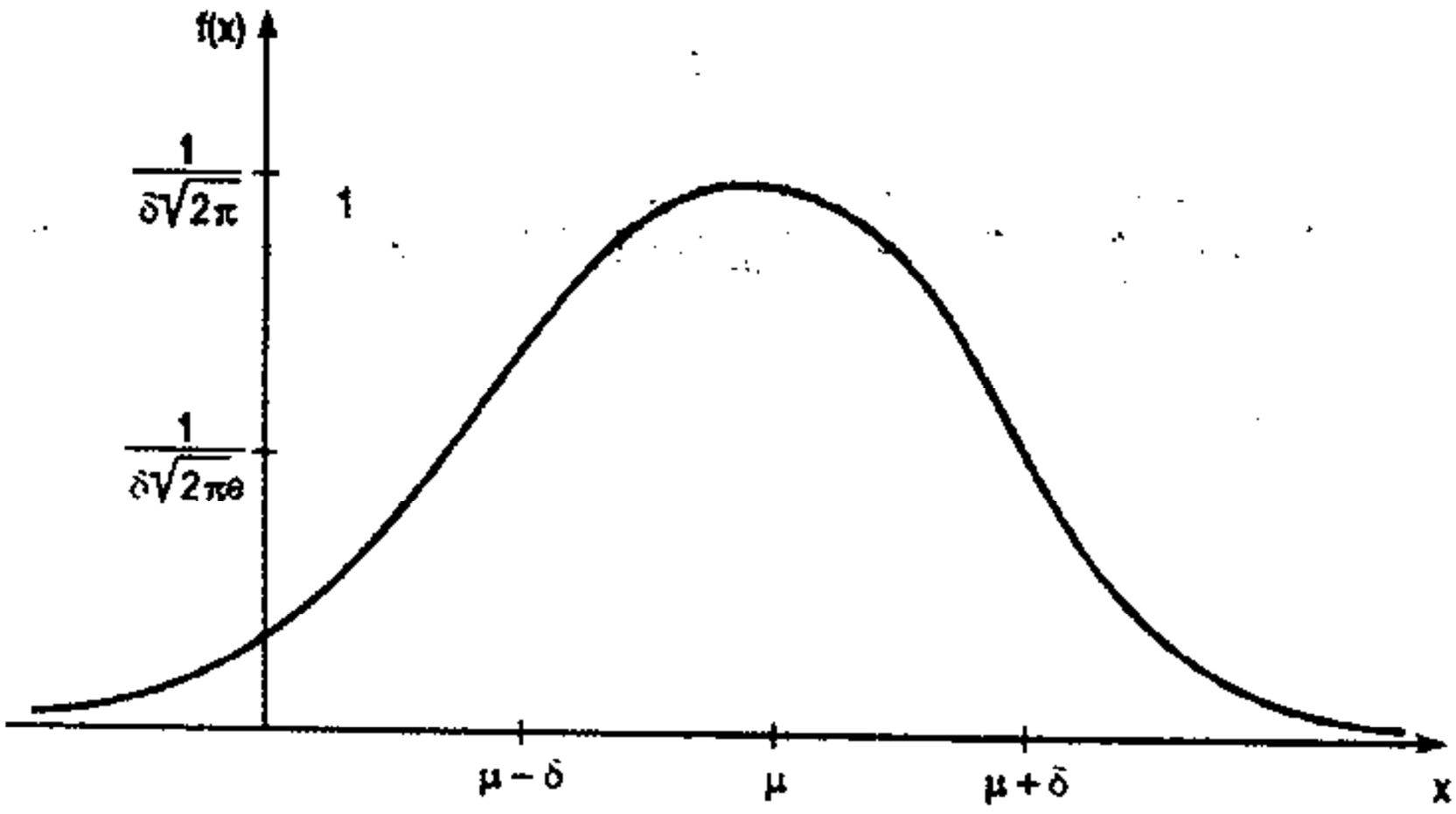
$$\left(\mu - \sigma; \frac{1}{\sigma\sqrt{2\pi e}}\right) \text{ và } \left(\mu + \sigma; \frac{1}{\sigma\sqrt{2\pi e}}\right)$$

là các điểm uốn.

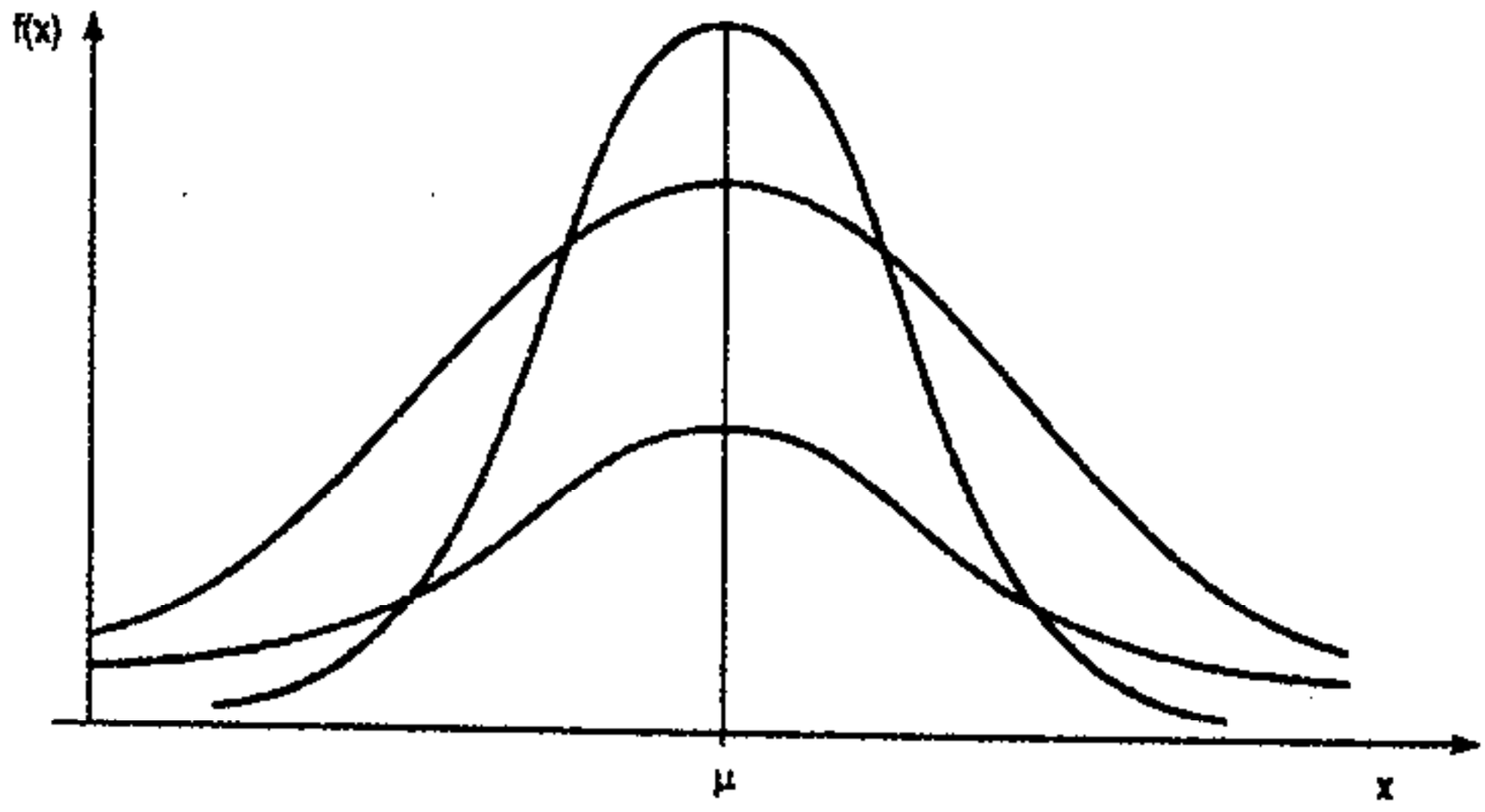
Vậy đồ thị của hàm mật độ xác suất của phân phối chuẩn có dạng như sau (hình 3.3) (Xem trang sau).

Hai tham số μ và σ có ý nghĩa rất quan trọng trong phân phối chuẩn (bản chất của nó sẽ được trình bày về sau). Khi μ và σ thay đổi, dạng đồ thị của hàm mật độ xác suất f(x) cũng

thay đổi như sau: Khi μ thay đổi thì dạng của đường cong f(x) không thay đổi song nó sẽ chuyển dịch sang phải hoặc sang trái theo trục Ox. Khi μ tăng lên đồ thị sẽ dịch sang phải, còn khi μ giảm đồ thị sẽ dịch sang trái. Trái lại, khi σ thay đổi thì dạng của đồ thị sẽ thay đổi theo. Nếu σ tăng lên thì đồ thị sẽ thấp xuống và phình ra, còn khi σ giảm đi thì đồ thị sẽ cao lên và nhọn thêm. Trên hình 3.4 ta minh họa đồ thị của f(x) với ba giá trị khác nhau của σ.



Hình 3.3. **Đồ thị hàm f(x) phân phối chuẩn**



Hình 3.4. **Sự thay đổi của f(x) theo σ**

Theo tính chất của hàm mật độ xác suất, ta có hàm phân bố xác suất của biến ngẫu nhiên X phân phối theo quy luật chuẩn được xác định bằng biểu thức:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

7.2. Các tham số đặc trưng của quy luật chuẩn

Ta sẽ chứng minh rằng trong quy luật chuẩn thì μ chính là kỳ vọng toán còn σ chính là độ lệch chuẩn của X. Thật vậy, theo định nghĩa kỳ vọng toán của biến ngẫu nhiên liên tục, ta có:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Ta thực hiện phép đổi biến số: $Z = \frac{x-\mu}{\sigma}$

Từ đó $x = \sigma Z + \mu$, $dx = \sigma dZ$. Chú ý rằng khi đổi biến các cận lấy tích phân không thay đổi, ta có:

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma Z + \mu) e^{-\frac{Z^2}{2}} dZ =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sigma Z e^{-\frac{Z^2}{2}} dZ + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{Z^2}{2}} dZ$$

Tích phân thứ nhất bằng không do hàm dưới dấu tích phân là hàm lẻ mà cận lấy tích phân lại đối xứng qua gốc tọa độ. Còn tích phân thứ hai bằng:

$$\int_{-\infty}^{+\infty} e^{-\frac{Z^2}{2}} \cdot Z = \sqrt{2\pi} \text{ (tích phân Poisson)}$$

Do đó: $E(X) = \mu$ (3.29)

Theo định nghĩa phương sai của biến ngẫu nhiên liên tục và do $E(X) = \mu$ ta có:

$$V(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Ta thực hiện phép đổi biến số $Z = \frac{x-\mu}{\sigma}$, từ đó $x - \mu = Z\sigma$, $dx = \sigma dZ$. Chú ý rằng cận lấy tích phân không thay đổi, ta có:

$$V(X) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} Z^2 e^{-\frac{Z^2}{2}} dZ$$

lấy tích phân từng phần bằng cách đặt $u = Z$, $dv = Ze^{-\frac{Z^2}{2}} dZ$ ta tìm được $V(X) = \sigma^2$

Do đó: $\sigma_x = \sqrt{V(X)} = \sigma$ (3.30)

Như vậy, kỳ vọng toán của biến ngẫu nhiên X phân phối chuẩn là $E(X) = \mu$ và phương sai là: $V(x) = \sigma^2$. Phân phối chuẩn được ký hiệu là $N(\mu, \sigma^2)$.

Có liên quan mật thiết với phân phối chuẩn là một phân phối khác gọi là phân phối chuẩn hóa.

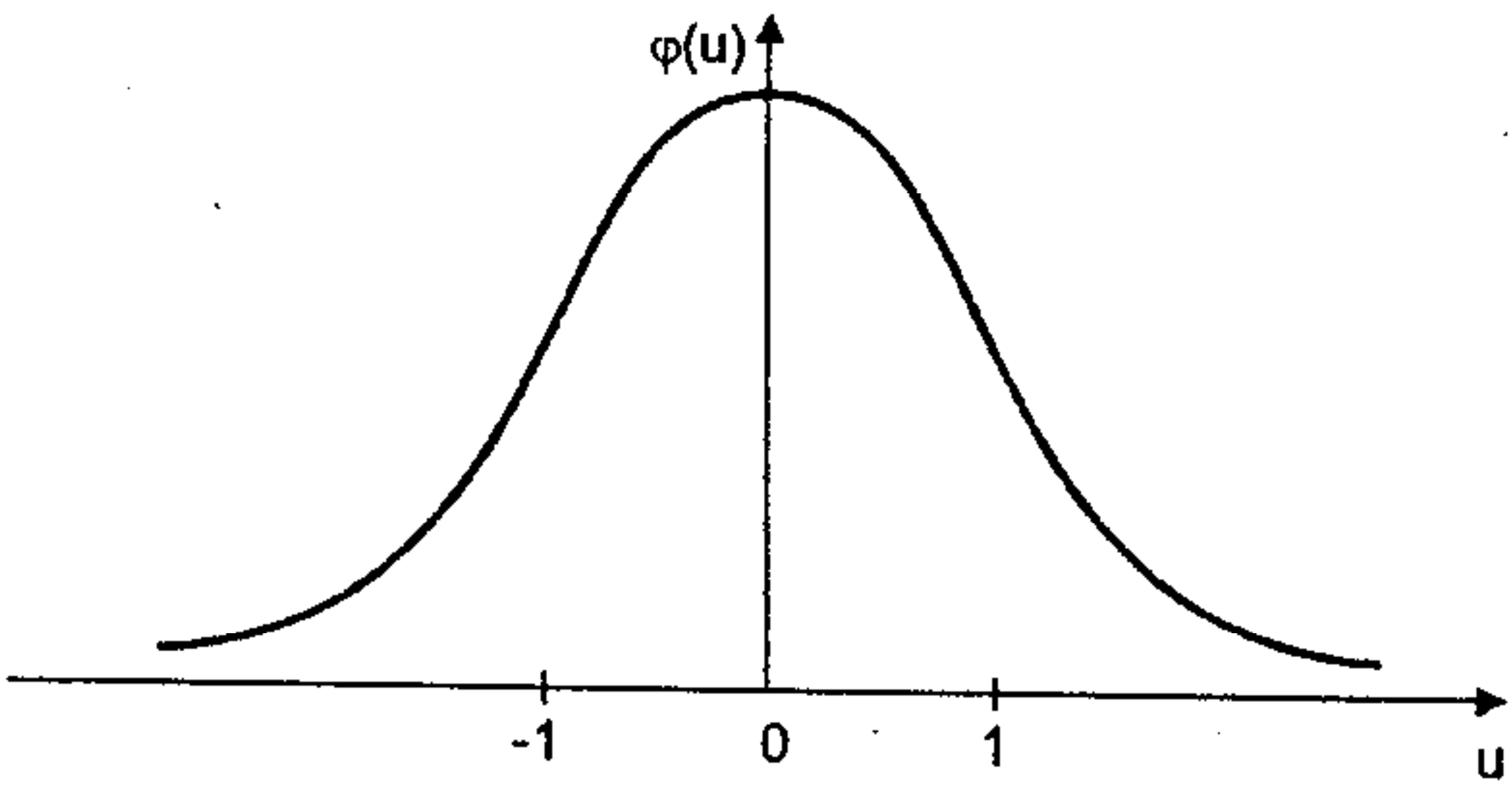
Giả sử biến ngẫu nhiên X phân phối chuẩn với kỳ vọng toán bằng μ và độ lệch chuẩn bằng σ. Xét biến ngẫu nhiên:

$$U = \frac{X-\mu}{\sigma}$$

7.3. Định nghĩa

Biến ngẫu nhiên U nhận các giá trị trong khoảng $(-\infty, +\infty)$ *gọi là tuân theo quy luật phân phối chuẩn hóa nếu hàm mật độ xác suất của nó có dạng:*

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \qquad (3.31)$$



***Hình 3.5.* Đồ thị hàm φ(u)**

Đồ thị của hàm φ(u) có dạng như hình vẽ.

Đặc điểm của đồ thị này là nó lấy trục tung làm trục đối xứng. Các giá trị của hàm φ(u) được tính sẵn thành bảng (Phụ lục 4).

Hàm phân bố xác suất của biến ngẫu nhiên U phân phối chuẩn hóa có dạng:

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{u^2}{2}} du$$

Các giá trị của hàm Φ(u) cũng được tính sẵn thành bảng (Phụ lục 5).

Ta tìm các tham số đặc trưng của biến ngẫu nhiên U phân phối chuẩn hóa:

$$E(U) = E\left(\frac{X-\mu}{\sigma}\right)$$

Theo tính chất của kỳ vọng toán ta có:

$$E(U) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}[E(X)-\mu]$$

Song $E(X) = \mu$, do đó $E(U) = 0$

$$V(U) = V\left(\frac{X-\mu}{\sigma}\right)$$

Cũng theo tính chất của phương sai ta có:

$$V(U) = \frac{1}{\sigma^2}V(X-\mu) = \frac{1}{\sigma^2}V(X)$$

Song $V(X) = \sigma^2$, do đó $V(U) = 1$

Phân phối chuẩn hóa được ký hiệu là N(0; 1).

Ngoài các tham số đặc trưng là kỳ vọng toán μ và phương sai σ^2, trong phân phối chuẩn có một tham số khác có nhiều ứng dụng trong thực tế, đó là giá trị tới hạn chuẩn.

7.4. Định nghĩa

Giá trị tới hạn chuẩn mức α (ký hiệu là u_α) là giá trị của biến ngẫu nhiên U có phân phối chuẩn hóa thỏa mãn điều kiện $P(U > u_\alpha) = \alpha$.

Vì U chuẩn hóa nên theo (3.31) hàm mật độ của U là:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Theo tính chất hàm mật độ thì:

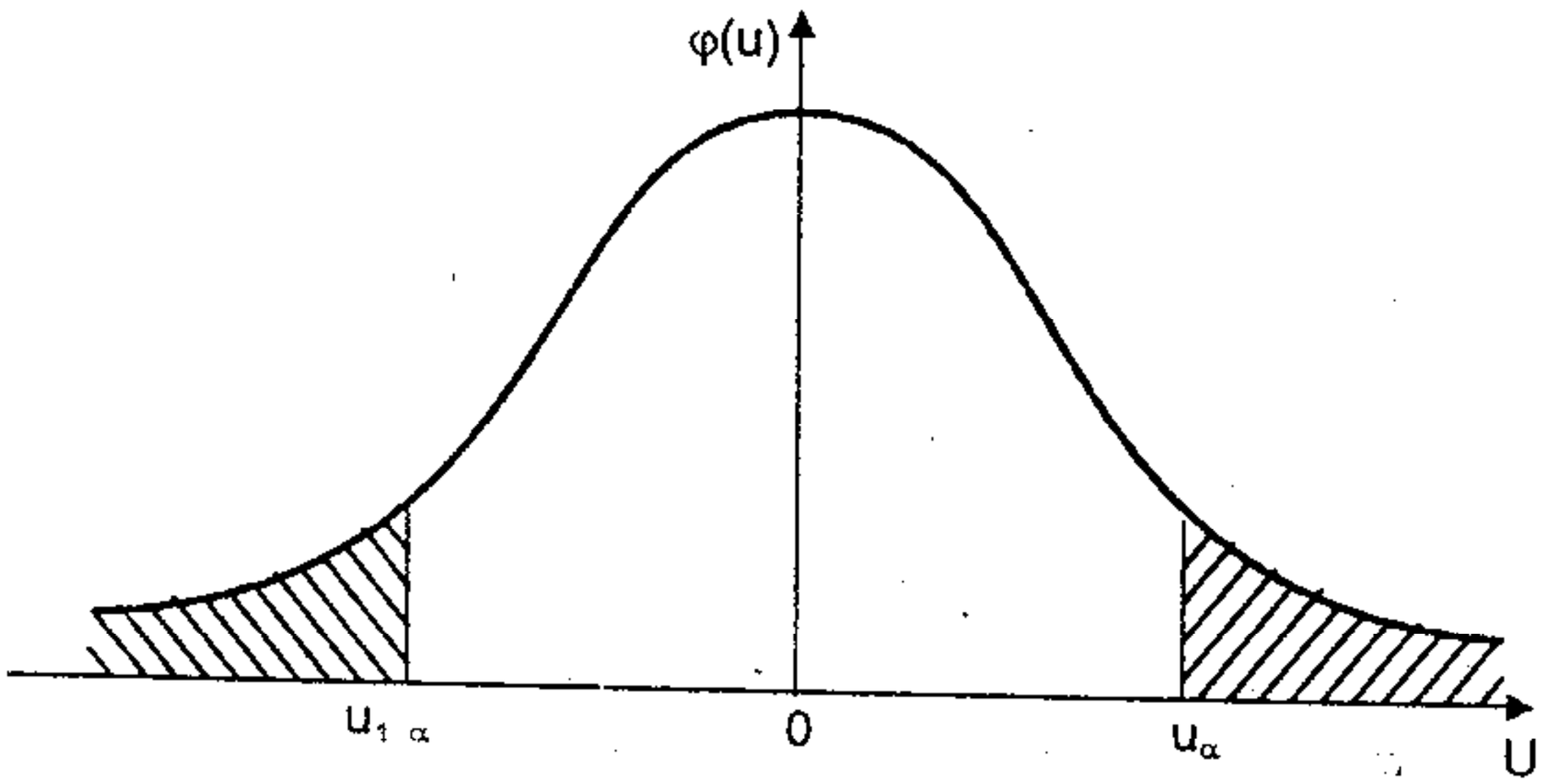
$$P(U > u_\alpha) = \int_{u_\alpha}^{+\infty} \varphi(u)du$$

Do đó: $P(U > u_\alpha) = \frac{1}{\sqrt{2\pi}} \int_{u_\alpha}^{+\infty} e^{-\frac{u^2}{2}} = \alpha$

Cho trước α, dựa vào biểu thức trên người ta tính được u_α và ngược lại.

Các giá trị của u_α được tính sẵn thành bảng (Phụ lục 6).

Trên đồ thị giá trị tới hạn chuẩn u_α là giá trị sao cho diện tích giới hạn bởi đường cong phân phối chuẩn hóa, trục 0U và đường thẳng $U = u_\alpha$ bằng α.



Hình 3.6. Giá trị tới hạn chuẩn u_α

Từ hình vẽ ta thấy ngay giá trị tới hạn chuẩn có tính chất sau đây:

$$u_\alpha = -u_{1-\alpha}$$

Sau đây ta sẽ xây dựng một số công thức có nhiều ứng dụng trong việc giải các bài toán thực tế.

7.5. Công thức tính xác suất để biến ngẫu nhiên X phân phối chuẩn nhận giá trị trong khoảng (a, b)

Ta biết rằng nếu biến ngẫu nhiên liên tục X có hàm mật độ xác suất là f(x) thì xác suất để X nhận giá trị trong khoảng (a, b) sẽ được tính bằng công thức:

$$P(a < X < b) = \int_a^b f(x)dx$$

Giả sử X phân phối chuẩn. Lúc đó:

$$P(a < X < b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Ta thay biến mới $Z = \frac{x-\mu}{\sigma}$, lúc đó $x = \sigma Z + \mu$ và $dx = \sigma dZ$.

Ta tìm cận lấy tích phân khi đổi biến. Khi $x = a$ thì $Z = \frac{a-\mu}{\sigma}$

và khi $x = b$ thì $Z = \frac{b-\mu}{\sigma}$. Như vậy ta có:

$$P(a < X < b) = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-\frac{Z^2}{2}} dZ =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{0} e^{-\frac{Z^2}{2}} dZ + \frac{1}{\sqrt{2\pi}} \int_0^{\frac{b-\mu}{\sigma}} e^{-\frac{Z^2}{2}} dZ =$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{\frac{b-\mu}{\sigma}} e^{-\frac{Z^2}{2}} dZ - \frac{1}{\sqrt{2\pi}} \int_0^{\frac{a-\mu}{\sigma}} e^{-\frac{Z^2}{2}} dZ =$$

$$= \Phi_0\left(\frac{b-\mu}{\sigma}\right) - \Phi_0\left(\frac{a-\mu}{\sigma}\right)$$

Trong đó:

$$\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{Z^2}{2}} dZ$$

Giá trị của hàm $\Phi_0(u)$ được tính sẵn thành bảng (Phụ lục 5).

Chú ý rằng hàm $\Phi_0(u)$ có các tính chất sau:

+ $\Phi_0(-u) = -\Phi_0(u)$

+ Với mọi $u > 5$ thì $\Phi_0(u) \approx \Phi_0(5) = 0,5$.

Các tính chất trên được vận dụng khi tra bảng giá trị hàm $\Phi_0(u)$.

Như vậy ta thu được công thức

$$P(a < X < b) = \Phi_0\left(\frac{b-\mu}{\sigma}\right) - \Phi_0\left(\frac{a-\mu}{\sigma}\right) \qquad (3.32)$$

Thí dụ 1. Một nhà sản xuất cần mua một loại gioáng cao su có độ dày từ 0,118cm đến 0,122cm. Có hai cửa hàng cùng bán loại gioăng này với độ dày là các biến ngẫu nhiên phân phối chuẩn với các đặc trưng được cho trong bảng sau:

	Độ dày trung bình	Độ lệch chuẩn	Giá bán
Cửa hàng A	0,12	0,001	3 USD/hộp/1000c
Cửa hàng B	0,12	0,0015	2,6 USD/hộp/1000c

Hỏi nhà sản xuất nên mua gioăng của cửa hàng nào?

Giải. Trước hết cần xác định tỷ lệ gioăng đáp ứng được yêu cầu của nhà sản xuất trong mỗi hộp sản phẩm của hai cửa hàng. Gọi X_A và X_B tương ứng là độ dày của gioăng do cửa hàng A và B bán. Theo giả thiết X_A và X_B đều phân phối chuẩn. Vì vậy, tỷ lệ gioăng dùng được của hai cửa hàng tương ứng là:

$$P(0,118 < X_A < 0,122) = \Phi_0\left(\frac{0,122-0,12}{0,001}\right) - \Phi_0\left(\frac{0,118-0,12}{0,001}\right)$$

$$= 2\Phi_0(2) = 0,9544$$

$$P(0,118 < X_B < 0,122) = \Phi_0\left(\frac{0,122-0,12}{0,0015}\right) - \Phi_0\left(\frac{0,118-0,12}{0,0015}\right)$$

$$= 2\Phi_0(1,33) = 0,8164$$

Vậy giá bán đối với mỗi chiếc gioăng dùng được của cửa hàng A là:

$$\frac{3}{954,4} = 0,00314\text{USD}$$

và của cửa hàng B là: $\frac{2,6}{816,4} = 0,00318\text{USD}$

Vậy nhà sản xuất nên mua gioăng của cửa hàng A.

7.6. Xác suất của sự sai lệch giữa biến ngẫu nhiên và kỳ vọng toán của nó

Trong thực tế nhiều khi ta phải tính xác suất để biến ngẫu nhiên X phân phối chuẩn nhận giá trị sai lệch so với kỳ vọng toán của nó về giá trị tuyệt đối nhỏ hơn một số dương cho trước, tức là ta phải tìm xác suất để xảy ra bất đẳng thức $|X - \mu| < \varepsilon$.

Ta thay bất đẳng thức trên bằng bất đẳng thức kép tương đương với nó.

$$\mu - \varepsilon < X < \mu + \varepsilon$$

Sử dụng công thức (3.32) ta có:

$$P(|X - \mu| < \varepsilon) = P(\mu - \varepsilon < X < \mu + \varepsilon)$$

$$= \Phi_0\left(\frac{\mu + \varepsilon - \mu}{\sigma}\right) - \Phi_0\left(\frac{\mu - \varepsilon - \mu}{\sigma}\right)$$

$$= \Phi_0\left(\frac{\varepsilon}{\sigma}\right) - \Phi_0\left(-\frac{\varepsilon}{\sigma}\right) = \Phi_0\left(\frac{\varepsilon}{\sigma}\right) + \Phi_0\left(\frac{\varepsilon}{\sigma}\right)$$

$$= 2\Phi_0\left(\frac{\varepsilon}{\sigma}\right)$$

Như vậy ta thu được công thức:

$$P(|X - \mu| < \varepsilon) = 2\Phi_0\left(\frac{\varepsilon}{\sigma}\right) \qquad (3.33)$$

Thí dụ 2. Các vòng bi do một máy tự động sản xuất ra được coi là đạt tiêu chuẩn nếu đường kính của nó sai lệch so với đường kính thiết kế không quá 0,7mm. Biết rằng sai lệch này là biến ngẫu nhiên phân phối chuẩn với $\mu = 0$ và $\sigma = 0,4$mm. Tìm tỷ lệ vòng bi đạt tiêu chuẩn của máy đó.

Giải. Ta thấy rằng tỷ lệ vòng bi đạt tiêu chuẩn chính là xác suất để lấy ngẫu nhiên một vòng bi được vòng bi đạt tiêu chuẩn. Nếu gọi X là sai lệch giữa đường kính của vòng bi được sản xuất ra so với đường kính thiết kế thì xác suất này chính là xác suất để xảy ra bất đẳng thức $|X - \mu| < 0,7$. Theo giả thiết X phân phối chuẩn với $\mu = 0$ và $\sigma = 0,4$ do đó theo công thức (3.33) ta có:

$$P(|X-\mu|<0,7)=P(|X|<0,7)=2\Phi_0\left(\frac{0,7}{0,4}\right)=2\Phi_0(1,75)$$

Tra bảng giá trị hàm $\Phi_0(U)$ ta có: $\Phi_0(1,75)=0,4599$

Do đó $P(|X|<0,7)=2.0,4599=0,9198$.

Vậy tỷ lệ vòng bi đạt tiêu chuẩn của máy đó gần bằng 92%.

Ta chú ý rằng do tính chất $\int_{-\infty}^{+\infty}\varphi(u)du=1$ và do hàm $\varphi(u)$ đối xứng qua gốc tọa độ do đó $\int_{-\infty}^{0}\varphi(u)du=0,5$ và như vậy:

$$P(-\infty<U<0)=0,5$$

Từ đó $\Phi(u)=P(-\infty<U<u)=P(-\infty<U<0)+P(0<U<u)$
$=0,5+\Phi_0(U)$

Như vậy ta có mối liên hệ sau $\Phi_0(u)=0,5+\Phi_0(U)$

7.7. Quy tắc hai xích ma và ba xích ma

Nếu trong công thức (3.33) ta đặt $\varepsilon=2\sigma$ tức là bằng hai lần độ lệch chuẩn của X thì ta có:

$$P(|X-\mu|<2\sigma)=P(\mu-2\sigma<X<\mu+2\sigma)$$
$$=P(\mu-2\sigma<X<\mu+2\sigma)$$
$$=0,9544 \qquad (3.34)$$

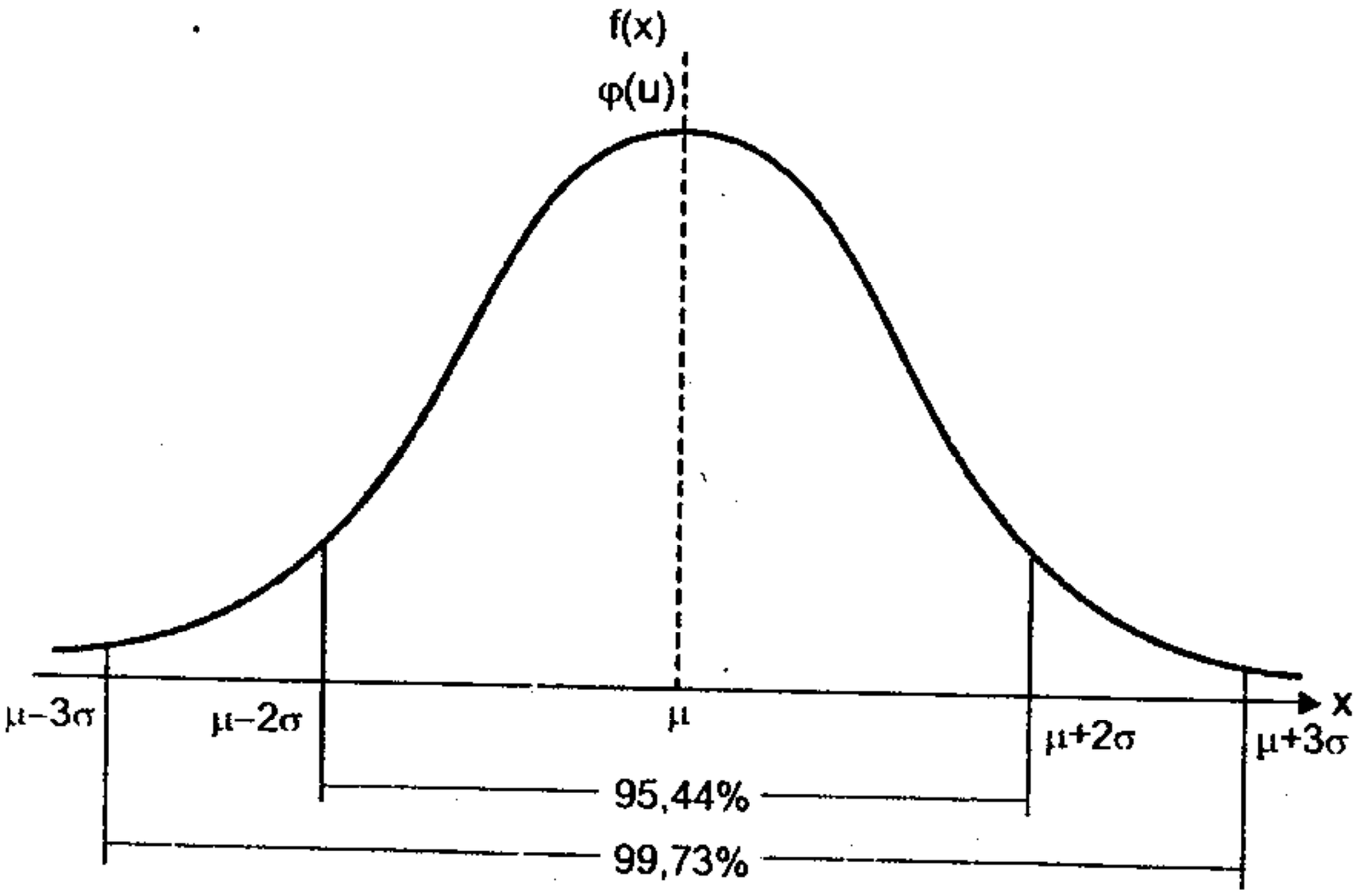
Biểu thức (3.34) được gọi là quy tắc hai xích ma. Tương tự nếu đặt $\varepsilon=3\sigma$ thì ta sẽ thu được biểu thức tương ứng.

$$P(\mu-3\sigma<X<\mu+3\sigma)\approx 0,9973 \qquad (3.35)$$

Các quy tắc trên cho thấy xác suất để biến ngẫu nhiên phân phối chuẩn nhận giá trị trong khoảng $(\mu-2\sigma;\ \mu+2\sigma)$

là 0,9544 hay 95,44% các giá trị của X sẽ nằm trong khoảng nói trên. Còn theo quy tắc ba xích ma thì hầu hết (tới 99,73%) các giá trị của X phân phối chuẩn sẽ nằm trong khoảng $(\mu - 3\sigma; \mu + 3\sigma)$. Điều này được thể hiện trên đồ thị như sau (hình 3.7).

Trong thực tế quy tắc hai xích ma và quy tắc ba xích ma được áp dụng như sau: Nếu quy luật phân phối xác suất của biến ngẫu nhiên được nghiên cứu chưa biết, song nó thỏa mãn điều kiện của quy tắc hai xích ma hoặc ba xích ma thì có thể xem như biến ngẫu nhiên đó phân phối chuẩn.



Hình 3.7. **Quy tắc hai xích ma và ba xích ma**

Mặt khác nếu biến ngẫu nhiên phân phối chuẩn thì 95,44% các giá trị của nó sẽ nằm trong khoảng $(\mu - 2\sigma; \mu + 2\sigma)$ còn

hầu như toàn bộ các giá trị của nó sẽ nằm trong khoảng $(\mu - 3\sigma; \mu + 3\sigma)$.

7.8. Phân phối xác suất của tổng các biến ngẫu nhiên độc lập tuân theo cùng một quy luật

Giả sử X_1 và X_2 là hai biến ngẫu nhiên độc lập. X_1 tuân theo quy luật chuẩn với kỳ vọng toán μ_1 và phương sai σ_1^2 còn X_2 cũng tuân theo quy luật chuẩn với kỳ vọng toán μ_2 và phương sai σ_2^2. Lúc đó tổng của chúng là biến ngẫu nhiên $X = X_1 + X_2$ cũng phân phối theo quy luật chuẩn với kỳ vọng toán là $\mu_1 + \mu_2$ và phương sai là $\sigma_1^2 + \sigma_2^2$ (xem thêm mục §9 chương IV). Tính chất trên cũng có thể mở rộng cho một số bất kỳ các biến ngẫu nhiên độc lập lẫn nhau và cùng phân phối chuẩn.

Mặt khác nếu $X_1X_2 \ldots X_n$ là n biến ngẫu nhiên độc lập lẫn nhau và cùng tuân theo một quy luật phân phối xác suất nào đó (không nhất thiết là quy luật chuẩn) với các kỳ vọng toán $E(X_1)$, $E(X_2)$ ... $E(X_n)$ và các phương sai $V(X_1)$, $V(X_2)$ ... $V(X_n)$ đã biết thì biến ngẫu nhiên $X = \sum_{i=1}^{n} X_i$ sẽ phân phối xấp xỉ chuẩn với $E(X) = \sum_{i=1}^{n} E(x_i)$ và $V(X) = \sum_{i=1}^{n} V(x_i)$ khi n khá lớn $(n > 30)$. Tính chất trên thường được gọi là định lý giới hạn trung tâm của Liapunốp (xem thêm mục §4 Chương V).

7.9. Sự hội tụ của quy luật nhị thức và quy luật Poisson về quy luật chuẩn

Khi sử dụng quy luật nhị thức, nếu n khá lớn thì việc tính toán theo công thức Bernoulli sẽ gặp khó khăn. Lúc đó nếu p nhỏ đến mức $np \approx npq$ thì có thể dùng quy luật Poisson thay thế cho quy luật nhị thức. Song nếu p lại không nhỏ

(p > 0,1) thì không thể dùng quy luật Poisson để thay thế được. Lúc đó có thể dùng quy luật chuẩn để thay thế cho quy luật nhị thức.

Trong thực tế quy luật chuẩn có thể thay thế cho quy luật nhị thức nếu thỏa mãn đồng thời hai điều kiện là:

$$n > 5 \text{ và } \left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \frac{1}{\sqrt{n}} < 0,3$$

Lúc đó thì biến ngẫu nhiên X phân phối nhị thức có thể coi như phân phối xấp xỉ chuẩn với kỳ vọng toán $\mu = np$ và phương sai $\sigma^2 = npq$.

Từ đó:

$$P(X = x) = C_n^x p^x q^{n-x} \approx \frac{1}{\sqrt{npq}} \varphi\left(\frac{x - np}{\sqrt{npq}}\right) \qquad (3.36)$$

Công thức (3.36) được gọi là định lý địa phương Laplace.

Mặt khác:

$$P(x \le X \le x + h) = P_x + P_{x+1} + \ldots + P_{x+h} \approx$$

$$\approx \Phi_0\left(\frac{x + h - np}{\sqrt{npq}}\right) - \Phi_0\left(\frac{x - np}{\sqrt{npq}}\right) \qquad (3.37)$$

Công thức (3.37) được gọi là định lý tích phân Laplace (xem thêm mục §4 chương V).

Thí dụ 3. Xác suất để sản phẩm sau khi sản xuất không được kiểm tra chất lượng bằng 0,2. Tìm xác suất để trong 400 sản phẩm được sản xuất ra có:

a - 80 sản phẩm không được kiểm tra chất lượng.

b - Có từ 70 đến 100 sản phẩm không được kiểm tra chất lượng.

Giải. Bài toán thỏa mãn lược đồ Bernoulli do đó nếu gọi X là số sản phẩm không được kiểm tra chất lượng thì $X \sim B$ $(n = 400; p = 0,2)$. Song vì $n = 400 > 5$ và

$$\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right|\frac{1}{\sqrt{n}} = \left|\frac{\sqrt{0,2}}{\sqrt{0,8}} - \frac{\sqrt{0,8}}{\sqrt{0,2}}\right|\frac{1}{\sqrt{400}} = 0,075 < 0,3$$

nên có thể coi $X \sim N$ $(\mu = np = 80; \sigma^2 = npq = 64)$

a. Từ đó $P(X = 80) \approx \frac{1}{\sqrt{400.0,2.0,8}}\varphi(u) = \frac{1}{8}\varphi(u)$

với $u = \frac{80 - 400.0,2}{\sqrt{400.0,2.0,8}} = 0$

$$\rightarrow P(X = 80) = \frac{1}{8}\varphi(0) = \frac{1}{8}.0,3989 = 0,04986$$

b. $P(70 \le X \le 100) \approx \Phi_0\left(\frac{100 - 400.0,2}{\sqrt{400.0,2.0,8}}\right) - \Phi_0\left(\frac{70 - 400.0,2}{\sqrt{400.0,2.0,8}}\right)$

$$= \Phi_0(2,5) - \Phi_0(-1,25)$$

$$= \Phi_0(2,5) + \Phi_0(1,25) = 0,8882$$

Đối với quy luật Poisson thì quá trình hội tụ của nó về quy luật chuẩn sẽ diễn ra khi λ trở nên lớn hơn 20. Vì vậy, nếu X phân phối Poisson song $\lambda > 20$ thì có thể xem là X phân phối xấp xỉ chuẩn với $\mu = \lambda$ và $\sigma^2 = \lambda$.

7.10. Ứng dụng của quy luật chuẩn

Quy luật phân phối chuẩn là quy luật phân phối xác suất được áp dụng rất rộng rãi trong thực tế. Trong nhiều lĩnh vực

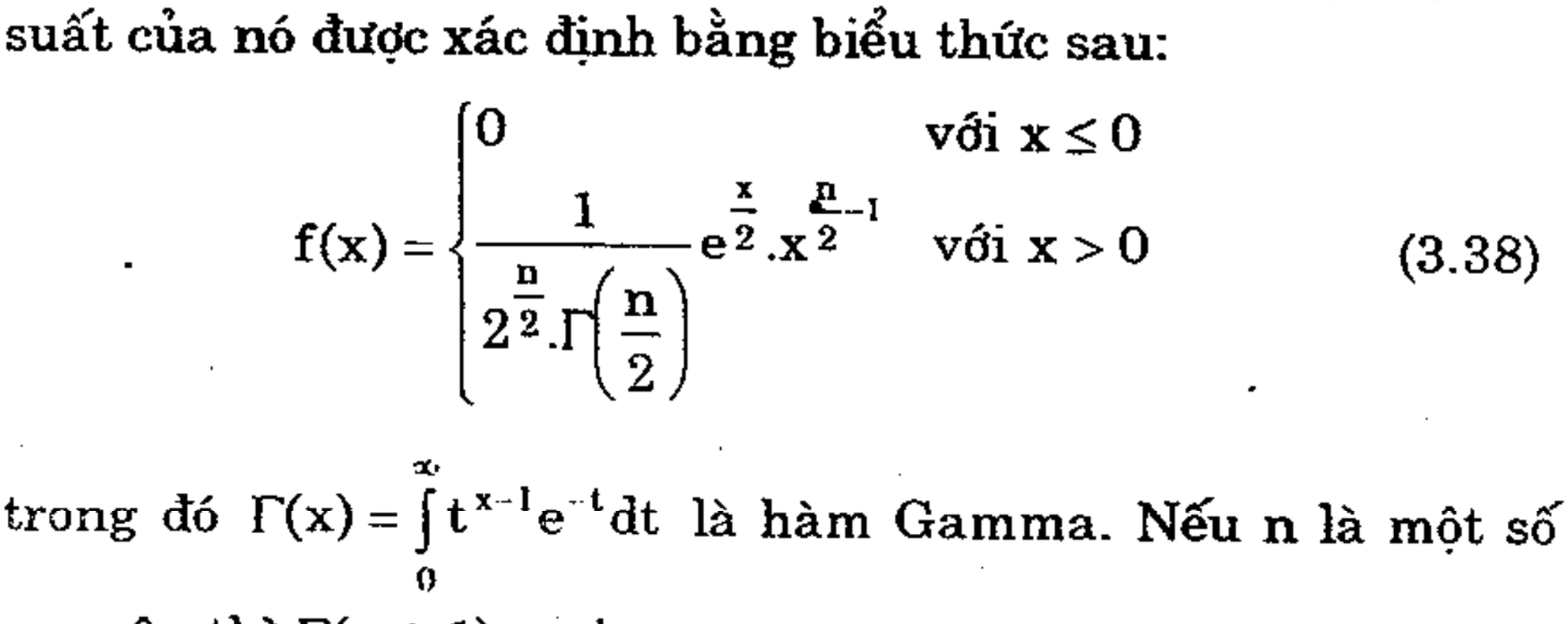
của khoa học và đời sống ta đều gặp các biến ngẫu nhiên phân phối chuẩn. Lý do của sự phổ biến đó không những đã được giải thích trong định lý giới hạn trung tâm như đã xét ở trên mà còn từ hệ quả của định lý đó: *Nếu biến ngẫu nhiên X là tổng của một số lớn các biến ngẫu nhiên độc lập và giá trị của mỗi biến chỉ chiếm vị trí rất nhỏ trong tổng đó thì X sẽ có phân phối xấp xỉ chuẩn.* Trong thực tế ta gặp chính các biến ngẫu nhiên như vậy. Chẳng hạn, trong công nghiệp người ta đã xác định được rằng kích thước của các chi tiết do các nhà máy sản xuất ra sẽ phân phối chuẩn nếu quá trình sản xuất diễn ra bình thường. Trong nông nghiệp năng suất của cùng một loại cây trồng tại các thửa ruộng khác nhau cũng phân phối chuẩn. Năng suất lao động của các công nhân có cùng tay nghề và làm cùng một công việc như nhau cũng phân phối chuẩn. Nhu cầu về các loại hàng hóa khác nhau cũng phân phối chuẩn v.v... Người ta ghi nhận rằng các năng lực về trí tuệ và thể lực con người cũng phân phối theo quy luật chuẩn. Thậm chí cả một số chỉ tiêu về sinh lý của những người cùng giới (chẳng hạn chiều cao, vòng ngực, chiều dài cánh tay v.v...) cũng phân phối theo quy luật chuẩn. Sự nhận biết này cho phép lập kế hoạch sản xuất quần áo may sẵn sản xuất hàng loạt sao cho đáp ứng một cách hợp lý nhất kích cỡ của người mua, tránh tình trạng thừa, thiếu do không vừa kích cỡ v.v... Tóm lại, khó có thể liệt kê được hết các hiện tượng và lĩnh vực trong đó có thể áp dụng quy luật phân phối chuẩn.

§8. QUY LUẬT KHI BÌNH PHƯƠNG - $\chi^2(n)$

Biến ngẫu nhiên liên tục χ^2 gọi là phân phối theo quy luật khi bình phương với n bậc tự do nếu hàm mật độ xác suất của nó được xác định bằng biểu thức sau:

$$f(x) = \begin{cases} 0 & \text{với } x \le 0 \\ \dfrac{1}{2^{\frac{n}{2}}.\Gamma\left(\frac{n}{2}\right)} e^{\frac{x}{2}}.x^{\frac{n}{2}-1} & \text{với } x > 0 \end{cases} \qquad (3.38)$$

trong đó $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt$ là hàm Gamma. Nếu n là một số nguyên thì $\Gamma(n+1) = n!$



Hình 3.8. **Đồ thị hàm $f(\chi^2)$ của quy luật "khi bình phương"**

Đồ thị của hàm $f(\chi)$ có dạng như ở hình 3.8. Có thể chứng minh được rằng nếu biến ngẫu nhiên χ^2 phân phối theo quy luật khi bình phương với n bậc tự do thì kỳ vọng toán:

$$E(\chi^2) = n$$

và phương sai

$$V(\chi^2) = 2n$$

Ngoài ra trong quy luật "khi bình phương" giá trị tới hạn χ^2 cũng là tham số được sử dụng nhiều. Giá trị tới hạn "khi bình phương" mức α ký hiệu là $\chi_\alpha^{2(n)}$ là giá trị của biến ngẫu nhiên χ^2 tuân theo quy luật phân phối "khi bình phương" với n bậc tự do thỏa mãn điều kiện:

$$P(\chi^2 > \chi_\alpha^{2(n)}) = \alpha$$

Các giá trị tới hạn $\chi_\alpha^{2(n)}$ được tính sẵn thành bảng (Phụ lục 7).

Ý nghĩa của nó được thấy rõ ở hình 3.8.

Khi số bậc tự do n tăng lên, quy luật "khi bình phương" sẽ xấp xỉ với quy luật chuẩn.

Quy luật khi bình phương có tính chất sau đây: Nếu χ_1^2 và χ_2^2 là các biến ngẫu nhiên độc lập cùng phân phối theo quy luật "khi bình phương" với số bậc tự do tương ứng là n_1 và n_2 thì tổng của chúng là biến ngẫu nhiên:

$$\chi^2 = \chi_1^2 + \chi_2^2$$

cũng phân phối theo quy luật khi bình phương với số bậc tự do là $n = n_1 + n_2$.

Trong thực tế quy luật khi bình phương thường được sử dụng trong trường hợp sau đây: Giả sử có các biến ngẫu nhiên x_i $(i = \overline{1,n})$ độc lập, cùng phân phối theo quy luật chuẩn hóa tức là có kỳ vọng toán bằng không và độ lệch

chuẩn bằng một. Nếu xét tổng bình phương của các biến ngẫu nhiên nói trên ta có:

$$\chi^2 = \sum_{i=1}^{n} x_i^2$$

Biến ngẫu nhiên χ^2 sẽ phân phối theo quy luật khi bình phương với n bậc tự do.

§9. QUY LUẬT STUDENT - T(n)

Biến ngẫu nhiên liên tục T gọi là phân phối theo quy luật Student với n bậc tự do nếu hàm mật độ xác suất của nó được xác định bằng biểu thức sau:

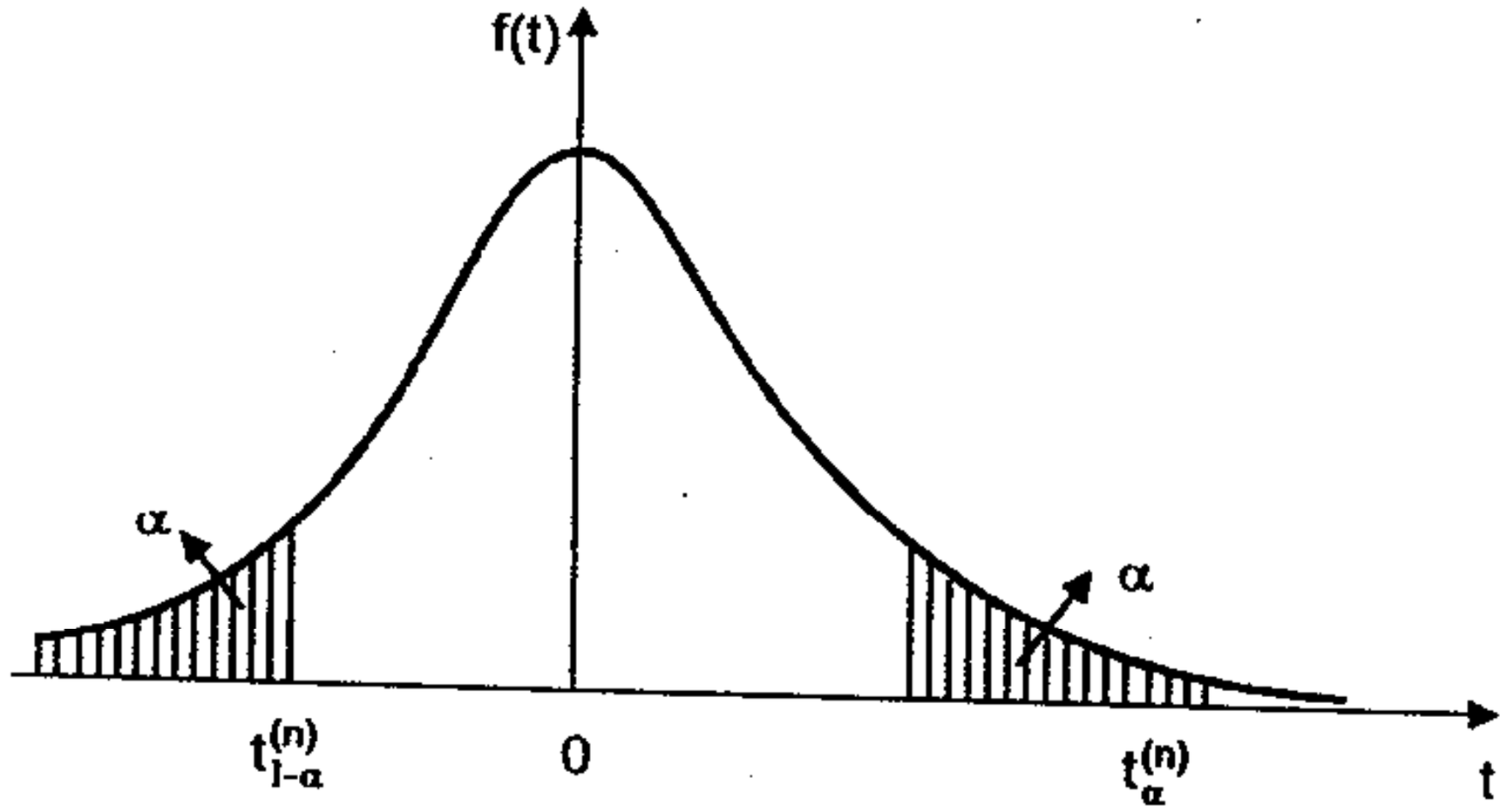
$$f(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)}\left[1+\frac{t^2}{n-1}\right]^{-\frac{n}{2}} \quad \forall t \qquad (3.39)$$

trong đó $\Gamma(x)$ là hàm Gamma.

Đồ thị của hàm f(t) có dạng như ở hình 3.9.

Có thể chứng minh được rằng nếu biến ngẫu nhiên T phân phối theo quy luật Student với n bậc tự do thì kỳ vọng toán $E(T) = 0$ và phương sai

$$V(T) = \frac{n}{n-2}$$



***Hình 3.9.* Đồ thị hàm f(t) của quy luật Student**

Giá trị tới hạn Student, ký hiệu là $t^{(n)}_{\alpha}$ là giá trị của biến ngẫu nhiên T phân phối theo quy luật Student với n bậc tự do, thỏa mãn điều kiện:

$$P(T > t^{(n)}_{\alpha}) = \alpha$$

Các giá trị tới hạn $t^{(n)}_{\alpha}$ được tính sẵn thành bảng (Phụ lục 8).

Giá trị tới hạn Student có tính chất sau đây:

$$t^{(n)}_{\alpha} = -t^{(n)}_{1-\alpha}$$

Ý nghĩa của nó được thể hiện trên hình 3.9.

Khi số bậc tự do n tăng lên, phân phối Student sẽ hội tụ rất nhanh về phân phối chuẩn hóa. Do đó nếu n khá lớn (n > 30) có thể dùng phân phối chuẩn hóa thay thế cho phân phối Student.

Tuy nhiên cần phải nhấn mạnh rằng với số bậc tự do nhỏ (n < 30) việc thay thế quy luật Student bằng quy luật

chuẩn có thể dẫn đến những sai sót rất lớn. Chẳng hạn với $n = 4$ và $\alpha = 0,05$ thì giá trị tới hạn Student $t^{(4)}_{0,05} = 4,6$ trong khi đó $U_{0,05} = 2,58$, tức là sai lệch nhau tới $4,6 - 2,58 = 2,02$.

Trong thực tế quy luật Student thường được sử dụng trong trường hợp sau đây: Giả sử có U là biến ngẫu nhiên phân phối chuẩn hóa N(0, 1) và biến ngẫu nhiên V độc lập với U, phân phối theo quy luật khi bình phương với n bậc tự do.

Nếu xét biến ngẫu nhiên

$$T = \frac{U}{\sqrt{\frac{V}{n}}}$$

thì biến ngẫu nhiên T sẽ phân phối theo quy luật Student với n bậc tự do.

§10. QUY LUẬT FISHER - SNEDECOR - $F(n_1, n_2)$

Biến ngẫu nhiên liên tục F gọi là phân phối theo quy luật Fisher - Snedecor với n_1 và n_2 bậc tự do nếu hàm mật độ xác suất của nó được xác định bằng biểu thức sau:

$$f(x) = \begin{cases} 0 & \text{với } x \le 0 \\ C \dfrac{x^{\frac{n_1 - n_2}{2}}}{(n_2 + n_1.x)^{\frac{(n_1 \cdot n_2)}{2}}} & \text{với } x > 0 \end{cases} \qquad (3.40)$$

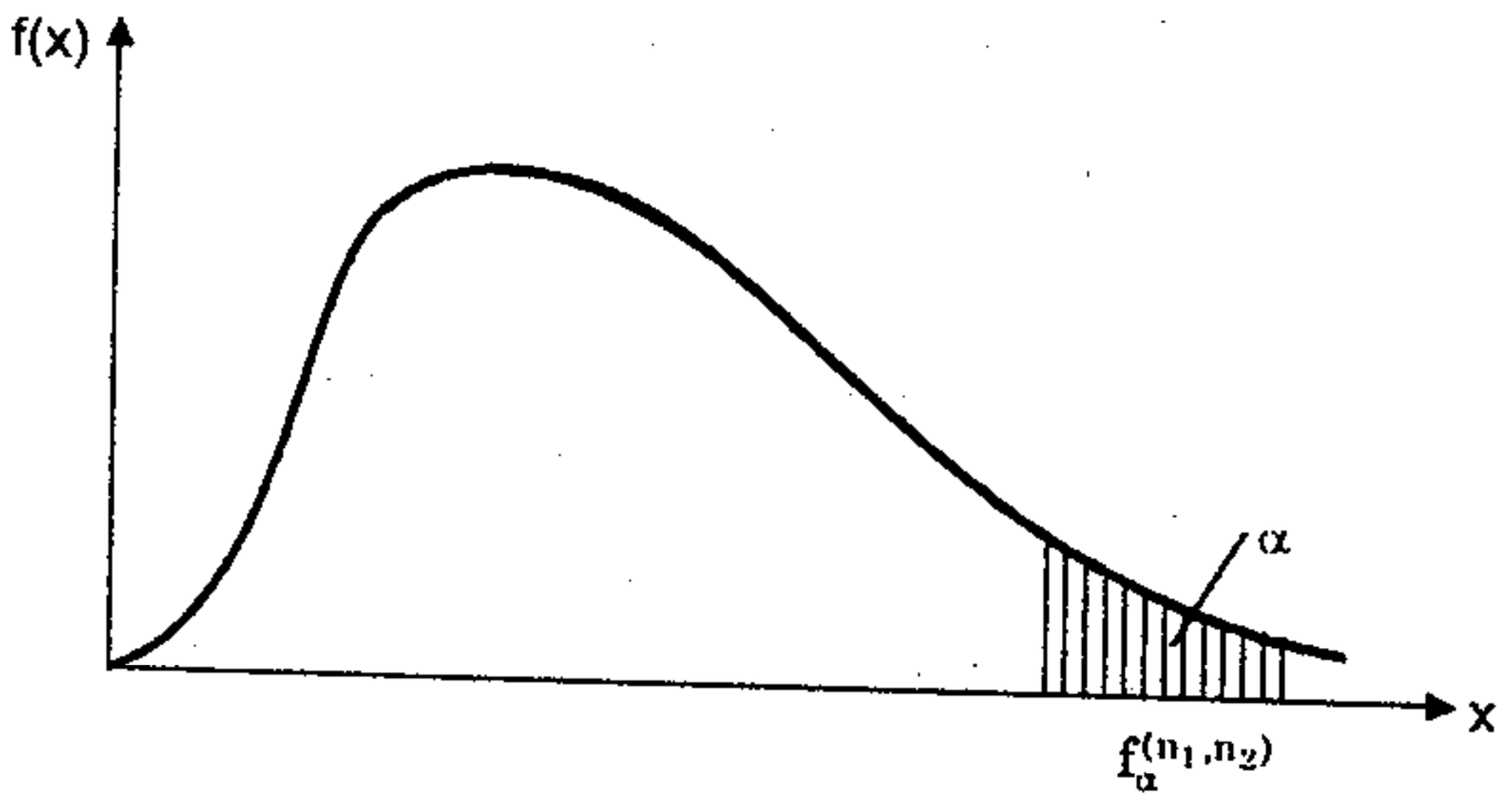
với $C = \dfrac{\Gamma\left(\dfrac{n_1 + n_2}{2}\right).n_1^{\frac{n_1}{2}}.n_2^{\frac{n_2}{2}}}{\Gamma\left(\dfrac{n_1}{2}\right).\Gamma\left(\dfrac{n_2}{2}\right)}$

Đồ thị của hàm f(x) có dạng sau (xem hình 3.10)

Có thể chứng minh được rằng trong quy luật Fisher - Snedecor thì:

$$E(F) = \frac{n_2}{n_2 - 2}$$

và $V(F) = \dfrac{2n_2^2(n_1 + n_2^2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}$



Hình 3.10. **Đồ thị của hàm f(x) của phân phối Fisher - Snedecor**

Giá trị tới hạn Fisher - Snedecor ký hiệu là $f_\alpha^{(n_1,n_2)}$, là giá trị của biến ngẫu nhiên F phân phối theo quy luật Fisher - Snedecor với n_1 và n_2 bậc tự do, thỏa mãn điều kiện:

$$P(F > f_{\alpha}^{(n_1,n_2)}) = \alpha$$

Giá trị $f_{\alpha}^{(n_1,n_2)}$ có tính chất sau đây:

$$f_{\alpha}^{(n_1,n_2)} = \frac{1}{f_{1-\alpha}^{(n_2,n_1)}}$$

Ý nghĩa của nó được thể hiện trên hình 3.10.

Các giá trị tới hạn $f_{\alpha}^{(n_1,n_2)}$ được tính sẵn thành bảng (Phụ lục 9).

Trong thực tế quy luật Fisher - Snedecor thường được sử dụng trong trường hợp sau: Giả sử có các biến ngẫu nhiên U và V độc lập với nhau và cùng phân phối theo quy luật khi bình phương với các bậc tự do tương ứng là n_1 và n_2. Lúc đó nếu xét biến ngẫu nhiên

$$F = \frac{\frac{U}{n_1}}{\frac{V}{n_2}}$$

thì F sẽ phân phối theo quy luật Fisher - Snedecor với n_1 và n_2 bậc tự do.

Các ký hiệu và công thức cơ bản

* Quy luật không - một - A(p)

$$P_x = p^x (1-p)^{1-x} \qquad x = \overline{0,1}$$

$$E(X) = p; \quad V(X) = p(1-p)$$

* Quy luật nhị thức - B(n, p)

$$P_x = C_n^x p^x (1-p)^{n-x} \qquad x = \overline{0,n}$$

$$E(X) = np;\ V(X) = np(1-p)$$

$$np + p - 1 \le m_0 \le np + p$$

* Quy luật Poisson - P(λ)

$$P_x = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2,...$$

$$E(X) = \lambda;\ V(X) = \lambda$$

$$\lambda - 1 \le m_0 \le \lambda$$

* Quy luật siêu bội - M(N, n)

$$P_x = \frac{C_M^x . C_{N-M}^{n-x}}{C_N^n} \qquad x = \overline{0,n}$$

$$E(X) = n.\frac{M}{N} = np;\ V(X) = np(1-p)\frac{N-n}{N-1}$$

* Quy luật phân phối đều - U(a, b)

$$f(x) = \begin{cases} 0 & x \notin (a,b) \\ \dfrac{1}{b-a} & x \in (a,b) \end{cases}$$

$$E(X) = \frac{a+b}{2},\ V(X) = \frac{(b-a)^2}{12}$$

* Quy luật phân phối lũy thừa - E(λ)

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \ge 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda},\ V(X) = \frac{1}{\lambda^2}$$

* Quy luật phân phối chuẩn - $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty)$$

$$E(X) = \mu;\ V(X) = \sigma^2$$

Nếu $U = \frac{X-\mu}{\sigma}$ thì U phân phối chuẩn hóa N(0, 1)

với $\varphi(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

và $E(U) = 0;\ V(U) = 1$

$$P(a < X < b) = \Phi_0\left(\frac{b-\mu}{\sigma}\right) - \Phi_0\left(\frac{a-\mu}{\sigma}\right)$$

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

Quy tắc 2σ và 3σ

$P(|X - \mu| < 2\sigma) = 2\Phi_0(2) = 0{,}9544$

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

Nếu $n > 5$ và $\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \times \frac{1}{\sqrt{n}} < 0{,}3$ thì quy luật nhị thức hội tụ về quy luật chuẩn N ($\mu = np$; $\sigma^2 = np(1 - p)$). Nếu $\lambda > 20$ thì quy luật Poisson hội tụ về quy luật chuẩn N ($\mu = \lambda$; $\sigma^2 = \lambda$).

* Quy luật khi bình phương - $\chi^2(n)$

Nếu $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên độc lập nhau và cùng tuân theo quy luật chuẩn hóa N(0, 1) thì:

$$\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

*** Quy luật Student - T(n)**

Nếu U và V độc lập U ~ N(0, 1) và V ~ $\chi^2(n)$ thì:

$$T = \frac{U}{\sqrt{\frac{V}{n}}} \sim T(n)$$

*** Quy luật Fisher - Snedecor - $F(n_1, n_2)$**

Nếu U và V độc lập, U ~ $\chi^2(n_1)$ và V ~ $\chi^2(n_2)$ thì:

$$F = \frac{\frac{U}{n_1}}{\frac{V}{n_2}} \sim F(n_1, n_2)$$

Câu hỏi ôn tập

1. Cho một thí dụ trong kinh tế hoặc kinh doanh về:

a. Một biến ngẫu nhiên phân phối A(p)

b. Một biến ngẫu nhiên phân phối B(n, p)

c. Một biến ngẫu nhiên phân phối $P(\lambda)$

d. Một biến ngẫu nhiên phân phối M(N, n)

2. Hai kiện tướng bóng bàn ngang sức thi đấu với nhau. Hỏi thắng 2 trong 4 ván dễ hơn hay thắng 3 trong 6 ván dễ hơn?

3. Một cầu thủ nổi tiếng về đá phạt đền với xác suất đá vào gôn là 4/5. Có người cho rằng vậy thì cứ sút 5 quả thì chắc chắn có 4 quả vào gôn. Điều khẳng định đó có đúng không? Tại sao?

4. Hãy cho biết các mệnh đề sau đây là đúng hay sai? Tại sao?

a. Tổng của một số hữu hạn n các biến ngẫu nhiên phân phối A(p) sẽ phân phối B(n, p)

b. Nếu $X_1 \sim B(n_1, p)$ và $X_2 \sim B(n_2, p)$ thì:

$$X_1 + X_2 \sim B(n_1 + n_2, p).$$

5. Với những điều kiện nào thì quy luật nhị thức B(n, p) sẽ hội tụ về quy luật Poisson $P(\lambda)$?

6. Với những điều kiện nào thì quy luật siêu bội M(N, n) xấp xỉ quy luật nhị thức B(n, p).

7. Cho một thí dụ trong kinh tế hoặc kinh doanh về:

a. Một biến ngẫu nhiên phân phối U(a, b)

b. Một biến ngẫu nhiên phân phối $N(\mu, \sigma^2)$.

8. Phân phối chuẩn là rời rạc hay liên tục? Tại sao?

9. a. Vẽ phác qua đồ thị hàm mật độ xác suất của phân phối chuẩn hóa.

b. Cho u_α là một giá trị thỏa mãn $\Phi(u_\alpha) = \alpha$. Hãy minh họa bằng đồ thị các giá trị α, u_α, $\Phi_0(u_\alpha)$, $\varphi(u_\alpha)$, $P(U > u_\alpha)$ và $P(a < U < b)$ với a và b là hai giá trị nào đó của U.

10. Cho hai biến ngẫu nhiên cùng phân phối chuẩn với kỳ vọng toán bằng nhau song phương sai khác nhau. Nếu cùng lấy ở mức 68,8% các giá trị của hai biến ngẫu nhiên ấy

thì khoảng giá trị của biến ngẫu nhiên nào sẽ rộng hơn. Hãy mô tả kết luận bằng đồ thị.

11. Biến ngẫu nhiên X tuân theo quy luật chuẩn với kỳ vọng toán bằng 10. Xác suất để X nhận giá trị trong khoảng (10; 20) là 0,3. Không cần tính toán, hãy cho biết xác suất để X nhận giá trị trong khoảng (0; 10) bằng bao nhiêu?

12. Biến ngẫu nhiên liên tục X tuân theo quy luật chuẩn với kỳ vọng toán $\mu = 3$ và độ lệch chuẩn $\sigma = 2$. Viết hàm mật độ xác suất của biến ngẫu nhiên đó.

13. Biến ngẫu nhiên liên tục X có hàm phân bố xác suất như sau:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\mu)^2/2\sigma^2} dx$$

a. Tìm hàm mật độ xác suất của X, từ đó cho biết X phân phối theo quy luật nào?

b. Dùng bảng tính $P(0 \le X \le 0,92)$ khi $\mu = 0$ và $\sigma^2 = 1$

14. Hãy sử dụng bảng giá trị tới hạn chuẩn để tìm các xác suất sau:

a. $P(U > 1,96)$ b. $P(U > 1,64)$

c. $P(U < -1,64)$ d. $P(U < 1,64)$

e. $P(1 < U < 1,5)$ f. $P(-1 < U < 2)$

15. Hãy dùng bảng giá trị hàm $\Phi_0(x)$ và bảng giá trị tới hạn chuẩn để tìm các xác suất sau nếu cho biến ngẫu nhiên X phân phối chuẩn với $\mu = 10$ và $\sigma = 5$. So sánh kết quả thu được bằng hai phương pháp:

a. $P(X > 20)$ b. $P(20 < X < 25)$

c. $P(X < 10)$ d. $P(12 < X < 24)$

16. Cho các biến ngẫu nhiên X_1, X_2, X_3 độc lập nhau, cùng phân phối chuẩn với các kỳ vọng toán tương ứng là 10; 15; 20 và các phương sai tương ứng là 3; 6; 9. Tìm quy luật phân phối xác suất và các tham số đặc trưng của biến ngẫu nhiên $X = X_1 + X_2 + X_3$.

17. Người ta kiểm tra chất lượng 900 chi tiết. Xác suất để chi tiết đạt tiêu chuẩn là 0,9. Hãy tìm với xác suất 0,9544 xem số sản phẩm đạt tiêu chuẩn nằm trong khoảng nào xung quanh số chi tiết đạt tiêu chuẩn trung bình?

18. Dùng bảng giá trị tới hạn χ^2 để tìm các xác suất sau đây:

a. $P(\chi^{2(9)} > 2,7)$ b. $P(\chi^{2(15)} > -27,49)$

c. $P(\chi^{2(30)} < 18,49)$ d. $P(\chi^{2(25)} < -44,31)$

19. Cho hai biến ngẫu nhiên độc lập và cùng phân phối theo quy luật khi bình phương với số bậc tự do tương ứng là 6 và 8. Tìm xác suất để tổng của hai biến ngẫu nhiên đó lớn hơn 21,06.

20. Dùng bảng giá trị tới hạn Student để tìm các xác suất sau:

a. $P(T(15) > 2,602)$ b. $P(T(25) > -2,06)$

c. $P(T(30) < 2,75)$ d. $P(T(10) < -2,228)$

21. Dùng bảng giá trị tới hạn Fisher để tìm các xác suất sau:

a. $P(F(3,7) > 5,89)$

b. $P(F(4,9) < 4,72)$

22. Dùng bảng giá trị tới hạn Fisher để tìm các giá trị sau:

a. $F_{0,05}^{(5,7)}$

b. $F_{0,025}^{(5,32)}$

c. $F_{0,95}^{(38,6)}$

Chương IV

BIẾN NGẪU NHIÊN HAI CHIỀU. HÀM CÁC BIẾN NGẪU NHIÊN

§1. KHÁI NIỆM VỀ BIẾN NGẪU NHIÊN NHIỀU CHIỀU

Ở các chương trước ta đã xét các biến ngẫu nhiên mà các giá trị có thể có của chúng được biểu diễn bằng một số. Đó là các biến ngẫu nhiên một chiều.

Ngoài các biến ngẫu nhiên một chiều, trong thực tế ta còn gặp các biến số mà các giá trị có thể có của chúng được xác định bằng hai, ba, ..., n số. Những biến số này được gọi một cách tương ứng là các biến ngẫu nhiên hai chiều, ba chiều, ..., n chiều.

Ta sẽ ký hiệu biến ngẫu nhiên hai chiều là (X, Y) trong đó X và Y được gọi là các thành phần của biến ngẫu nhiên hai chiều mà thực chất mỗi thành phần lại là một biến ngẫu nhiên một chiều. Như vậy, nhiên X và Y được xét một cách đồng thời. Tương tự như vậy, biến ngẫu nhiên n chiều có thể xem xét như hệ của n biến ngẫu nhiên.

Thí dụ. Một máy sản xuất một loại sản phẩm. Nếu kích thước của sản phẩm được đo bằng chiều dài X và chiều rộng

Y thì ta có biến ngẫu nhiên hai chiều, còn nếu tính thêm cả chiều cao Z nữa thì ta có biến ngẫu nhiên ba chiều.

Trong thực tế người ta cũng phân biệt các biến ngẫu nhiên nhiều chiều thành hai loại: Rời rạc và liên tục. Các biến ngẫu nhiên nhiều chiều gọi là rời rạc nếu các thành phần của nó là rời rạc và gọi là liên tục nếu các thành phần của nó là liên tục.

Sau đây ta sẽ chỉ xét các biến ngẫu nhiên hai chiều.

§2. BẢNG PHÂN PHỐI XÁC SUẤT CỦA BIẾN NGẪU NHIÊN HAI CHIỀU

Đối với các biến ngẫu nhiên hai chiều người ta cũng dùng bảng phân phối xác suất, hàm phân bố xác suất và hàm mật độ xác suất để mô tả quy luật phân phối xác suất của chúng.

Bảng phân phối xác suất đồng thời của biến ngẫu nhiên hai chiều rời rạc liệt kê các giá trị có thể có của nó và các xác suất tương ứng. Nó có dạng sau đây:

Y \ X	x_1	x_2	...	x_i	...	x_n
y_1	$P(x_1, y_1)$	$P(x_2, y_1)$	...	$P(x_i, y_1)$	...	$P(x_n, y_1)$
y_2	$P(x_1, y_2)$	$P(x_2, y_2)$	...	$P(x_i, y_2)$	...	$P(x_n, y_2)$
...	...	...	...	...	...	...
y_j	$P(x_1, y_j)$	$P(x_2, y_j)$	...	$P(x_i, y_j)$	...	$P(x_n, y_j)$
...	...	...	...	...	...	...
y_m	$P(x_1, y_m)$	$P(x_2, y_n)$	...	$P(x_i, y_m)$	...	$P(x_n, y_m)$

Trong đó x_i $(i = \overline{1,n})$ là các giá trị có thể có của thành phần X; y_j $(j = \overline{1,m})$ là các giá trị có thể có của thành phần Y; còn $P(x_i, y_j)$ là xác suất đồng thời để biến ngẫu nhiên hai chiều (X, Y) nhận giá trị (x_i, y_j).

Ta chú ý rằng để tạo nên một quy luật phân phối xác suất thì các xác suất đồng thời $P(x_i, y_j)$ phải thỏa mãn điều kiện:

$$\begin{cases} P(x_i, y_j) \geq 0 \qquad \text{với } i = \overline{1,n};\ j = \overline{1,m} \\ \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) = 1 \end{cases} \tag{4.1}$$

Biết bảng phân phối xác suất đồng thời của biến ngẫu nhiên hai chiều bao giờ cũng có thể tìm được bảng phân phối xác suất biên của mỗi thành phần. Bảng phân phối xác suất biên của thành phần X có dạng:

X	x_1	x_2	...	x_i	...	x_n
P	$P(x_1)$	$P(x_2)$	...	$P(x_i)$	...	$P(x_n)$

trong đó:

$$P(x_i) = \sum_{j=1}^{m} P(x_i, y_j) \qquad \text{với } i = \overline{1;n} \tag{4.2}$$

được gọi là xác suất biên của thành phần X.

Rõ ràng là: $\sum_{i=1}^{n} P(x_i) = 1$

Bảng phân phối xác suất biên của thành phần Y có dạng:

Y	y_1	y_2	...	y_j	...	y_m
P	$P(y_1)$	$P(y_2)$	...	$P(y_j)$	...	$P(y_m)$

trong đó:

$$P(y_j) = \sum_{i=1}^{n} P(x_i, y_j) \qquad \text{với } j = \overline{1;m} \tag{4.3}$$

được gọi là xác suất biên của thành phần Y.

Và $\sum_{j=1}^{m} P(y_j) = 1$

Thí dụ. Thu nhập hàng năm của các cặp vợ chồng có bảng phân phối xác suất đồng thời như sau:

Bảng 4.1

X \ Y	10	20	30	40
10	0,2	0,04	0,01	0
20	0,1	0,36	0,09	0
30	0	0,05	0,1	0
40	0	0	0	0,05

Trong đó X là thu nhập của chồng (triệu đồng/năm)

Y là thu nhập của vợ (triệu đồng/năm)

Tìm phân phối biên của mỗi thành phần.

Giải. Cộng các xác suất theo dòng ta thu được các xác suất tương ứng của các giá trị của thành phần X.

$P(X_1) = 0,2 + 0,04 + 0,01 = 0,25$

$P(X_2) = 0,1 + 0,36 + 0,09 = 0,55$

$P(X_3) = 0,05 + 0,1 = 0,15$

$P(X_4) = 0,05$

Vậy phân phối biên của thu nhập của chồng như sau:

X	10	20	30	40
P	0,25	0,55	0,15	0,05

Tiến hành tương tự đối với từng cột của bảng 4.1 ta thu được phân phối biên của thu nhập của vợ như sau:

Y	10	20	30	40
P	0,3	0,45	0,2	0,05

§3. HÀM PHÂN BỐ XÁC SUẤT CỦA BIẾN NGẪU NHIÊN HAI CHIỀU

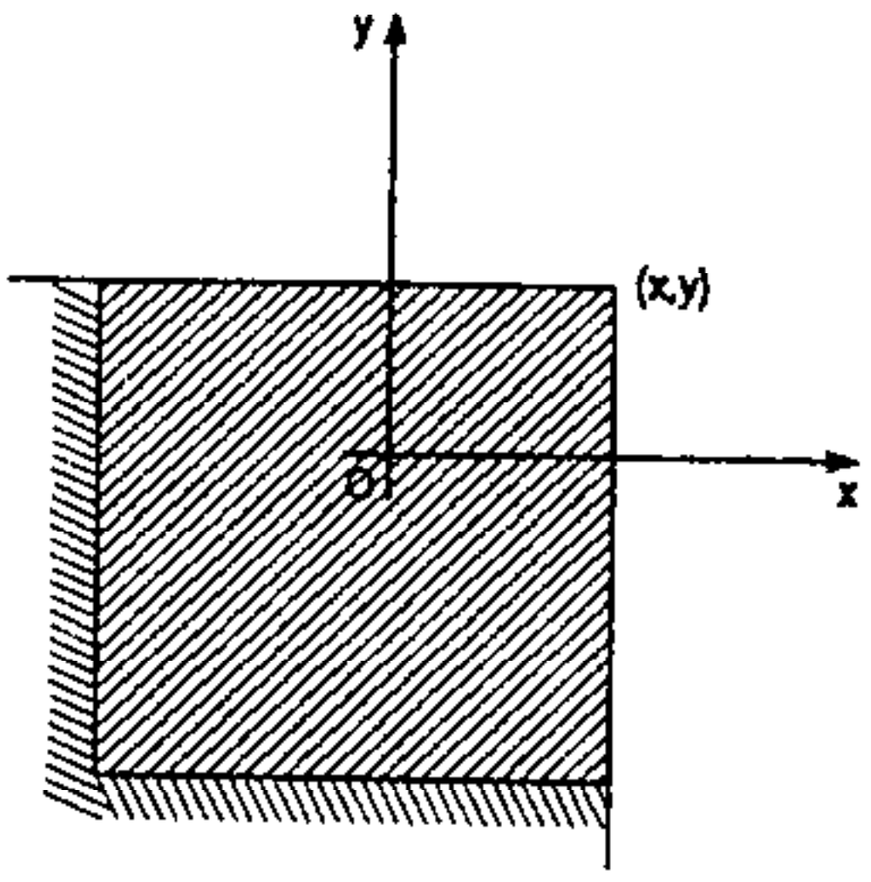
Xét biến ngẫu nhiên hai chiều (X, Y) có thể là rời rạc hoặc liên tục.

Giả sử x và y là một cặp số thực bất kỳ. Xét biến cố (X < x; Y < y) là biến cố để X nhận giá trị nhỏ hơn x và Y nhận giá trị nhỏ hơn y. Hiển nhiên là khi x và y thay đổi thì xác suất của biến cố trên cũng thay đổi theo, tức là nó là một hàm số của x, y.

Hàm phân bố xác suất đồng thời của biến ngẫu nhiên hai chiều (X, Y), ký hiệu là F(x, y) là xác suất để thành phần X nhận giá trị nhỏ hơn x và thành phần Y nhận giá trị nhỏ hơn y với x, y là các số thực tùy ý:

$$F(x, y) = P(X < x, Y < y) \qquad (4.4)$$

Về mặt hình học giá trị của hàm phân bố xác suất đồng thời tại mỗi điểm (x, y) là xác suất để biến ngẫu nhiên (X, Y) nhận giá trị tại một góc phẳng có đỉnh là (x, y) và nằm ở bên dưới và bên trái đỉnh đó.



***Hình 4.1.* Đồ thị hàm F(x, y)**

Thí dụ 1. Tìm xác suất để trong kết quả của phép thử thành phần X của biến ngẫu nhiên hai chiều (X, Y) nhận giá trị $X < 2$ và thành phần Y nhận giá trị $Y < 3$ nếu biết hàm phân bố xác suất của nó có dạng:

$$F(x, y) = \left(\frac{1}{\pi}\text{arctg}\frac{x}{2} + \frac{1}{2}\right) \times \left(\frac{1}{\pi}\text{arctg}\frac{y}{3} + \frac{1}{2}\right)$$

Giải. Theo định nghĩa hàm phân bố xác suất của biến ngẫu nhiên hai chiều, ta có:

$$P(X < 2, Y < 3) = F(2, 3) = \left(\frac{1}{\pi}\text{arctg}\frac{2}{2} + \frac{1}{2}\right) \times \left(\frac{1}{\pi}\text{arctg}\frac{3}{3} + \frac{1}{2}\right)$$

$$= \left(\frac{1}{\pi}\frac{\pi}{4} + \frac{1}{2}\right) \times \left(\frac{1}{\pi}\frac{\pi}{4} + \frac{1}{2}\right) = \frac{3}{4} \cdot \frac{3}{4} = \frac{9}{16}$$

Hàm phân bố xác suất đồng thời của biến ngẫu nhiên hai chiều có các tính chất sau đây:

Tính chất 1. Giá trị của hàm phân bố xác suất đồng thời luôn nằm trong đoạn [0, 1].

Tính chất 2. Hàm phân bố xác suất đồng thời là hàm không giảm theo từng đối số, tức là:

$$F(x_2, y) \geq F(x_1, y) \text{ nếu } x_2 > x_1$$

$$F(x, y_2) \geq F(x, y_1) \text{ nếu } y_2 > y_1$$

Ta sẽ chứng minh rằng F(x, y) là hàm không giảm theo x, còn việc chứng minh nó là hàm không giảm theo y cũng tiến hành tương tự. Chúng ta xét biến cố $(X < x_2, Y < y)$. Biến cố này có thể tách ra thành tổng của hai biến cố xung khắc là $(X < x_1, Y < y)$ và $(x_1 \leq X < x_2, Y < y)$. Do đó theo định lí cộng xác suất ta có:

$$P(X < x_2, Y < y) = P(X < x_1, Y < y) + P(x_1 \leq X < x_2, Y < y).$$

Từ đó:

$$P(X < x_2, Y < y) - P(X < x_1, Y < y) = P(x_1 \leq X < x_2, Y < y)$$

Hay:

$$F(x_2, y) - F(x_1, y) = P(x_1 \leq X < x_2, Y < y) \geq 0$$

Do đó: $F(x_2, y) \geq F(x_1, y)$.

Tính chất 3. Ta có các biểu thức giới hạn sau:

$$F(-\infty, y) = 0;\ F(x, -\infty) = 0$$

$$F(-\infty, -\infty) = 0;\ F(+\infty, +\infty) = 1$$

Thật vậy, $F(-\infty, y)$ là xác suất của biến cố $(X < -\infty).(Y < y)$, song biến cố này là biến cố không thể có, do đó xác suất của nó bằng không. Tương tự như vậy:

$$F(x, -\infty) = 0 \text{ và } F(-\infty, \infty) = 0.$$

Biến cố $(X < +\infty).(Y < +\infty)$ là biến cố chắc chắn, do đó xác suất của nó bằng một: $F(+\infty, +\infty) = 1$.

Tính chất 4. Khi $y = +\infty$ hàm phân bố xác suất đồng thời của hệ hai biến ngẫu nhiên trở thành hàm phân bố xác suất biên của riêng thành phần X:

$$F(x, +\infty) = F_1(x).$$

Và khi $x = +\infty$ hàm phân bố xác suất đồng thời của hai biến ngẫu nhiên trở thành hàm phân bố xác suất biên của riêng thành phần Y

$$F(+\infty, y) = F_2(y)$$

Thật vậy, vì biến cố $(Y < +\infty)$ là biến cố chắc chắn do đó $F(x, +\infty)$ chỉ xác định xác suất của biến cố $X < x$ tức là nó trở thành hàm phân bố xác suất của thành phần X.

Tương tự, biến cố $(X < +\infty)$ là biến cố chắc chắn do đó $F(+\infty, y)$ chỉ còn là xác suất để $Y < y$ và nó trở thành hàm phân bố xác suất của riêng thành phần Y.

Tính chất trên cho phép ta luôn luôn có thể tìm được hàm phân bố xác suất biên của mỗi thành phần của biến ngẫu nhiên hai chiều khi biết hàm phân bố xác suất đồng thời của hệ hai biến ngẫu nhiên.

Từ các tính chất trên ta có thể suy ra một vài công thức sau đây:

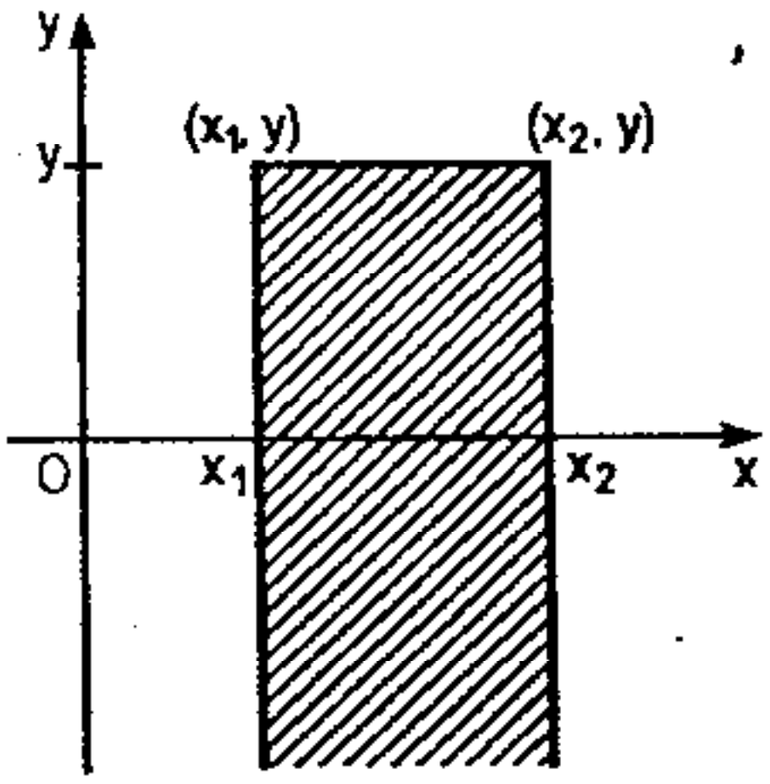
- Xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong một dải $(x_1 < X < x_2, Y < y)$ bằng $P(x_1 < X < x_2, Y < y) = F(x_2, y) - F(x_1, y)$ (hình 4.2) và trong dải $(X < x, y_1 < Y < y_2)$ bằng:

$$P(X < x, y_1 < Y < y_2) = F(x, y_2) - F(x, y_1).$$

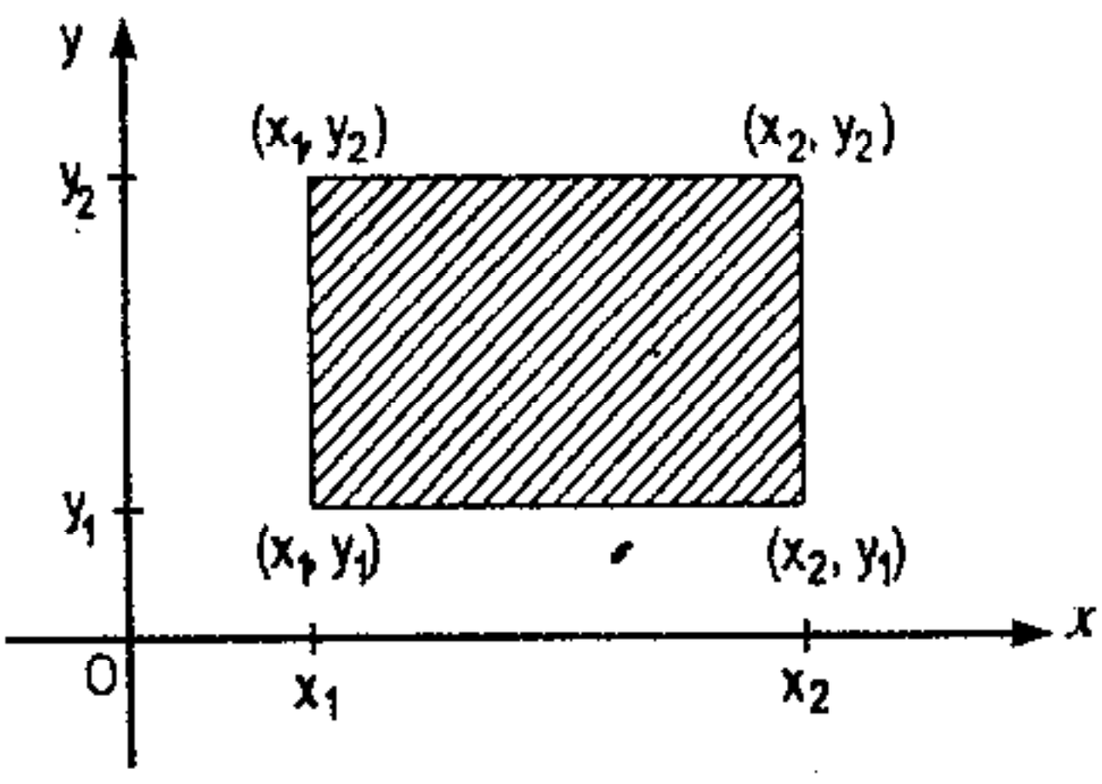
- Xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong hình chữ nhật $(x_1 < X < x_2, y_1 < Y < y_2)$ bằng:

$$P(x_1 < X < x_2, y_1 < Y < y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$$

(hình 4.3).



Hình 4.2



Hình 4.3

Thí dụ 2. Tìm xác suất để biến ngẫu nhiên hai chiều (X, Y) nhận giá trị trong hình chữ nhật giới hạn bởi các đường thẳng

$$x_1 = \frac{\pi}{6}, x_2 = \frac{\pi}{2}, y_1 = \frac{\pi}{4}, y_2 = \frac{\pi}{3}$$

nếu biết hàm phân bố xác suất đồng thời:

$$F(x, y) = \sin x.\sin y \ (0 \le x \le \frac{\pi}{2}, 0 \le y \le \frac{\pi}{2})$$

Giải. Theo công thức tính xác suất để (X, Y) nhận giá trị trong hình chữ nhật, ta có:

$$P\left(\frac{\pi}{6}<X<\frac{\pi}{2}, \frac{\pi}{4}<Y<\frac{\pi}{3}\right) = F\left(\frac{\pi}{2},\frac{\pi}{3}\right) - F\left(\frac{\pi}{2},\frac{\pi}{4}\right) - F\left(\frac{\pi}{6},\frac{\pi}{3}\right) + F\left(\frac{\pi}{6},\frac{\pi}{4}\right)$$

$$= \sin\frac{\pi}{2}.\sin\frac{\pi}{3} - \sin\frac{\pi}{2}.\sin\frac{\pi}{4} - \sin\frac{\pi}{6}.\sin\frac{\pi}{3} + \sin\frac{\pi}{6}.\sin\frac{\pi}{4}$$

$$= \frac{\sqrt{3}}{2} - \frac{\sqrt{2}}{2} - \frac{1}{2}.\frac{\sqrt{3}}{2} + \frac{1}{2}.\frac{\sqrt{2}}{2} = \frac{\sqrt{3}-\sqrt{2}}{4} = 0{,}08$$

§4. HÀM MẬT ĐỘ XÁC SUẤT CỦA BIẾN NGẪU NHIÊN HAI CHIỀU

Đối với biến ngẫu nhiên liên tục (X, Y) **ngoài hàm phân bố xác suất** ra còn có thể dùng hàm **mật độ xác suất** để biểu diễn quy luật phân phối xác suất của nó. **Ta sẽ giả thiết rằng** với biến ngẫu nhiên liên tục (X, Y) **hàm phân bố xác suất** luôn liên tục và có đạo hàm riêng hỗn hợp bậc hai ở mọi đường cong (có thể trừ một số đường cong nhất định).

Hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liên tục (X, Y), ký hiệu là f(x, y), là đạo hàm riêng hỗn hợp bậc 2 của hàm phân bố xác suất đồng thời:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x.\partial y} \qquad (4.5)$$

Về mặt hình học, hàm f(x, y) có thể **xem như** một mặt, được gọi là mặt phân phối xác suất.

Thí dụ 1. Tìm hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liên tục (X, Y) nếu biết hàm phân bố xác suất đồng thời của nó:

$$F(x, y) = \sin x.\sin y \ (0 \le x \le \frac{\pi}{2},\ 0 \le y \le \frac{\pi}{2})$$

ật

Giải. Theo định nghĩa hàm mật độ xác suất đồng thời, trước hết ta tìm đạo hàm riêng của hàm phân bố xác suất đồng thời theo x:

$$\frac{\partial F(x,y)}{\partial x} = \cos x.\sin y$$

Từ kết quả thu được, ta tìm tiếp đạo hàm riêng của nó theo y, ta được:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x.\partial y} = \cos x.\cos y$$

với $(0 \le x \le \frac{\pi}{2}, 0 \le y \le \frac{\pi}{2})$

Hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liên tục có các tính chất sau đây:

Tính chất 1. Hàm mật độ xác suất đồng thời luôn không âm

$$f(x, y) \ge 0$$

Tính chất này được suy ra trực tiếp từ chỗ nó là đạo hàm của hàm không giảm F(x, y) theo các thành phần của nó.

Tính chất 2. Xác suất để biến ngẫu nhiên hai chiều liên tục (X, Y) nhận giá trị trong một miền D được xác định bằng công thức:

$$P[(X, Y) \in D] = \iint_D f(x,y)dxdy$$

Thật vậy, theo tính chất của hàm phân bố xác suất đồng thời, ta có xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong một hình chữ nhật bằng:

$$P(x_1 < X < x_2, y_1 < Y < y_2)$$

$$= [F(x_2, y_2) - F(x_1, y_2)] - [F(x_2, y_1) - F(x_1, y_1)]$$

Gọi vế trái là P_{ABCD} và áp dụng định lí Lagrange đối với vế phải ta có:

$$P_{ABCD} = F''_{xy}(\xi, \eta)\Delta_x\Delta_y$$

Song theo định nghĩa của hàm mật độ xác suất thì:

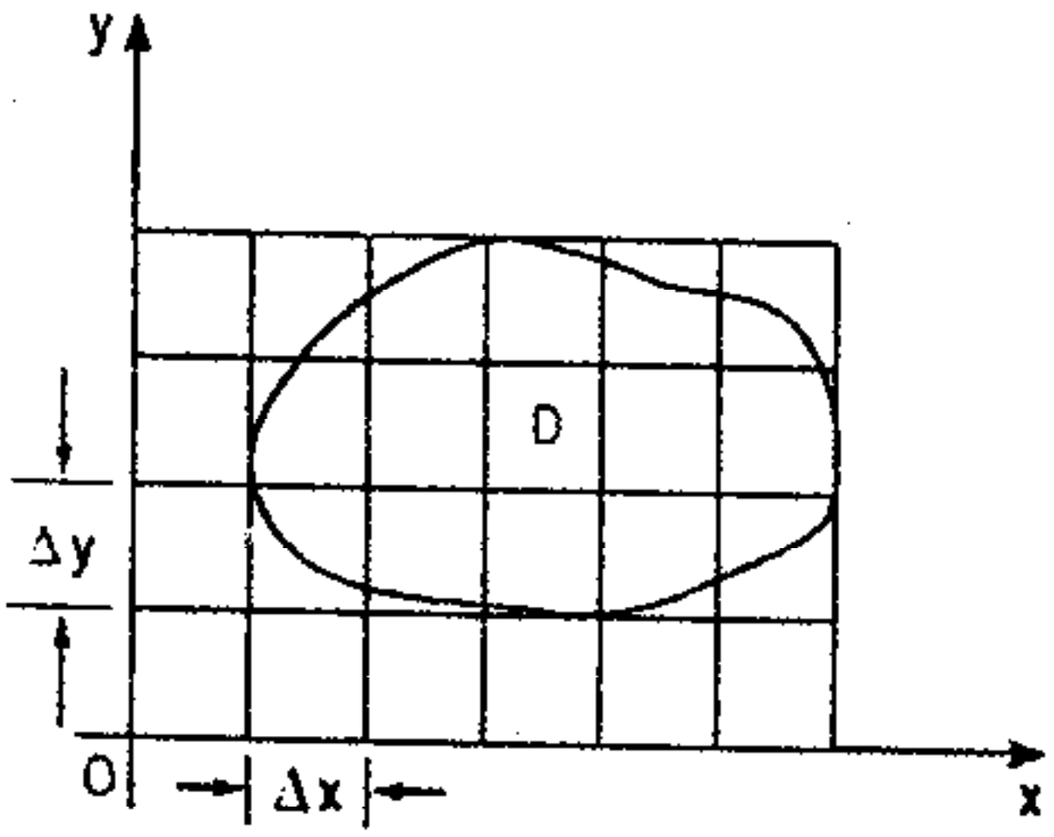
$$F''_{xy}(\xi, \eta) = f(\xi, \eta)$$

Do đó ta có:

$$P_{ABCD} = f(\xi, \eta).\Delta_x\Delta_y$$

Như vậy tích $f(\xi, \eta).\Delta_x\Delta_y$ chính là xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong hình chữ nhật ABCD với các cạnh là Δ_x và Δ_y.

Chia D thành các miền bằng các đường thẳng song song với trục OY cách nhau một đoạn Δ_x và các đường thẳng song song với trục OX cách nhau một đoạn Δ_y (hình 4.4).



Hình 4.4

Vì các biến cố để biến ngẫu nhiên (X, Y) nhận giá trị trong n miền nói trên là xung khắc từng đôi nên xác suất để (X, Y) nhận giá trị trong miền D có thể tính xấp xỉ như sau:

$$P[(X, Y) \in D] \approx \sum_{i=1}^{n} f(\xi_i, \eta_i)\Delta x.\Delta y$$

Lấy giới hạn khi $\Delta_x \to 0$ và $\Delta_y \to 0$ ta có:

$$P[(X, Y) \in D] = \iint_D f(x, y)dxdy$$

Về mặt hình học, biểu thức trên có thể xem như xác suất để biến ngẫu nhiên (X, Y) nhận giá trị trong miền D bằng thể tích của khối giới hạn bởi mặt xác suất f(x, y) mà đáy của nó là hình chiếu của mặt đó trên mặt phẳng X0Y (hình 4.4).

Tính chất 3. Hàm phân bố xác suất đồng thời được xác định thông qua hàm mật độ xác suất đồng thời bằng biểu thức sau:

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy$$

Tính chất này suy ra trực tiếp từ định nghĩa của hàm mật độ xác suất đồng thời.

Tính chất 4. Tích phân suy rộng hai lớp của hàm mật độ xác suất đồng thời bằng một:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$$

Tích phân suy rộng từ $-\infty$ đến $+\infty$ cho thấy các miền lấy tích phân là toàn bộ mặt phẳng x0y. Mà biến cố để biến ngẫu nhiên (X, Y) nhận giá trị trên toàn mặt phẳng x0y là biến cố chắc chắn, nên xác suất của nó bằng một.

Thí dụ 2. Tìm hàm phân bố xác suất đồng thời của biến ngẫu nhiên (X, Y) theo hàm mật độ xác suất đồng thời sau đây:

$$f(x,y) = \frac{1}{\pi^2(1+x^2)(1+y^2)}$$

Giải. Theo tính chất của hàm mật độ xác suất đồng thời, ta có:

$$F(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(x,y)dxdy = \frac{1}{\pi^2}\int_{-\infty}^{y}\frac{1}{1+y^2}\int_{-\infty}^{x}\frac{dx}{1+x^2}dy$$

$$= \frac{1}{\pi^2}\int_{-\infty}^{y}\frac{1}{1+y^2}\left(\text{arctgx}+\frac{\pi}{2}\right)dy = \left(\frac{1}{\pi}\text{arctgx}+\frac{1}{2}\right)\frac{1}{\pi}\int_{-\infty}^{y}\frac{dy}{1+y^2} =$$

$$= \left(\frac{1}{\pi}\text{arctgx}+\frac{1}{2}\right).\left(\frac{1}{\pi}\text{arctgy}+\frac{1}{2}\right)$$

Thí dụ 3. Với hàm mật độ xác suất đồng thời đã cho ở thí dụ trước, tìm xác suất để (X, Y) nhận giá trị trong hình chữ nhật với các đỉnh A(1, 1), B($\sqrt{3}$,1), C(1, 0) và D($\sqrt{3}$,0).

Giải. Theo tính chất của hàm mật độ xác suất đồng thời:

$$P[(X, Y) \in D] = \iint_{D} f(x,y)dxdy = \frac{1}{\pi^2}\int_{0}^{1}\left(\frac{1}{1+y^2}\int_{1}^{\sqrt{3}}\frac{dx}{1+x^2}\right)dy$$

$$= \frac{1}{\pi^2}\text{arctgx}\Big|_{1}^{\sqrt{3}}\int_{0}^{1}\frac{dy}{1+y^2} = \frac{1}{\pi^2}\left(\frac{\pi}{3}-\frac{\pi}{4}\right)\text{arctgy}\Big|_{0}^{1} = \frac{1}{\pi}-\frac{\pi}{12}.\frac{\pi}{4} = \frac{1}{48}$$

Khi đã biết hàm mật độ xác suất đồng thời của hệ hai biến ngẫu nhiên, bao giờ ta cũng có thể xác định được hàm mật độ xác suất biên của từng thành phần của nó. Thực vậy, gọi hàm mật độ xác suất biên của thành phần X là $f_1(x)$ thì theo định nghĩa hàm mật độ xác suất, ta có:

$$f_1(x) = \frac{dF_1(x)}{dx} = \frac{dF(x,\infty)}{dx} = \frac{d}{dx}\int_{-\infty}^{x}\int_{-\infty}^{+\infty} f(x,y)dxdy = \int_{-\infty}^{+\infty} f(x,y)dy$$

Như vậy:
$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy \qquad (4.6)$$

Tương tự, có thể chứng minh được rằng:

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx \qquad (4.7)$$

Như vậy, **hàm mật độ xác suất biên của một thành phần** nào đó bằng tích phân suy rộng của **hàm mật độ xác suất** đồng thời của hệ trong đó biến lấy tích phân là thành phần kia của hệ.

Thí dụ 4. Biến ngẫu nhiên hai chiều (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y)=\begin{cases}\dfrac{1}{6\pi} & \text{với } \dfrac{x^2}{9}+\dfrac{y^2}{4}<1\\[2ex] 0 & \text{với } \dfrac{x^2}{9}+\dfrac{y^2}{4}\geq 1\end{cases}$$

Tìm hàm mật độ xác suất biên của các thành phần X và Y.

Giải. Theo công thức ta có:

$$f_1(x)=\int_{-\infty}^{+\infty} f(x,y)dy=\frac{1}{6\pi}\int_{-2\sqrt{1-\frac{x^2}{9}}}^{2\sqrt{1-\frac{x^2}{9}}} dy=\frac{2}{6\pi}\int_{0}^{2\sqrt{1-\frac{x^2}{9}}} dy=\frac{2}{9\pi}\sqrt{9-x^2}$$

Như vậy:

$$f_1(x)=\begin{cases}\dfrac{2}{9\pi}\sqrt{9-x^2} & \text{với} |x|<3\\[2ex] 0 & \text{với} |x|\geq 3\end{cases}$$

Tương tự, ta tìm được

$$f_2(y)=\begin{cases}\dfrac{1}{2\pi}\sqrt{4-y^2} & \text{với} |y|<2\\[2ex] 0 & \text{với} |y|\geq 2\end{cases}$$

§5. QUY LUẬT PHÂN PHỐI XÁC SUẤT CÓ ĐIỀU KIỆN CỦA CÁC THÀNH PHẦN CỦA HỆ HAI BIẾN NGẪU NHIÊN

Ở các phần trên ta đã thấy quá trình phân tích hệ hai biến ngẫu nhiên thành các thành phần của nó luôn luôn có thể thực hiện được, tức là biết quy luật phân phối xác suất đồng thời của hệ hai biến ngẫu nhiên, ta luôn tìm được quy luật phân phối xác suất biên của từng thành phần của nó.

Vấn đề tổng hợp hai biến ngẫu nhiên một chiều thành hệ hai biến ngẫu nhiên được tiến hành phức tạp hơn. Để làm điều đó ta phải sử dụng khái niệm *phân phối xác suất có điều kiện*.

Xét biến ngẫu nhiên rời rạc (X, Y) trong đó các giá trị có thể có của thành phần X là $x_1, x_2, ..., x_n$, còn các giá trị có thể có của thành phần Y là $y_1, y_2, ..., y_m$. Gọi $P(x_i/y_j)$ $(i = \overline{1,n};\ j = \overline{1,m})$ là xác suất có điều kiện để thành phần X nhận giá trị bằng x_i với điều kiện thành phần Y nhận giá trị bằng y_j.

Bảng phân phối xác suất có điều kiện của thành phần X với điều kiện $Y = y_j$ có dạng:

X/y_j	x_1	x_2	...	x_i	...	x_n
P	$P(x_1/y_j)$	$P(x_2/y_j)$	...	$P(x_i/y_j)$	...	$P(x_n/y_j)$

trong đó các xác suất có điều kiện được tính bằng công thức:

$$P(x_i / y_j) = \frac{P(x_i, x_j)}{P(y_j)} \quad i = \overline{1,n};\ \ j = \overline{1,m} \tag{4.8}$$

Ta chú ý rằng:

$$\sum_{i=1}^{n} P(x_i / y_j) = \sum_{i=1}^{n} \frac{P(x_i, x_j)}{P(y_j)} = \frac{P(y_j)}{P(y_j)} = 1 \qquad j = \overline{1, m}$$

Tức là các xác suất có điều kiện $P(x_i/y_j)$ cũng phải thỏa mãn các yêu cầu của một quy luật phân phối xác suất.

Tương tự, bảng phân phối xác suất có điều kiện của thành phần Y với điều kiện $X = x_i$ có dạng:

X/x_i	y_1	y_2	...	y_j	...	y_m
P	$P(y_1/x_i)$	$P(y_2/x_i)$	...	$P(y_j/x_i)$	...	$P(y_m/x_i)$

Trong đó

$$P(y_i / x_j) = \frac{P(x_i, y_j)}{P(x_i)} \quad i = \overline{1, n};\ j = \overline{1, m} \tag{4.9}$$

và $$\sum_{j=1}^{m} P(y_i / x_j) = \sum_{j=1}^{m} \frac{P(x_i, y_j)}{P(x_i)} = \frac{P(x_i)}{P(x_i)} = 1 \qquad i = \overline{1, n}$$

Thí dụ 1. Phân phối xác suất của lương tháng Y (triệu đồng) và giới tính X của công nhân một công ty như sau:

Bảng 4.2

X \ Y	0,5	1	1,5
Nữ : 0	0,1	0,3	0,2
Nam : 1	0,06	0,18	0,16

Tìm phân phối xác suất của lương tháng của nữ công nhân.

Giải. Trước hết ta tìm $P(x_1) = 0,1 + 0,3 + 0,2 = 0,6$

Từ đó:

$$P(y_1 / x_1) = \frac{P(x_1, y_1)}{P(x_1)} = \frac{0,10}{0,60} = \frac{1}{6}$$

$$P(y_2 / x_1) = \frac{P(x_1, y_2)}{P(x_1)} = \frac{0,30}{0,60} = \frac{1}{2}$$

$$P(y_3 / x_1) = \frac{P(x_1, y_3)}{P(x_1)} = \frac{0,20}{0,60} = \frac{1}{3}$$

Vậy bảng phân phối xác suất của lương tháng của nữ công nhân là:

Y/X=x_1	0,5	1	1,5
P	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Giả sử (X, Y) là biến ngẫu nhiên hai chiều liên tục. Hàm mật độ xác suất có điều kiện của thành phần X với Y = y, ký hiệu f(x/y) là biểu thức:

$$f(x/y) = \frac{f(x,y)}{f_2(y)} = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dx} \qquad (4.10)$$

Tương tự, hàm mật độ xác suất có điều kiện của thành phần Y với X = x, ký hiệu f(y/x) là biểu thức:

$$f(y/x) = \frac{f(x,y)}{f_1(x)} = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy} \qquad (4.11)$$

Ta chú ý rằng cũng như mọi hàm mật độ xác suất khác, các hàm mật độ xác suất có điều kiện cũng thỏa mãn các điều kiện:

$$f(x/y) \geq 0; \int_{-\infty}^{+\infty} f(x/y)dx = 1$$

$$f(y/x) \geq 0; \int_{-\infty}^{+\infty} f(y/x)dy = 1$$

Thí dụ 2. Biến ngẫu nhiên liên tục (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y) = \begin{cases} \dfrac{1}{\pi r^2} & \text{với } x^2 + y^2 < r^2 \\ 0 & \text{với } x^2 + y^2 > r^2 \end{cases}$$

Tìm các hàm mật độ xác suất có điều kiện của các thành phần.

Giải. Theo công thức:

$$f(x/y) = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dx} = \frac{\dfrac{1}{\pi r^2}}{\dfrac{1}{\pi r^2}\int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} 1dr} = \frac{1}{2\sqrt{r^2-y^2}}$$

với $|x| < \sqrt{r^2 - y^2}$ vì $f(x, y) = 0$ với $x^2 + y^2 > r^2$ nên $f(x, y) = 0$

với $|x| > \sqrt{r^2 - y^2}$

Tương tự, ta tìm được:

$$f(y/x) = \begin{cases} \dfrac{1}{2\sqrt{r^2 - x^2}} & \text{với } |y| < \sqrt{r^2 - x^2} \\ 0 & \text{với } |y| > \sqrt{r^2 - x^2} \end{cases}$$

Trên cơ sở các phân phối xác suất có điều kiện, ta có các công thức tổng hợp hệ hai biến ngẫu nhiên theo phân phối xác suất của các thành phần như sau:

- Nếu (X, Y) là biến ngẫu nhiên hai chiều rời rạc thì:

$$P(x_i, y_j) = P(x_i)\, P(y_j/x_i)$$

$$= P(y_j)P(x_i/y_j) \quad i = \overline{1,n},\ j = \overline{1,m} \tag{4.12}$$

- Còn nếu (X, Y) là biến ngẫu nhiên hai chiều liên tục thì:

$$f(x, y) = f_1(x)f\,(y/x) = f_2(y)f(x/y) \tag{4.13}$$

Như vậy phân phối xác suất của hệ hai biến ngẫu nhiên bằng tích giữa phân phối xác suất biên của một thành phần với phân phối xác suất có điều kiện của thành phần còn lại.

- Nếu hai thành phần X và Y độc lập với nhau thì phân phối xác suất có điều kiện cũng bằng phân phối xác suất không điều kiện, lúc đó ta có các công thức:

$$P(x_i, y_j) = P(x_i).P(y_j) \quad i = \overline{1,n},\ j = \overline{1,m} \tag{4.14}$$

nếu (X, Y) là biến ngẫu nhiên rời rạc và

$$f(x, y) = f_1(x)\ .\ f_2(y) \tag{4.15}$$

nếu (X, Y) là biến ngẫu nhiên liên tục.

Chú ý rằng các hệ thức (4.14) và (4.15) là điều kiện cần và đủ để X và Y độc lập.

§6. CÁC THAM SỐ ĐẶC TRƯNG CỦA HỆ HAI BIẾN NGẪU NHIÊN

Đối với hệ hai biến ngẫu nhiên, các tham số đặc trưng cơ bản của nó trước hết là các kỳ vọng toán và phương sai của các thành phần. Các kỳ vọng toán được xác định bằng các công thức sau đây:

$$E(X) = \sum_{i=1}^{n} x_i P(x_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i P(x_i, y_j) \qquad (4.16)$$

$$E(Y) = \sum_{j=1}^{m} y_j P(y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_j P(x_i, y_j) \qquad (4.17)$$

nếu (X, Y) là biến ngẫu nhiên rời rạc, và

$$E(X) = \int_{-\infty}^{+\infty} x f_1(x)dx = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x,y)dxdy \qquad (4.18)$$

$$E(Y) = \int_{-\infty}^{+\infty} y f_2(y)dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x,y)dxdy \qquad (4.19)$$

nếu (X, Y) là biến ngẫu nhiên liên tục. Các phương sai được xác định bằng công thức:

$$V(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 P(x_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i^2 p(x_i, y_j) - [E(X)]^2 \qquad (4.20)$$

$$V(Y) = \sum_{j=1}^{m} [y_j - E(Y)]^2 P(y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_j^2 p(x_i, y_j) - [E(Y)]^2 \qquad (4.21)$$

nếu biến ngẫu nhiên (X, Y) rời rạc, và

$$V(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 f_1(x)dx = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^2 f(x,y)dxdy - [E(X)]^2 \qquad (4.22)$$

$$V(Y) = \int_{-\infty}^{+\infty} [y - E(Y)]^2 f_2(y)dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y^2 f(x,y)dxdy - [E(Y)]^2 \qquad (4.23)$$

nếu (X, Y) liên tục.

Ngoài các tham số trên, người ta còn thường xác định các tham số quan trọng khác là hiệp phương sai và hệ số tương quan.

Hiệp phương sai, ký hiệu Cov(X, Y) của các biến ngẫu nhiên X và Y là kỳ vọng toán của tích các sai lệch của các biến ngẫu nhiên đó với kỳ vọng toán của chúng:

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\} \qquad (4.24)$$

Để tìm hiệp phương sai của các biến ngẫu nhiên rời rạc, người ta thường dùng công thức sau:

$$Cov(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(x_i, y_j) - E(X).E(Y) \qquad (4.25)$$

và đối với các biến ngẫu nhiên liên tục:

$$Cov(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x.yf(x,y)dxdy - E(X).E(Y) \qquad (4.26)$$

Từ định nghĩa trên, ta thấy hiệp phương sai có đơn vị đo lường bằng tích đơn vị đo lường của các biến ngẫu nhiên X và Y. Do đó, hiệp phương sai sẽ có các giá trị khác nhau tùy thuộc vào đơn vị đo lường của các biến đó.

Để khắc phục hạn chế này người ta đưa ra một tham số khác là hệ số tương quan.

Hệ số tương quan, ký hiệu ρ_{xy} là tỷ số giữa hiệp phương sai và tích các độ lệch chuẩn của các biến ngẫu nhiên đó.

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x.\sigma_y} \qquad (4.27)$$

Hệ số tương quan không có đơn vị đo và có các tính chất cơ bản sau đây:

1. $\rho_{xy} = \rho_{yx}$
2. $-1 \le \rho_{xy} \le 1$
3. Nếu X và Y độc lập thì $\rho_{xy} = 0$

4. Nếu $\rho_{xy} = \pm 1$ thì X và Y phụ thuộc hàm số với nhau.

Hiệp phương sai và hệ số tương quan được dùng để đặc trưng cho mức độ chặt chẽ của mối liên hệ phụ thuộc giữa các biến ngẫu nhiên X và Y. Chú ý rằng nếu X và Y độc lập thì $\rho_{xy} = 0$ song điều ngược lại chưa chắc đã đúng, tức là nếu $\rho_{xy} = 0$ thì X và Y có thể độc lập hoặc phụ thuộc ở một dạng thức nào đó. Trong thực tế trường hợp được quan tâm hơn cả là sự phụ thuộc tương quan.

Hai biến ngẫu nhiên gọi là tương quan với nhau nếu hiệp phương sai (cũng tức là hệ số tương quan) khác không và hai biến nói trên gọi là không tương quan nếu hiệp phương sai (cũng tức là hệ số tương quan) bằng không.

Ta chú ý rằng nếu hai biến ngẫu nhiên tương quan với nhau thì cũng phụ thuộc nhau, song điều ngược lại chưa chắc đã đúng, tức là nếu các biến ngẫu nhiên phụ thuộc thì chúng có thể tương quan nhưng cũng có thể không tương quan với nhau.

Thí dụ 1. Với các số liệu của thí dụ 1 (bảng 4.2) hãy cho biết lương tháng của công nhân có tương quan với giới tính của công nhân hay không và mức đô tương quan chặt chẽ đến đâu.

Giải.

Dễ thấy rằng

$$P(x_1, y_1) = 0,1 \neq 0,2.0,06 = P(x_1).P(y_1)$$

do đó X và Y phụ thuộc nhau.

Ta tìm hiệp phương sai giữa X và Y (bảng 4.3)

Bảng 4.3

	0,5	1	1,5	P(x)
0	0,1 0	0,3 0	0,2 0	0,6
1	0,06 0,03	0,18 0,18	0,16 0,24	0,4
P(y)	0,16	0,48	0,36	1

$$\sum_i \sum_j x_i y_j P(x_i, y_j) = 0,03 + 0,18 + 0,24 = 0,45$$

$$E(X) = 0.0,6 + 1.0,4 = 0,4$$

$$E(Y) = 0,5.0,16 + 1.0,48 + 1,5.0,36 = 1,1$$

Từ đó $Cov(X, Y) = 0,45 - 0,4.1,1 = 0,01$

$$V(X) = 0^2.0,6 + 1^2.0,4 - (0,4)^2 = 0,24$$

$\rightarrow \sigma x = 0,4899$

$$V(Y) = 0,5^2.0,16 + 1^2.0,48 + 1,5^2.0,36 - (1,1)^2 = 0,12$$

$\rightarrow \sigma y = 0,3464$

$$\text{Vậy } \rho_{xy} = \frac{0,01}{0,4899.0,3464} = 0,0589$$

Kết quả cho thấy Y và X có tương quan với nhau song mức độ phụ thuộc tương quan không chặt chẽ.

Thí dụ 2. Biến ngẫu nhiên hai chiều (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y) = \begin{cases} \dfrac{1}{6\pi} & \text{nếu } \dfrac{x^2}{9} + \dfrac{y^4}{4} < 1 \\ 0 & \text{nếu } \dfrac{x^2}{9} + \dfrac{y^4}{4} > 1 \end{cases}$$

Hãy chứng tỏ rằng X và Y phụ thuộc và không tương quan với nhau.

Giải. Theo hàm mật độ xác suất đồng thời đã cho có thể tìm được các hàm mật độ xác suất biên của các thành phần như sau:

$$f_1(x) = \begin{cases} \dfrac{2}{9\pi}\sqrt{9-x^2} & \text{nếu } |x| < 3 \\ 0 & \text{nếu } |x| > 3 \end{cases}$$

và

$$f_2(y) = \begin{cases} \dfrac{1}{2\pi}\sqrt{4-y^2} & \text{nếu } |y| < 2 \\ 0 & \text{nếu } |y| > 2 \end{cases}$$

Vì $f(x, y) \neq f_1(x).f_2(y)$ nên X và Y phụ thuộc nhau. Còn để chứng minh rằng X và Y không tương quan với nhau, ta chỉ cần chứng minh rằng $Cov(X, Y) = 0$.

Thật vậy theo công thức ta có:

$$Cov(X,Y) = \int\int xyf(x,y)dx.dy - E(X).E(Y)$$

Vì các hàm mật độ xác suất biên $f_1(x)$ và $f_2(y)$ đối xứng qua các trục tọa độ, do đó $E(X) = 0$ và $E(Y) = 0$, vì thế:

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dx.dy =$$

$$= \frac{1}{6\pi}\int_{\frac{2}{3}\sqrt{9-x^2}}^{\frac{2}{3}\sqrt{9-x^2}} y\left(\int_{\frac{3}{2}\sqrt{4-y^2}}^{\frac{3}{2}\sqrt{4-y^2}} xdx\right)dy$$

Tích phân trong ngoặc bằng không (hàm dưới dấu tích phân là hàm lẻ, cận lấy tích phân đối xứng qua gốc tọa độ), do đó Cov(X, Y) = 0 tức là X và Y không tương quan với nhau.

Với khái niệm hiệp phương sai ta có thể xét thêm tính chất của phương sai của tổng hai biến ngẫu nhiên phụ thuộc. Nếu X và Y là hai biến ngẫu nhiên phụ thuộc thì phương sai của tổng hoặc hiệu các biến ngẫu nhiên đó được xác định bằng biểu thức sau:

$$V(X \pm Y) = V(X) + V(Y) \pm 2\mathrm{Cov}(X, Y) \qquad (4.28)$$

Trong trường hợp tổ hợp tuyến tính của các biến đó ta có công thức

$$V(aX \pm bY) = a^2 V(X) + b^2 V(Y) \pm 2ab\mathrm{Cov}(X, Y) \qquad (4.29)$$

Nếu X và Y là hai biến ngẫu nhiên độc lập thì phương sai của tích được xác định bằng biểu thức

$$V(X.Y) = [E(Y)]^2.V(X) + [E(X)]^2.V(Y) + V(X).V(Y) \qquad (4.30)$$

Thí dụ 3. Lãi suất hàng năm của trái phiếu T và cổ phiếu S của một công ty có bảng phân phối xác suất như sau (bảng 4.4):

Bảng 4.4

	-10%	0	10%	20%	P(T)
6%	0	0	0,1	0,1	0,2
8%	0	0,1	0,3	0,2	0,6
10%	0,1	0,1	0	0	0,2
P(S)	0,1	0,2	0,4	0,3	1

Nếu muốn đầu tư tiền vào cả trái phiếu và cổ phiếu thì nên đầu tư theo tỉ lệ bao nhiêu để:

a - Lãi suất kỳ vọng thu được là lớn nhất

b - Độ rủi ro về lãi suất là nhỏ nhất.

Giải. Theo bảng phân phối xác suất đồng thời ta tìm được:

$$E(T) = 6.0,2 + 8.0,6 + 10.0,2 = 8\%$$

$$V(T) = 6^2.0,2 + 8^2.0,6 + 10^2.0,2 - 8^2 = 1,6$$

$$\sigma_T = 1,2649\%$$

$$E(S) = -10.0,1 + 0.0,2 + 10.0,4 + 20.0,3 = 9\%$$

$$V(S) = (-10)^2.0,1 + 0^2.0,2 + 10^2.0,4 + 20^2.0,3 - 9^2 = 89$$

$$\sigma_S = 9,4339\%$$

Từ đó $Cov(S, T) = -8$

a. Gọi p là tỷ lệ đầu tư cho trái phiếu ($0 \le p \le 1$) và gọi X là lãi suất thu được khi đầu tư cho cả trái phiếu và cổ phiếu thì

$$X = pT + (1 - p)S$$

Từ đó

$E(X) = p.E(T) + (1 - p).E(S) = 8p + 9(1 - p) = -p + 9$

Hiển nhiên E(X) sẽ đạt cực đại khi p = 0 tức là khi đầu tư toàn bộ tiền cho cổ phiếu.

b. Độ rủi ro được đặc trưng bởi phương sai hoặc độ lệch chuẩn của X.

Ta có $V(X) = p^2V(T) + (1 - p)^2 V(S) + 2p(1 - p).Cov(S, T)$

$= 106,6p^2 - 194p + 89$

V(X) sẽ đạt cực tiểu khi p = 0,9099%.

§7. KỲ VỌNG TOÁN CÓ ĐIỀU KIỆN - HÀM HỒI QUY

Giá trị của hệ số tương quan chỉ cho ta biết mức đô chặt chẽ của sự phụ thuộc tương quan tuyến tính giữa X và Y. Để biểu diễn sự phụ thuộc tương quan này người ta sử dụng các hàm hồi quy.

Trước hết ta xét khái niệm kỳ vọng toán có điều kiện. *Kỳ vọng toán có điều kiện* của biến ngẫu nhiên rời rạc Y với X = x (x là một giá trị xác định của X) là tổng các tích giữa các giá trị có thể có của Y với các xác suất có điều kiện tương ứng:

$$E(Y / X = x) = \sum_{j=1}^{m} y_j P(y_j / x) \qquad (4.31)$$

Đối với các biến ngẫu nhiên liên tục, kỳ vọng toán có điều kiện được xác định bằng công thức:

$$E(Y / X = x) = \int_{-\infty}^{+\infty} yf(y / x)dy \qquad (4.32)$$

Trong đó f(y/x) là hàm mật độ xác suất có điều kiện của Y với X = x. Tương tự, ta có định nghĩa kỳ vọng toán có điều kiện của X khi Y = y

$$E(X/Y=y)=\sum_{i=1}^{n} x_i P(x_i/y) \qquad (4.33)$$

đối với biến ngẫu nhiên rời rạc và

$$E(X/Y=y)=\int_{-\infty}^{+\infty} xf(x/y)dx$$

đối với biến ngẫu nhiên liên tục. Hàm hồi quy của Y đối với X là kỳ vọng toán có điều kiện của Y đối với X:

$$f(x)=E(Y/x) \qquad (4.34)$$

Tương tự, hàm hồi quy của X đối với Y là kỳ vọng toán có điều kiện của X đối với Y :

$$f(y)=E(X/y) \qquad (4.35)$$

Các hàm hồi quy cho biết giá trị trung bình của biến ngẫu nhiên này phụ thuộc vào biến kia như thế nào.

Thí dụ. Thống kê dân số của một nước ở độ tuổi trưởng thành theo trình độ học vấn X và lứa tuổi Y thu được kết quả sau (bảng 4.5)

Tìm học vấn trung bình theo lứa tuổi.

Bảng 4.5

	25-35 30	35-55 45	55-100 70
Thất học: 0	0,01	0,02	0,05
Tiểu học: 1	0,03	0,06	0,10
Trung học: 2	0,18	0,21	0,15
Đại học : 3	0,07	0,08	0,04

Giải. Học vấn trung bình theo lứa tuổi là kỳ vọng toán có điều kiện của X theo Y. Với Y = 30 ta có bảng phân phối xác suất có điều kiện sau:

X/Y=30	0	1	2	3
P	$\frac{0,01}{0,29}$	$\frac{0,03}{0,29}$	$\frac{0,18}{0,29}$	$\frac{0,07}{0,29}$

Từ đó

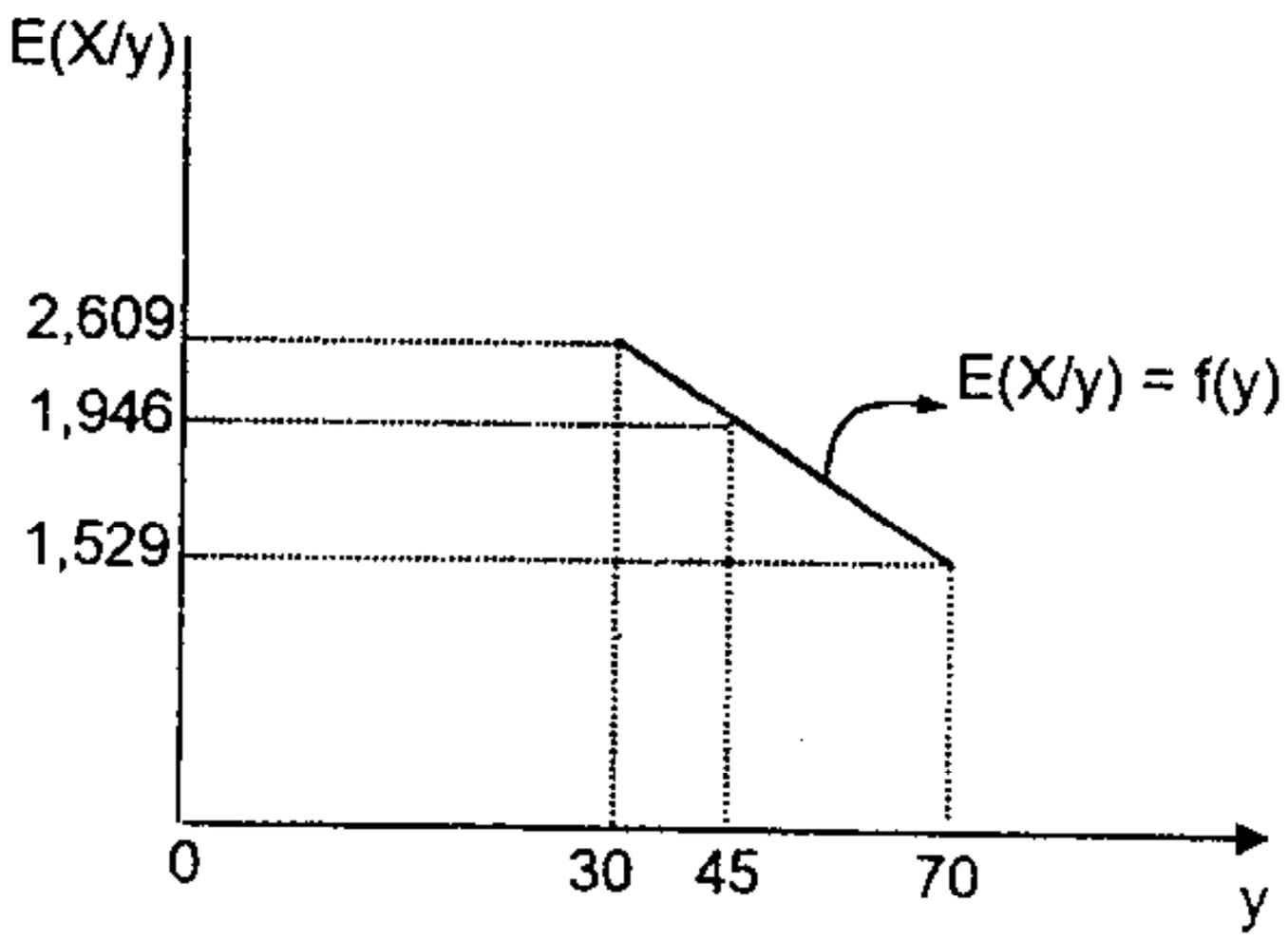
$$E(X/Y=30) = 2,069$$

Tương tự

$$E(X/Y=45) = 1,946$$

$$E(X/Y=70) = 1,529$$

Hàm hồi quy E(X/y) = f(y) có thể mô tả trên đồ thi sau (hình 4.5).



Hình 4.5

§8. PHÂN PHỐI CHUẨN HAI CHIỀU

Đối với các biến *ngẫu nhiên hai chiều* thì trường hợp phổ biến trong thực tế là phân phối theo quy luật chuẩn.

Biến ngẫu nhiên (X, Y) gọi là *phân phối theo quy luật chuẩn hai chiều* nếu hàm mật độ xác suất đồng thời của nó có dạng

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} . \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \right.\right.$$

$$\left.\left. - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\} \qquad (4.36)$$

Ta thấy rằng quy luật chuẩn hai chiều có 5 tham số đặc trưng là μ_x, μ_y, σ_x, σ_y và ρ. Có thể chứng minh được rằng μ_x và μ_y chính là các kỳ vọng toán của các thành phần X và Y. Còn σ_x và σ_y tương ứng là các độ lệch chuẩn, ρ là hệ số tương quan giữa X và Y.

Đặc điểm của biến ngẫu nhiên phân phối theo quy luật chuẩn hai chiều là nếu các thành phần của nó không tương quan thì chúng cũng độc lập với nhau. Thật vậy, giả sử X và Y không tương quan với nhau, lúc đó $\rho = 0$ và biểu thức của hàm mật độ xác suất có dạng:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]}$$

$$= \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} . \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}} = f_1(x).f_2(y) \quad (4.37)$$

Như vậy, nếu các thành phần của biến ngẫu nhiên hai chiều phân phối chuẩn mà không tương quan với nhau thì hàm mật độ xác suất đồng thời của hệ hai biến ngẫu nhiên ấy bằng tích các hàm mật độ xác suất biên, từ đó suy ra tính độc lập của các thành phần đó. Điều ngược lại cũng đúng như vậy.

Như vậy là trong quy luật chuẩn hai chiều khái niệm độc lập và không tương quan là tương đương với nhau.

§9. QUY LUẬT PHÂN PHỐI XÁC SUẤT CỦA HÀM CÁC BIẾN NGẪU NHIÊN

Trong thực tế người ta thường gặp trường hợp một biến ngẫu nhiên là hàm số của một hoặc nhiều biến ngẫu nhiên khác. Lúc đó, khi biết quy luật phân phối xác suất của các đối số, ta có thể tìm được quy luật phân phối xác suất của hàm số tương ứng.

9.1. Quy luật phân phối xác suất của hàm một biến ngẫu nhiên

Nếu mỗi giá trị có thể có của biến ngẫu nhiên X tương ứng với một giá trị có thể có của biến ngẫu nhiên Y thì Y được gọi là hàm của biến ngẫu nhiên X:

$$Y = \varphi(X) \qquad (4.38)$$

Giả sử X là biến ngẫu nhiên rời rạc thì ứng với các giá trị khác nhau của X ta có các giá trị khác nhau của Y và xác suất tương ứng với các giá trị đó bằng nhau.

Thí dụ 1. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	2	3
P	0,6	0,4

Tìm quy luật phân phối xác suất của $Y = X^2$.

Giải. Ta tìm các giá trị có thể có của Y:

$$y_1 = 2^2 = 4 ; \qquad y_2 = 3^2 = 9$$

Vậy bảng phân phối xác suất của Y có dạng:

Y	4	9
P	0,6	0,4

Nếu trong số các giá trị có thể có của Y có các giá trị giống nhau thì phải cộng các xác suất tương ứng lại.

Thí dụ 2. Biến ngẫu nhiên X có bảng phân phối xác suất như sau:

X	–2	2	3
P	0,4	0,5	0,1

Tìm quy luật phân phối xác suất của $Y = X^2$.

Giải. Xác suất tương ứng với giá trị $y_1 = 4$ bằng tổng xác suất của các biến cố xung khắc $x = -2$ và $x = 2$ tức là bằng $0,4 + 0,5 = 0,9$; xác suất để $y_2 = 9$ bằng 0,1 do đó Y có bảng phân phối xác suất như sau:

Y	4	9
P	0,9	0,1

Giả sử biến ngẫu nhiên X là liên tục với hàm mật độ xác suất f(x) đã biết và giả sử $Y = \varphi(X)$. Có thể chứng minh được

rằng nếu $Y = \varphi(X)$ là khả vi, đơn điệu tăng hoặc đơn điệu giảm, có hàm ngược là $X = \psi(Y)$ thì hàm mật độ xác suất $g(y)$ của biến ngẫu nhiên Y được xác định bằng biểu thức:

$$g(y) = f[\psi(y)]\,|\psi'(y)| \qquad (4.39)$$

Thí dụ 3. Biến ngẫu nhiên X phân phối chuẩn với kỳ vọng toán $\mu = 0$. Tìm quy luật phân phối xác suất của biến ngẫu nhiên $Y = X^3$.

Giải. Vì hàm $y = X^3$ là khả vi và đơn điệu tăng, do đó có thể áp dụng công thức:

$$g(y) = f[\psi(y)]\,|\psi'(y)|$$

Ta tìm hàm ngược của hàm $y = x^3$

$$\psi(y) = x = y^{1/3}$$

Ta tìm $f[\psi(y)]$. Theo điều kiện

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{x^2}{2\sigma^2}}$$

do đó:

$$f[\psi(y)] = f[y^{1/3}] = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-y^{2/3}}{2\sigma^2}}$$

Ta tìm đạo hàm của hàm ngược $\psi(y)$

$$|\psi'(y)| = |(y^{1/3})'| = \left|\frac{1}{3y^{2/3}}\right|$$

Từ đó ta có hàm mật độ xác suất của biến ngẫu nhiên Y như sau:

$$g(y) = \frac{1}{|3\sigma y^{2/3}|\sqrt{2\pi}} e^{\frac{y^{2/3}}{2\sigma^2}}$$

Thí dụ 4. Cho biến ngẫu nhiên X phân phối chuẩn với kì vọng μ. Tìm quy luật phân phối xác suất của biến ngẫu nhiên: $Y = aX + b$ trong đó a và b là các hệ số.

Giải. Vì hàm số $Y = aX + b$ là khả vi và đơn điệu, do đó có thể áp dụng công thức:

$$g(y) = f[\psi(y)]\,|\psi'(y)|$$

Hàm ngược $\psi(y)$ có dạng:

$$\psi(y) = x = \frac{y-b}{a}$$

Vậy $$|\psi'(y)| = \left|\frac{1}{a}\right| = \frac{1}{|a|}$$

$$f[\psi(y)] = f\left[\frac{y-b}{a}\right]; g(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-\left(\frac{y-b}{a} - \mu\right)^2}{2(\sigma)^2}\right] \cdot \frac{1}{|a|}$$

$$= \frac{1}{\sigma|a|\sqrt{2\pi}} \exp\left\{\frac{-[y-(a\mu+b)]^2}{2(a\sigma)^2}\right\}$$

Như vậy hàm mật độ xác suất g(y) có dạng:

$$g(y) = \frac{1}{\sigma|a|\sqrt{2\pi}} \exp\left\{\frac{-[y-(a\mu+b)]^2}{2(a\sigma)^2}\right\}$$

Đó lại chính là hàm mật độ xác suất của quy luật chuẩn với các tham số đặc trưng là $E(Y) = a\mu + b$ và $V(Y) = (|a|\sigma)^2$.

Như vậy, ta đã chứng tỏ được rằng nếu biến ngẫu nhiên X phân phối chuẩn thì một hàm tuyến tính bất kỳ của nó cũng phân phối theo quy luật chuẩn.

9.2. Quy luật phân phối xác suất của hàm hai biến ngẫu nhiên

Nếu ứng với mỗi cặp giá trị có thể có của biến ngẫu nhiên X và Y có một giá trị có thể có của Z thì Z được gọi là hàm của hai biến ngẫu nhiên $Z = \varphi(X, Y)$.

Để tìm quy luật phân phối xác suất của Z ta xét một trường hợp cụ thể là $Z = X + Y$ khi đã biết các quy luật phân phối xác suất của X và Y.

Giả sử X và Y là các biến ngẫu nhiên rời rạc và độc lập với nhau. Để xây dựng quy luật phân phối xác suất của $Z = X + Y$ ta phải tìm tất cả các giá trị có thể có của Z và các xác suất tương ứng.

Thí dụ 5. Các biến ngẫu nhiên X và Y rời rạc độc lập với nhau và có các bảng phân phối xác suất như sau:

X	1	2
P	0,4	0,6

Y	3	4
P	0,2	0,8

Tìm quy luật phân phối xác suất của $Z = X + Y$.

Giải. Các giá trị có thể có của Z là tổng của mỗi giá trị có thể có của X với mỗi giá trị có thể có của Y:

$$Z_1 = 1 + 3 = 4; \qquad Z_2 = 1 + 4 = 5;$$

$$Z_3 = 2 + 3 = 5; \qquad Z_4 = 2 + 4 = 6;$$

Ta đi tìm các xác suất tương ứng. Xác suất để $Z_1 = 4$ là xác suất để X nhận giá trị bằng 1 và Y nhận giá trị bằng 3. Vì X và Y độc lập, do đó theo định lý nhân xác suất, ta có:

$$P(Z_1 = 4) = P(x = 1).P(y = 3) = 0,4.0,2 = 0,08$$

Tương tự, ta được:

$$P(Z_2 = 5) = P(x = 1).P(y = 4) = 0,4.0,8 = 0,32$$

$$P(Z_3 = 5) = P(x = 2).P(y = 3) = 0,6.0,2 = 0,12$$

$$P(Z_4 = 6) = P(x = 2).P(y = 4) = 0,6.0,8 = 0,48$$

Vì hai giá trị Z_2 và Z_3 đều bằng 5 do đó xác suất tương ứng của chúng phải được cộng lại:

$$P(Z = 5) = 0,32 + 0,12 = 0,44$$

Vậy Z có bảng phân phối xác suất như sau:

Z	4	5	6
P	0,08	0,44	0,48

Giả sử X và Y là các biến ngẫu nhiên liên tục. Có thể chứng minh được rằng khi X và Y độc lập thì hàm mật độ xác suất g(z) của tổng Z = X + Y (với điều kiện là khi mật độ xác suất của ít nhất một trong hai đối số xác định trong khoảng $(-\infty ; +\infty)$ bằng một biểu thức) có thể tìm được theo công thức:

$$g(z) = \int_{-\infty}^{z} f_1(x)f_2(z - x)dx \qquad (4.40)$$

hoặc

$$g(z) = \int_{-\infty}^{z} f_1(z - y)f_2(y)dy \qquad (4.41)$$

Trong đó f_1 và f_2 tương ứng là các hàm mật độ xác suất biên của X và Y.

Thí dụ 6. Các biến ngẫu nhiên X và Y độc lập, có các hàm mật độ xác suất như sau:

$$f_1(x) = \frac{1}{3}.e^{-x/3} \text{ với } x \geq 0$$

$$f_2(y) = \frac{1}{4}.e^{-y/4} \text{ với } y \geq 0$$

Tìm quy luật phân phối xác suất của Z = X + Y.

Giải. Vì các giá trị có thể có của X và Y không âm, do đó hàm mật độ xác suất g(z) được xác định bằng biểu thức:

$$g(z) = \int_0^z f_1(x) f_2(z-x)dx = \int_0^z \left[\frac{1}{3}e^{-\frac{x}{3}}\right]\left[\frac{1}{4}e^{-\frac{z-x}{4}}\right]dx$$

$$= \frac{1}{12}e^{-\frac{z}{4}} \int_0^z e^{\frac{x}{12}}dx = e^{-\frac{z}{4}}\left(1 - e^{-\frac{z}{12}}\right)$$

Ta chú ý rằng ở đây $Z \geq 0$ vì Z = X + Y mà theo đầu bài thì X và Y đều không âm.

9.3. Các tham số đặc trưng của hàm các biến ngẫu nhiên

Ở trên ta đã xét các phương pháp để tìm quy luật phân phối xác suất của hàm các biến ngẫu nhiên. Khi ta đã biết được quy luật phân phối xác suất dưới dạng bảng phân phối xác suất hoặc hàm mật độ xác suất thì việc tìm các tham số đặc trưng không còn trở ngại gì. Song như đã thấy việc tìm quy luật phân phối xác suất của các biến ngẫu nhiên nhiều khi khá phức tạp. Mặt khác, trong thực tế có nhiều trường hợp ta chỉ quan tâm đến các tham số đặc trưng của hàm các biến ngẫu nhiên chứ không quan tâm đến bản thân quy luật phân phối xác suất của nó. Trong mục này ta xét cách xác định các tham số đặc trưng của hàm các biến ngẫu nhiên mà không cần xác định trước quy luật phân phối xác suất của nó. Giả sử có biến ngẫu nhiên X rời rạc với bảng phân phối xác suất đã biết:

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

Ta phải tìm kỳ vọng toán và phương sai của biến ngẫu nhiên $Y = \varphi(X)$. Các tham số đặc trưng này được xác định bằng các công thức sau đây:

$$E(Y) = E[\varphi(X)] = \sum_{i=1}^{n} \varphi(x_i)p_i \qquad (4.42)$$

$$V(Y) = V[\varphi(X)] = \sum_{i=1}^{n} \{\varphi(x_i) - E[\varphi(x_i)]\}^2 p_i =$$

$$= \sum_{i=1}^{n} \varphi^2(x_i)p_i - \{E[\varphi(x)]\}^2 \qquad (4.43)$$

Thí dụ 7. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

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

Tìm kỳ vọng toán và phương sai của biến ngẫu nhiên

$$Y = \varphi(X) = X^2 + 1$$

Giải. Các giá trị có thể có của Y là:

$$\varphi(1) = 1^2 + 1 = 2 \; ; \varphi(3) = 3^2 + 1 = 10;$$

$$\varphi(5) = 5^2 + 1 = 26.$$

Do đó:

$$E(Y) = E(X^2 + 1) = 2.0,2 + 10.0,5 + 26.0,3 = 13,2$$

$$V(Y) = 2^2.0,2 + 10^2.0,5 + 26^2.0,3 - [13,2]^2 = 79,36$$

Nếu X là biến ngẫu nhiên liên tục với hàm mật độ xác

suất là f(x) thì kỳ vọng toán và phương sai của biến ngẫu nhiên Y = φ(X) được xác định bằng các công thức:

$$E(Y) = E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x).f(x)dx \qquad (4.44)$$

$$V(Y) = V[\varphi(x)] = \int_{-\infty}^{+\infty} \{\varphi(x) - E[\varphi(X)]\}^2 f(x)dx$$

$$= \int_{-\infty}^{+\infty} \varphi^2(x)f(x)dx - \{E[\varphi(X)]\}^2 \qquad (4.45)$$

Thí dụ 8. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \sin x & x \in \left(0, \frac{\pi}{2}\right) \\ 0 & x \notin \left(0, \frac{\pi}{2}\right) \end{cases}$$

Tìm kỳ vọng toán và phương sai của hàm $Y = X^2$.

Giải. Vì X là biến ngẫu nhiên liên tục có hàm mật độ xác suất khác không trong khoảng $(0, \frac{\pi}{2})$ do đó:

$$E(Y) = E(X^2) = \int_0^{\frac{\pi}{2}} x^2 \sin x dx$$

Lấy tích phân từng phần ta được

$$E(Y) = \pi - 2$$

$$V(Y) = V(X^2) = \int_0^{\frac{\pi}{2}} x^4 \sin x dx - (\pi - 2)^2 = \frac{\pi^3}{2} - \pi^2 - 8\pi + 20$$

Tương tự, có thể xây dựng các công thức tìm các tham số đặc trưng của hàm hai biến ngẫu nhiên Z = (X.Y). Chẳng hạn kỳ vọng toán E(Z) được xác định bằng các công thức sau:

Nếu X và Y là các biến ngẫu nhiên rời rạc thì;

$$E(Z) = E[(X.Y)] = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i, y_j)P_{ij} \qquad (4.46)$$

Còn nếu X và Y là các biến ngẫu nhiên liên tục thì:

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

Các ký hiệu và công thức cơ bản

* Xác suất đồng thời của biến ngẫu nhiên (X, Y) rời rạc

$$p(x_i, y_j) = P[(X = x_i)(Y = y_j)] \quad i = \overline{1,n}; \quad j = \overline{1,m}$$

* Xác suất biên của thành phần X

$$P(x_i) = \sum_{j=1}^{m} P(x_i, y_j) \qquad i = \overline{1,n}$$

* Xác suất biên của thành phần Y

$$P(y_j) = \sum_{i=1}^{n} P(x_i, y_j) \qquad j = \overline{1,m}$$

* Hàm phân bố xác suất đồng thời của (X, Y)

$$F(x, y) = P(X < x, Y < y)$$

* Hàm phân bố xác suất biên của X

$$F_1(x) = F(x, +\infty)$$

*** Hàm phân bố xác suất biên của Y**

$$F_2(y) = F(+\infty, y)$$

* $P(x_1 < X < x_2, y_1 < Y < y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$

*** Hàm mật độ xác suất đồng thời của biến ngẫu nhiên (X, Y) liên tục**

$$f(x, y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \Rightarrow \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$$

* $P[(X, Y) \in D] = \iint_D f(x,y)dxdy$

*** Hàm mật độ xác suất biên của X**

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

*** Hàm mật độ xác suất biên của Y**

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

*** Xác suất có điều kiện của (X, Y) rời rạc**

$$P(x_i/y_j) = \frac{P(x_i, y_j)}{P(y_j)} \qquad i = \overline{1,n}; \quad j = \overline{1,m}$$

$$P(y_j/x_i) = \frac{P(x_i, y_j)}{P(x_i)} \qquad i = \overline{1,n}; \quad j = \overline{1,m}$$

*** Hàm mật độ xác suất có điều kiện của (X, Y) liên tục**

$$f(x/y) = \frac{f(x,y)}{f_2(y)} \quad \text{và} \quad f(y/x) = \frac{f(x,y)}{f_1(x)}$$

* Các tham số đặc trưng của biến ngẫu nhiên hai chiều

+ Kỳ vọng toán

$$E(X) = \sum_{i=1}^{n} x_i P(x_i)$$

$$E(Y) = \sum_{j=1}^{m} y_j P(y_j) \qquad \text{nếu (X, Y) rời rạc}$$

$$E(X) = \int_{-\infty}^{+\infty} x f_1(x)dx$$

$$E(Y) = \int_{-\infty}^{+\infty} y f_2(y)dy \qquad \text{nếu (X, Y) liên tục}$$

+ Phương sai

$$V(X) = \sum_{i=1}^{n} x_i^2 P(x_i) - [E(X)]^2$$

$$V(Y) = \sum_{j=1}^{m} y_j^2 P(y_j) - [E(Y)]^2 \qquad \text{nếu (X, Y) rời rạc}$$

$$V(X) = \int_{-\infty}^{+\infty} x^2 f_1(x)dx - [E(X)]^2$$

$$V(Y) = \int_{-\infty}^{+\infty} y^2 f_2(y)dy - [E(Y)]^2 \qquad \text{nếu (X, Y) liên tục}$$

+ Hiệp phương sai

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

$$= E(X.Y) - E(X).E(Y)$$

trong đó $E(X.Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j P(x_i y_j)$

nếu (X, Y) rời rạc và $E(X.Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x.y.f(x,y)dxdy$

nếu (X, Y) liên tục.

+ Hệ số tương quan

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$$

* Kỳ vọng toán có điều kiện

$$E(Y/x_i) = \sum_{j=1}^{m} y_j P(y_j / x_i)$$

$$E(X/y_j) = \sum_{i=1}^{n} x_i P(x_i / y_j) \quad \text{nếu (X, Y) rời rạc}$$

$$E(Y/x) = \int_{-\infty}^{+\infty} yf(y/x)dy$$

$$E(X/y) = \int_{-\infty}^{+\infty} xf(x/y)dx \quad \text{nếu (X, Y) liên tục}$$

* Hàm hồi quy của Y đối với X

$$E(Y/x) = f(x)$$

* Hàm hồi quy của X đối với Y

$$E(X/y) = g(y)$$

* Phân phối chuẩn hai chiều

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X.\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}$$

* Hàm các biến ngẫu nhiên

+ Nếu X là biến ngẫu nhiên rời rạc và $Y = \varphi(X)$ thì bảng phân bố xác suất của Y là:

Y	$\varphi(x_1)$	$\varphi(x_2)$	...	$\varphi(x_n)$
P	p_1	p_2	...	p_n

+ Nếu X là biến ngẫu nhiên liên tục, $Y = \varphi(X)$ và hàm ngược là $X = \psi(Y)$ thì hàm mật độ xác suất

$$g(y) = f[\psi(y)].|\ \psi'(y)|$$

+ Các tham số đặc trưng

$$E(Y) = E[\varphi(X)] = \sum_{i=1}^{n} \varphi(x_i)P_i$$

$$V(Y) = V[\varphi(X)] = \sum_{i=1}^{n} \varphi^2(x_i)P_i - \{E[\varphi(X)]\}^2$$

* Các tính chất của phương sai

+ $V(aX \pm bY) = a^2.V(X) + b^2.V(Y) \pm 2abCov(X, Y)$

+ Nếu X và Y độc lập thì

$$V(X.Y) = [E(Y)]^2.V(X) + [E(X)]^2.V(Y) + V(X).V(Y)$$

Câu hỏi ôn tập

1. Cho một ví dụ thực tế trong kinh tế hoặc kinh doanh về:

a. Một biến ngẫu nhiên nhiều chiều rời rạc.

b. Một biến ngẫu nhiên nhiều chiều liên tục.

2. Cho bảng phân phối xác suất đồng thời của biến ngẫu nhiên rời rạc (X, Y) như sau:

Y \ X	-1	0	1	Σ
0		0,25	0,15	
1	0,15	0,2		0,45
Σ				

a. Hãy điền vào các giá trị còn thiếu trong bảng.

b. Tìm các phân phối biên của X và Y

c. Tìm phân phối có điều kiện của X khi Y = 1

3. Cho biến ngẫu nhiên hai chiều rời rạc (X, Y). Nếu X và Y độc lập thì các dòng của bảng phân phối xác suất đồng thời có tỉ lệ với nhau không? Câu hỏi tương tự đối với các cột của bảng.

4. Cho hai biến ngẫu nhiên rời rạc X và Y độc lập nhau và có các bảng phân phối xác suất như sau:

X	1	2	3
P_X	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{5}{12}$

Y	-2	-1
P_Y	$\frac{1}{3}$	$\frac{2}{3}$

Hãy tìm bảng phân phối xác suất đồng thời của (X, Y).

5. Hãy cho biết các mệnh đề sau đây là đúng hay sai? Giải thích:

a. Nếu $P(x_i, y_j) = P(x_i).P(y_j)$ $\forall i, j$ thì X và Y độc lập.

b. Nếu X và Y độc lập thì $\rho_{XY} = 0$

c. Nếu $\rho_{XY} = 0$ thì X và Y độc lập.

6. Có hai hộp. Hộp thứ nhất có 3 bi đỏ và 2 bi xanh. Hộp thứ hai có 2 bi đỏ và 3 bi xanh. Lấy ngẫu nhiên 1 bi từ hộp thứ nhất bỏ sang hộp thứ hai, sau đó từ hộp thứ hai lấy ra 1 bi. Gọi X và Y tương ứng là số bi đỏ được lấy ra từ hộp thứ nhất và thứ hai. Lập bảng phân phối xác suất đồng thời của (X, Y).

7. Thống kê về giá thành sản phẩm (Y) và sản lượng (X) của một ngành sản xuất thu được bảng phân phối xác suất như sau:

Y \ X	30	50	80	100
6	0,05	0,06	0,08	0,11
7	0,06	0,15	0,04	0,08
8	0,07	0,09	0,1	0,11

a. Tìm giá thành sản phẩm trung bình và mức độ phân tán của nó.

b. Tìm sản lượng trung bình khi giá thành bằng 8.

c. X và Y có độc lập không?

d. X và Y có tương quan với nhau không?

8. Biến ngẫu nhiên rời rạc (X, Y) có bảng phân phối xác suất đồng thời như sau:

Y \ X	1	3	4	8
3	0,15	0,06	0,25	0,04
6	0,3	0,1	0,03	0,07

Tìm hàm phân bố xác suất đồng thời F(x, y).

9. Tìm hàm mật độ xác suất đồng thời của biến ngẫu nhiên hai chiều liện tục (X, Y) nếu hàm phân bố xác suất đồng thời có dạng:

$$F(x, y) = (1 - e^{-ax})(1 - e^{-by}) \qquad x \geq 0,\ y \geq 0$$

10. Biến ngẫu nhiên hai chiều liên tục (X, Y) có hàm mật độ xác suất đồng thời như sau:

$$f(x,y) = \frac{A}{\pi^2(16 + x^2)(25 + y^2)}$$

a. Tìm giá trị của A.

b. Tìm hàm phân bố xác suất đồng thời F(x, y).

11. Hàm mật độ xác suất đồng thời của biến ngẫu nhiên liên tục (X, Y) có dạng:

$$f(x,y) = Ae^{-ax^2 + bxy - cy^2} \qquad a > 0,\ c > 0$$

a. Tìm các hàm mật độ xác suất biên $f_1(x)$, $f_2(y)$

b. Với điều kiện nào thì X và Y độc lập?

12. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	20	30	40	50
P_X	0,1	0,3	0,4	0,2

và biến ngẫu nhiên $Y = 0,2X - 2$.

a. Lập bảng phân phối xác suất của Y.

b. Tìm E(Y) và V(Y) theo hai cách:

+ Dùng bảng phân phối xác suất của Y

+ Dùng tính chất của kỳ vọng toán và phương sai.

13. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	0	1	2	3
P_X	0,14	0,39	0,36	0,11

và biến ngẫu nhiên $Y = -100X^2 + 300X + 500$.

Tìm E(Y) và V(Y) bằng hai cách:

+ Dùng bảng phân phối xác suất của Y

+ Dùng tính chất của kỳ vọng toán và phương sai.

14. Cho X là biến ngẫu nhiên và $Y = \varphi(X)$. Để tìm kỳ vọng toán E(Y) và phương sai V(Y) thì dùng cách nào dễ hơn.

a. Nếu $\varphi(X)$ là hàm tuyến tính

b. Nếu $\varphi(X)$ là hàm phi tuyến.

15. Cho $Z = X^2 + Y^2$ với X và Y có bảng phân phối xác suất đồng thời như sau:

Y \ X	0	2	4
0	0,1	0,1	0
2	0,1	0,4	0,1
4	0	0,1	0,1

a. Tìm E(Z) bằng cách trực tiếp dùng bảng phân phối xác suất đồng thời.

b. Tìm E(Z) thông qua bảng phân phối xác suất của Z.

c. Tìm các kỳ vọng toán sau đây bằng cách đơn giản nhất.

+ $E[(X - 2)(Y - 2)]$

+ $E[(X - 2)^2]$

+ $E(4X + 2Y)$

Chương V

CÁC ĐỊNH LÝ GIỚI HẠN

Như ta đã thấy ở các chương trước, không thể dự đoán trước được một cách chắc chắn xem biến ngẫu nhiên sẽ nhận giá trị nào trong các giá trị có thể có của nó khi thực hiện phép thử. Điều đó phụ thuộc vào rất nhiều nhân tố mà ta không thể tính hết được. Tuy nhiên vấn đề sẽ khác đi khi ta xét cùng một lúc một số lớn các biến ngẫu nhiên.Với một số điều kiện khá rộng rãi, hành vi tổng thể của một số lớn các biến ngẫu nhiên lại gần như mất đi tính ngẫu nhiên và trở nên có quy luật. Nói cách khác, khi ta tổng hợp một số lượng lớn các biến ngẫu nhiên thì tính ngẫu nhiên của hiện tượng mất đi và quy luật tất nhiên của nó được bộc lộ.

Đối với thực tiễn thì điều quan trọng là phải xác định các điều kiện trong đó tác động đồng thời của rất nhiều nguyên nhân ngẫu nhiên sẽ dẫn đến kết quả gần như không phụ thuộc gì vào các yếu tố ngẫu nhiên nữa vì lúc đó ta có thể dự đoán được tiến trình của hiện tượng. Các điều kiện này được chỉ ra trong các định lý có tên là các định lý giới hạn mà một vài kết luận của chúng đã được đề cập ở các chương trước. Ở đây ta sẽ chỉ xét một số định lý có nhiều ứng dụng hơn cả trong thực tế bao gồm một số định lý của luật số lớn và định lý giới hạn trung tâm.

Trước hết, ta xét một công cụ bổ trợ là bất đẳng thức Trêbưsép.

§1. BẤT ĐẲNG THỨC TRÊBUSÉP

Nếu X là biến ngẫu nhiên có kỳ vọng toán và phương sai hữu hạn thì với mọi số dương ε tùy ý ta đều có:

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2} \qquad (5.1)$$

Chứng minh. Ta sẽ chứng minh cho trường hợp X là biến ngẫu nhiên rời rạc. Việc chứng minh cho trường hợp X là biến ngẫu nhiên liên tục cũng tiến hành tương tự.

Giả sử X là biến ngẫu nhiên rời rạc với các giá trị có thể có là $x_1, x_2, ..., x_n$ với các xác suất tương ứng $p_1, p_2, ..., p_n$. Ta giả thiết thêm là k giá trị đầu tiên của X thỏa mãn điều kiện $|x_i - E(X)| < \varepsilon$, còn n – k giá trị còn lại thỏa mãn điều kiện $|x_i - E(X)| \geq \varepsilon$. Vì các biến cố để thực hiện các bất đẳng thức $|X - E(X)| < \varepsilon$ và $|X - E(X)| \geq \varepsilon$ đối lập với nhau, do đó:

$$P(|(X - E(X)| < \varepsilon) = 1 - P(|X - E(X)| \geq \varepsilon) \ (*)$$

Theo định nghĩa của phương sai của biến ngẫu nhiên rời rạc ta có:

$$V(X) = [x_1 - E(X)]^2 p_1 + ... + [x_k - E(X)]^2 p_k + \\ + [x_{k+1} - E(X)]^2 p_{k+1} + ... + [x_n - E(X)]^2 p_n$$

Rõ ràng là tất cả các số hạng của tổng trên đều không âm. Nếu ta bỏ bớt đi k số hạng đầu của tổng trên thì tổng chỉ có thể giảm đi. Do đó:

$$V(X) \geq [x_{k+1} - E(X)]^2 p_{k+1} + ... + [x_n - E(X)]^2 p_n$$

Chú ý rằng đối với các giá trị x_i; ($i = \overline{k+1,n}$) theo giả thiết ta đều có $|x_i - E(X)| \geq \varepsilon$; $i = k+1,n$ do đó $[x_i - E(X)]^2 \geq \varepsilon_i$; nếu trong mỗi số hạng ta thay một thừa số $[x_i - E(X)]^2$ bằng ε^2; thì bất đẳng thức đã cho tiếp tục mạnh thêm. Ta có:

$$V(X) \geq (p_{k+1} + ... + p_n)\varepsilon^2$$

Theo định lý cộng xác suất tổng $p_{k+1} + ... + p_n$ là xác suất để biến ngẫu nhiên X nhận một trong các giá trị x_{k+1}, ..., x_n song mọi giá trị nói trên đều thỏa mãn bất đẳng thức ($|x_i - E(X)| \geq \varepsilon$). Từ đó suy ra là tổng $p_{k+1} + ... + p_n$ chính là xác suất

$$P(|X - E(X)| \geq \varepsilon)$$

Từ đó ta có:

$$V(X) \geq P(|X - E(X)| \geq \varepsilon)\varepsilon^2$$

hay
$$P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2} \qquad (5.2)$$

Thay (5.2) vào (*) ta thu được:

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

Trong thực tế biểu thức (5.2) cũng được sử dụng như một dạng khác của bất đẳng thức Trêbưsép.

Về mặt thực tiễn bất đẳng thức Trêbưsép chỉ cho phép đánh giá cận trên hoặc cận dưới xác suất để biến ngẫu nhiên X nhận giá trị sai lệch so với kỳ vọng toán của nó lớn hơn hoặc nhỏ hơn ε. Đôi khi sự đánh giá đó là hiển nhiên và không có ý nghĩa. Chẳng hạn nếu $V(X) \geq \varepsilon^2$ thì bất đẳng thức Trêbưsép cho kết quả hiển nhiên song nó lại có ưu điểm là áp dụng được đối với mọi biến ngẫu nhiên mà không cần biết

quy luật phân phối xác suất của nó. Về mặt lý thuyết, bất đẳng thức Trêbưsép có ý nghĩa rất to lớn. Nó được sử dụng để chứng minh các định lý của luật số lớn.

Thí dụ. Thu nhập trung bình hàng năm của dân cư một vùng là 700 USD và độ lệch chuẩn là 120 USD. Hãy xác định một khoảng thu nhập hàng năm xung quanh giá trị trung bình của ít nhất 95% dân cư vùng đó.

Giải. Gọi X là thu nhập hàng năm của dân cư vùng đó thì X là biến ngẫu nhiên với quy luật phân phối xác suất chưa biết song có kỳ vọng toán E(X) = 700 và độ lệch chuẩn $\sigma_x = 120$. Do đó theo bất đẳng thức Trêbưsép:

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

$$\rightarrow P(|X - 700| < \varepsilon) \geq 1 - \frac{120^2}{\varepsilon^2} = 0,95$$

Từ đó $\varepsilon = 536,656$

Vậy ít nhất 95% dân cư vùng đó có thu nhập hàng năm nằm trong khoảng (700 – 536,656; 700 + 536,656) tức là khoảng (163,344; 1236,656).

§2. ĐỊNH LÝ TRÊBUSÉP

Nếu các biến ngẫu nhiên $X_1, X_2, ..., X_n...$ độc lập từng đôi, có các kỳ vọng toán hữu hạn và các phương sai đều bị chặn trên bởi hằng số C ($V(X_i) \leq C; i=\overline{1,n}$) thì với mọi ε dương bé tùy ý ta luôn có:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n}\right| < \varepsilon\right) = 1 \qquad (5.3)$$

Chứng minh. Xét biến ngẫu nhiên $\overline{X}$ là trung bình số học của các biến ngẫu nhiên nói trên:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Tìm kỳ vọng toán và phương sai của $\overline{X}$:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$

Áp dụng bất đẳng thức Trêbưsép đối với biến ngẫu nhiên $\overline{X}$:

$$P\left(\left|\overline{X} - E(\overline{X})\right| < \varepsilon\right) \geq 1 - \frac{V(\overline{X})}{\varepsilon^2} = 1 - \frac{\sum_{i=1}^{n} V(X_i)}{n^2\varepsilon^2}$$

Theo giả thiết $V(X_i) \leq C$; $i = \overline{1,n}$ do đó trong biểu thức trên nếu ta thay mỗi $V(X_i)$; $i = \overline{1,n}$ bằng C thì bất đẳng thức sẽ chỉ mạnh thêm:

$$P\left(\left|\overline{X} - E(\overline{X})\right| < \varepsilon\right) \geq 1 - \frac{nC}{n^2\varepsilon^2} = 1 - \frac{C}{n\varepsilon^2}$$

Lấy giới hạn của cả hai vế khi $n \to \infty$, ta có:

$$\lim_{n\to\infty} P\left(\left|\overline{X} - E(\overline{X})\right| < \varepsilon\right) \geq \lim_{n\to\infty}\left(1 - \frac{C}{n\varepsilon^2}\right) = 1$$

Sau nữa ta chú ý rằng xác suất của một biến cố không thể lớn hơn một, do đó $\lim_{n\to\infty} P\left(\left|\overline{X} - E(\overline{X})\right| < \varepsilon\right) = 1$

Ở trên ta giả thiết là các biến ngẫu nhiên $X_1, X_2, ..., X_n$ có các kỳ vọng toán khác nhau. Trong thực tế thường gặp trường hợp các biến ngẫu nhiên có cùng một kỳ vọng toán. Lúc đó có thể áp dụng trường hợp riêng của định lý Trêbusép như sau:

Nếu $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên độc lập từng đôi, có cùng kỳ vọng toán ($E(X_i) = m$; $i = \overline{1,n}$) và các phương sai cùng bị chặn trên ($V(X_i) \le C$; $i = \overline{1,n}$) thì với mọi ε dương bé tùy ý ta luôn có:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - m\right| < \varepsilon\right) = 1 \qquad (5.4)$$

Định lý trên còn được gọi là *luật số lớn của Trêbusép.*

Bản chất của định lý Trêbusép là nó chứng minh sự hội tụ theo xác suất của trung bình số học của một số lớn các biến ngẫu nhiên về trung bình số học của các kỳ vọng toán tương ứng. Nói cách khác nó chứng tỏ sự ổn định của trung bình số học của một số lớn các biến ngẫu nhiên xung quanh trung bình số học của các kỳ vọng toán của các biến ngẫu nhiên ấy.

Như vậy mặc dù từng biến ngẫu nhiên độc lập có thể nhận giá trị khác nhiều so với kỳ vọng toán của chúng, song trung bình số học của một số lớn các biến ngẫu nhiên lại nhận giá trị gần bằng trung bình số học của các kỳ vọng toán của chúng với xác suất rất lớn. Điều đó cho phép dự đoán giá trị của trung bình số học của các biến ngẫu nhiên.

Trong thực tế định lý Trêbusép có ứng dụng rộng rãi trong nhiều lĩnh vực, chẳng hạn trường hợp riêng của nó chính là cơ sở cho phương pháp đo lường trong vật lý. Để xác định giá trị của một đại lượng vật lý nào đó người ta thường tiến hành đo n lần và lấy trung bình số học của các kết quả đo làm giá trị thực của đại lượng cần đo. Thật vậy, giả sử xem kết quả của n lần đo là các biến ngẫu nhiên $X_1, X_2, ..., X_n$. Ta thấy rằng đối với các biến ngẫu nhiên này có thể áp dụng được trường hợp riêng của định lý Trêbusép vì chúng độc lập với nhau (từ đó chúng cũng độc lập từng đôi một với nhau), có cùng kỳ vọng toán vì nếu không có sai số hệ thống thì kỳ vọng toán của các biến ngẫu nhiên ấy chính bằng giá trị thực của đại lượng vật lý. Cuối cùng các phương sai của chúng đều bị chặn trên bằng chính độ chính xác của thiết bị đo. Do đó theo định lý Trêbusép ta có thể chứng tỏ rằng trung bình số học của các kết quả đo sẽ sai lệch rất ít so với giá trị thực của đại lượng vật lý và điều đó xảy ra với xác suất gần như bằng một.

Định lý Trêbusép còn là cơ sở cho một phương pháp được áp dụng rộng rãi trong thống kê là phương pháp mẫu mà thực chất của nó là dựa vào một mẫu ngẫu nhiên khá nhỏ có thể kết luận về toàn bộ tập hợp tổng quát của các đối tượng được nghiên cứu.

Chẳng hạn để đánh giá năng suất cây trồng của một vùng nào đó người ta không cần phải điều tra trên toàn bộ diện tích của vùng đó mà chỉ cần dựa vào kết quả thu hoạch của một mẫu ngẫu nhiên khá nhỏ mà vẫn đưa ra được các kết luận đủ chính xác về năng suất cây trồng của vùng đó.

Qua vài thí dụ như vậy có thể thấy được ý nghĩa to lớn của định lý Trêbusép đối với thực tiễn.

§3. ĐỊNH LÝ BERNOULLI

Nếu f là tần suất xuất hiện biến cố A trong n phép thử độc lập và p là xác suất xuất hiện biến cố đó trong mỗi phép thử thì với mọi ε dương bé tùy ý ta luôn có

$$\lim_{n\to\infty} P\left(|f - p| < \varepsilon\right) = 1 \qquad (5.5)$$

Chứng minh. Xét biến ngẫu nhiên $f = \frac{X}{n}$ là tần suất xuất hiện biến cố A trong n phép thử độc lập.

Ta tìm kỳ vọng toán và phương sai của nó:

$$E(f) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X)$$

$$V(f) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X)$$

Song X là số lần xuất hiện biến cố A trong n phép thử độc lập, mà theo giả thiết xác suất xuất hiện biến cố trong mỗi phép thử bằng p. Như vậy X phân phối theo qui luật nhị thức với các tham số là n và p. Do đó ta có E(X) = np và V(X) = np(1 – p), từ đó:

$$E(f) = \frac{1}{n}np = p$$

$$V(f) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Áp dụng bất đẳng thức Trêbusép đối với biến ngẫu nhiên f ta có:

$$P(|f-p|<\varepsilon)\geq 1-\frac{p(1-p)}{n\varepsilon^2}$$

Lấy giới hạn của hai vế khi $n\to\infty$:

$$\lim_{n\to\infty}P(|f-p|<\varepsilon)\geq\lim_{n\to\infty}\left(1-\frac{p(1-p)}{n\varepsilon^2}\right)=1$$

Mặt khác, vì xác suất không thể lớn hơn một, do đó:

$$\lim_{n\to\infty}P(|f-p|<\varepsilon)=1$$

Định lý trên còn được gọi là *luật số lớn của Bernoulli.*

Định lý Bernoulli chứng minh sự hội tụ theo xác suất của tần suất xuất hiện biến cố trong n phép thử độc lập về xác suất xuất hiện biến cố đó trong mỗi phép thử khi số phép thử tăng lên vô hạn. Nó chứng tỏ sự ổn định của tần suất xung quanh giá trị xác suất của biến cố đó.

Định lý Bernoulli là cơ sở lý thuyết của định nghĩa thống kê về xác suất, do đó nó cũng là cơ sở cho mọi áp dụng của định nghĩa thống kê về xác suất trong thực tế.

Chú ý rằng trong các định lý của luật số lớn ta chỉ đề cập đến sự hội tụ theo xác suất chứ không phải sự hội tụ theo nghĩa thông thường của giải tích toán học. Chẳng hạn theo định lý Bernoulli không thể kết luận rằng $\lim f=p$ tức là không thể kết luận khi n đủ lớn thì f sẽ luôn luôn sai lệch không đáng kể so với p. Sự hội tụ theo xác suất chỉ có nghĩa là khi n đủ lớn thì việc f và p sai lệch nhau không đáng kể sẽ có thể xem như có xác suất bằng 1. Như vậy thì với từng giá

trị riêng biệt của n, bất đẳng thức có thể vẫn không thỏa mãn, tức là f và p vẫn có thể sai lệch nhau đáng kể. Vì vậy định lý Bernoulli có thể viết ngắn gọn như sau:

$$f \xrightarrow[n\to\infty]{\text{Hội tụ theo xác suất}} p$$

§4. ĐỊNH LÝ GIỚI HẠN TRUNG TÂM

Định lý giới hạn tổng quát hơn cả là định lý giới hạn trung tâm của Liapunov (xem sử dụng của nó ở mục 7.8 chương III). Ở đây ta chỉ xét một trường hợp của nó được sử dụng nhiều trong thống kê. Trước hết ta xét khái niệm hàm đặc trưng.

4.1. Hàm đặc trưng

1. Định nghĩa. *Hàm đặc trưng của biến ngẫu nhiên X là kỳ vọng toán của biến ngẫu nhiên e^{itx} và được ký hiệu là $\varphi_X(t)$. Tức là:*

$$\varphi_X(t) = E[e^{itX}] = E(\cos tX) + iE(\sin tX)$$

Như vậy nếu X là biến ngẫu nhiên rời rạc thì:

$$\varphi_X(t) = \sum_j e^{itx_j} P_j$$

Còn nếu X là biến ngẫu nhiên liên tục thì:

$$\varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx$$

2. Các tính chất của hàm đặc trưng

a. $|\varphi_X(t)| \le 1$

Thật vậy vì $|e^{itx}| = 1$ nên:

$$\left|\int_{-\infty}^{+\infty} e^{itx} f(x)dx\right| \le \int_{-\infty}^{+\infty} \left|e^{itx}\right| f(x)dx = 1$$

b. Nếu $Y = ax + b$ thì $\varphi_Y(t) = e^{ibt}\varphi_X(at)$

Thật vậy, theo định nghĩa:

$$\varphi_Y(t) = E\left[e^{itY}\right] = E\left[e^{it(ax+b)}\right] = e^{itb}.E\left[e^{itax}\right] = e^{itb}.\varphi_X(at)$$

c. F(x) xác định một cách duy nhất hàm đặc trưng $\varphi_X(t)$

d. Nếu $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên độc lập thì

$$\varphi_{X_1+...+X_n}(t) = \prod_{k=1}^{n} \varphi_{X_k}(t)$$

Thật vậy, dễ dàng thấy rằng vì các biến ngẫu nhiên X_1, X_2, ..., X_n độc lập nên theo tính chất của kỳ vọng toán ta có:

$$\varphi_{X_1+X_2+...+X_n}(t) = E\left[e^{it(X_1+X_2+...+X_n)}\right]$$

$$= E\left[e^{itX_1}.e^{itX_2}. ...e^{itX_n}\right]$$

$$= \prod_{k=1}^{n} E\left[e^{itX_k}\right] = \prod_{k=1}^{n} \varphi_{X_k}(t)$$

e. Nếu tồn tại $E|X|^k$ thì hàm đặc trưng $\varphi_X(t)$ cũng tồn tại đạo hàm đến bậc k tại mọi điểm t (khả vi đến bậc k tại mọi điểm t).

Hệ quả: Nếu tồn tại $E|X|^k$ thì $\varphi_X(t)$ sẽ có khai triển Taylor như sau:

$$\varphi_X(t) = 1 + m_1 it + m_2 \frac{(it)^2}{2!} + \ldots + m_k \frac{(it)^k}{k!} + O(t^k)$$

trong đó $m_l = E[X^l]$ $(l = \overline{1, k})$.

f. Nếu $\{F_n(x)\}$ là dãy hàm phân bố xác suất và $\{\varphi_n(t)\}$ là dãy các hàm đặc trưng tương ứng thì điều kiện cần và đủ để $\{F_n(x)\}$ hội tụ yếu (tức là hội tụ tại các điểm $F_n(x)$ liên tục) tới hàm phân bố xác suất $F(x)$ là $\{\varphi_n(t)\}$ hội tụ tại mọi t đến hàm đặc trưng $\varphi(t)$ tương ứng với $F(x)$.

Thí dụ 1. Cho biến ngẫu nhiên $X \sim A(p)$. Tìm hàm đặc trưng $\varphi_X(t)$.

Giải. Theo định nghĩa của quy luật A(p)

$$P_x = P^x(1-p)^{1-x} \text{ với } x = \overline{0, 1}$$

nên

$$\varphi_X(t) = E[e^{itx}] = e^{it0}(1-p) + e^{it1}p = p.e^{it} + (1-p)$$

Thí dụ 2. Cho biến ngẫu nhiên $X \sim B(n, p)$. Tìm $\varphi_X(t)$.

Giải. Theo định nghĩa của quy luật B(n, p)

$$P_x = C_n^x p^x (1-p)^{n-x} \qquad x = \overline{0, n}$$

nên

$$\varphi_X(t) = E[e^{itx}] = \sum_{x=0}^{n} e^{itx} C_n^x p^x (1-p)^{n-x}$$

$$= \sum_{i=0}^{n} C_n^x (pe^{it})^x (1-p)^{n-x} = [pe^{it} + (1-p)]^n$$

Qua đó chứng minh mối liên hệ giữa quy luật nhị thức và quy luật không - một.

Thí dụ 3. Cho biến ngẫu nhiên $X \sim P(\lambda)$. Tìm $\varphi_X(t)$.

Giải. Theo định nghĩa của quy luật Poisson

$$P_x = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0,1...$$

nên

$$\varphi_X(t) = \sum_{x=0}^{\infty} e^{itx} e^{-\lambda}\frac{\lambda^x}{x!} = \sum_{x=0}^{\infty} e^{-\lambda}\frac{(\lambda e^{it})^x}{x!} = e^{-\lambda}.e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

Ngoài ra có thể chứng minh được rằng nếu biến ngẫu nhiên $X \sim N(\mu, \sigma^2)$ thì hàm đặc trưng bằng:

$$\varphi_X(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}$$

và đối với phân phối N(0, 1) thì:

$$\varphi_U(t) = e^{-\frac{t^2}{2}}$$

Với khái niệm hàm đặc trưng ta sẽ xét định lý giới hạn trung tâm sau đây:

4.2. Định lý Lindenberg - Lewi

Nếu $X_1, X_2, ..., X_n, ...$ là một dãy các biến ngẫu nhiên độc lập cùng tuân theo một quy luật phân phối xác suất nào đó với kỳ vọng toán và phương sai hữu hạn:

$$E(X_K) = a,\ V(X_K) = \sigma^2;\ \forall k$$

thì quy luật phân phối xác suất của biến ngẫu nhiên

$$U_n^c = \frac{U_n - E(U_n)}{\sqrt{V(U_n)}} \quad \text{với } U_n = \sum_{K=1}^{n} X_K$$

sẽ hội tụ khi $n \to \infty$ tới quy luật chuẩn hóa N(0, 1). Tức là:

$$P(U_n^c < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Chứng minh. Theo tính chất của hàm đặc trưng, ta chỉ cần chỉ ra rằng hàm đặc trưng $\varphi_{U_n^c}(t)$ hội tụ đến hàm đặc trưng của phân phối chuẩn hóa là $\varphi_U(t) = e^{-\frac{t^2}{2}}$ khi $n \to \infty$.

Xét biến ngẫu nhiên

$$U_n - E(U_n) = (X_1 - E(X_1)) + (X_2 - E(X_2)) + \ldots + (X_n - E(X_n))$$

nếu ký hiệu $Y_K = X_K - E(X_K)$ ta có:

$$U_n - E(U_n) = Y_1 + Y_2 + \ldots + Y_n$$

Mặt khác ta lại có

$$E(Y_K) = E[X_K - E(X_K)] = 0 \ \forall K$$

$$V(Y_K) = E(Y_K^2) = \sigma^2 = V(X_K) \ \forall K$$

Do đó nếu đặt $\varphi(t) = \varphi_{YK}(t)$ thì theo hệ quả trên ta có :

$$\varphi(t) = 1 + \frac{(it)^2}{2!}\sigma^2 + 0(t^2)$$

Hơn nữa, do $U_n - E(U_n) = \sum_{K=1}^{n} Y_K$ nên

$$\varphi_{U_n - E(U_n)}(t) = [\varphi(t)]^n$$

từ đó theo tính chất của hàm đặc trưng ta thu được hàm đặc trưng của U_n^c là:

$$\varphi_{U_n^c}(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = \left[1 - \frac{\sigma^2}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^2 + 0\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^2\right)\Bigg]^n =$$

$$= \left[1 - \frac{t^2}{2n} + 0\left(\frac{t^2}{\sigma^2 n}\right)\right]^n$$

do đó:

$$\lim_{n\to\infty} \varphi_{U_n^c}(t) = \lim_{n\to\infty}\left[1 - \frac{t^2}{2n} + 0\left(\frac{t^2}{\sigma^2 n}\right)\right]^n$$

$$= e^{\lim\limits_{n\to\infty} n\left(\frac{-t^2}{2n} + 0\left(\frac{t^2}{\sigma^2 n}\right)\right)} = e^{-\frac{t^2}{2}}$$

Về mặt thực hành người ta có thể ứng dụng định lý giới hạn trung tâm như sau: Với n đủ lớn ta có thể cho rằng:

$$P(U_n^c < x) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

hoặc

$$P(a < Un < b) \approx \Phi_0\left(\frac{b - E(U_n)}{\sqrt{V(U_n)}}\right) - \Phi_0\left(\frac{a - E(U_n)}{\sqrt{V(U_n)}}\right)$$

Thí dụ 4. **Chọn ngẫu nhiên 192 số trên đoạn [0, 1]. Tìm xác suất để tổng số điểm thu được (X) nằm trong khoảng (88, 104).**

Giải. **Ta có thể coi như** $X = \sum_{i=1}^{192} X_i$, **trong đó mọi biến ngẫu nhiên** X_i **độc lập và cùng tuân theo quy luật phân phối đều U(0, 1). Từ đó:**

$$E(X_i) = \frac{1}{2} \text{ và } V(X_i) = \frac{1}{12} \quad \forall i$$

Từ đó $E(X) = nE(X_i) = 96$ và

$$V(X) = nV(X_i) = 16$$

$$\rightarrow \sqrt{V(X)} = 4 = \sigma_X$$

Vì vậy:

$$P(88 < X < 104); \Phi_0\left(\frac{104-96}{4}\right) - \Phi_0\left(\frac{88-96}{4}\right)$$

$$= 2\Phi_0(2) = 0,9544$$

Thí dụ 5. Cho biến ngẫu nhiên $X \sim B$ ($n = 1000$, $p = 0,02$). Tìm xác suất để X nhận giá trị trong khoảng (40 ; 50).

Giải. Có thể coi $X = \sum_{i=1}^{1000} X_i$ trong đó các X_i độc lập và cùng phân phối $A(p = 0,02)$. Từ đó theo định lý giới hạn trung tâm suy ra:

$$X \sim N(\mu = np = 20; \sigma = np(1-p) = 19,6)$$

do đó:

$$P(40 < X < 50) \approx \Phi_0\left(\frac{50-20}{\sqrt{19,6}}\right) - \Phi_0\left(\frac{40-20}{\sqrt{19,6}}\right) =$$

$$= \Phi_0(6,77) \cdot \Phi_0(4,51) =$$

$$= 0,5 - 0,4999 = 0,0001$$

Các ký hiệu và công thức cơ bản

* Bất đẳng thức Trêbusép:

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

* Định lý Trêbusép:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n}\right| < \varepsilon\right) = 1$$

* Định lý Bernoulli

$$\lim_{n\to\infty} P(|f - P| < \varepsilon) = 1$$

* Hàm đặc trưng $\varphi_X(t) = E[e^{itX}]$

Nếu $U \sim N(0, 1) \Rightarrow \varphi_U(t) = e^{-\frac{t^2}{2}}$

* Định lý giới hạn trung tâm

Nếu $X_1, X_2, \ldots, X_n, \ldots$ độc lập và cùng tuân theo một quy luật phân phối xác suất với kỳ vọng toán và phương sai hữu hạn thì:

$$\lim_{n\to\infty} P(U_n^c < x) = \lim_{n\to\infty} F_{U_n^c}(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Trong đó: $U_n = \sum_{i=1}^{n} X_i$; $U_n^c = \frac{U_n - E(U_n)}{\sqrt{V(U_n)}}$

+ Với n đủ lớn ta có:

$$P(a < U_n < b) = \Phi_0\left(\frac{b - E(U_n)}{\sqrt{V(U_n)}}\right) - \Phi_0\left(\frac{a - E(U_n)}{\sqrt{V(U_n)}}\right)$$

Câu hỏi ôn tập

1. Khi nào thì có thể áp dụng được bất đẳng thức Trêbưsép đối với biến ngẫu nhiên X?

2. Khi nào thì việc áp dụng bất đẳng thức Trêbưsép sẽ không có ý nghĩa?

3. Dùng bất đẳng thức Trêbưsép hãy xây dựng quy tắc 3σ cho một biến ngẫu nhiên bất kỳ.

4. Cho biến ngẫu nhiên X với kỳ vọng toán bằng 20 và phương sai bằng 2. Dùng bất đẳng thức Trêbưsép hãy đánh giá xác suất để X nhận giá trị trong khoảng (17; 24).

5. Với những điều kiện nào thì có thể áp dụng Định lý Trêbưsép đối với dãy các biến ngẫu nhiên $X_1, X_2, ..., X_n, ...$?

6. Khái niệm hội tụ theo xác suất khác với khái niệm hội tụ trong giải tích ở điểm nào?

7. Dãy các biến ngẫu nhiên độc lập X_i (i = 1, 2, ...) có bảng phân phối xác suất như sau:

X_i	$\sqrt{2}$	0	$-\sqrt{2}$
P_{X_i}	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Hỏi có thể áp dụng được luật số lớn của Trêbusép đối với dãy các biến ngẫu nhiên nói trên hay không?

8. Gọi X_K (K = 1, 2, ..., n) là số mặt sấp ở lần tung thứ k của một đồng xu và $S_n = X_1 + X_2 + ... + X_K$. Hãy chứng tỏ rằng $\forall \varepsilon > 0$ bé tùy ý thì:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - 0{,}5\right| \geq \varepsilon\right) = 0$$

9. Tìm hàm đặc trưng của biến ngẫu nhiên liên tục X phân phối đều trên đoạn [a, b].

10. Tìm hàm đặc trưng của biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} e^{-x} & \text{với } x \geq 0 \\ 0 & \text{với } x < 0 \end{cases}$$

11. Cho dãy các biến ngẫu nhiên độc lập $X_1, X_2, ... X_n$ cùng phân phối theo một quy luật nào đó với kỳ vọng toán đều bằng a và phương sai đều bằng b^2. Hãy áp dụng định lý giới hạn trung tâm của Lindenberg - Lewi để tìm quy luật phân phối xác suất của biến ngẫu nhiên

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

khi n đủ lớn.

Phần thứ hai

THỐNG KÊ TOÁN

Thống kê toán là bộ môn toán học nghiên cứu qui luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý các số liệu thống kê - các kết quả quan sát. Như vậy, nội dung chủ yếu của thống kê toán là xây dựng các phương pháp thu thập và xử lý các số liệu thống kê nhằm rút ra các kết luận khoa học và thực tiễn.

Các phương pháp thống kê toán là công cụ để giải quyết nhiều vấn đề khoa học và thực tiễn nảy sinh trong các lĩnh vực khác nhau của tự nhiên và kinh tế - xã hội.

Chương VI

CƠ SỞ LÍ THUYẾT MẪU

§1. KHÁI NIỆM VỀ PHƯƠNG PHÁP MẪU

Trong thực tế thường phải nghiên cứu một tập hợp các phần tử đồng nhất theo một hay nhiều dấu hiệu định tính hoặc định lượng đặc trưng cho các phần tử đó. Chẳng hạn một doanh nghiệp phải nghiên cứu tập hợp các khách hàng của nó thì dấu hiệu định tính có thể là mức độ hài lòng của khách hàng đối với sản phẩm hoặc dịch vụ của doanh nghiệp, còn dấu hiệu định lượng là nhu cầu của khách hàng về số lượng sản phẩm của doanh nghiệp.

Để nghiên cứu tập hợp các phần tử này theo một dấu hiệu nhất định đôi khi người ta sử dụng phương pháp nghiên cứu toàn bộ, tức là thống kê toàn bộ tập hợp đó và phân tích từng phần tử của nó theo dấu hiệu nghiên cứu. Chẳng hạn để nghiên cứu dân số của một nước theo các dấu hiệu như tuổi tác, trình độ văn hóa, địa bàn cư trú, cơ cấu nghề nghiệp... có thể tiến hành tổng điều tra dân số và phân tích từng người theo các dấu hiệu trên, từ đó tổng hợp thành dấu hiệu chung cho toàn bộ dân số của nước đó. Tuy nhiên trong thực tế việc áp dụng phương pháp này gặp phải những khó khăn chủ yếu sau:

- Nếu quy mô của tập hợp quá lớn thì việc nghiên cứu toàn bộ sẽ đòi hỏi nhiều chi phí vật chất và thời gian.

- Nhiều khi cũng do quy mô của tập hợp quá lớn nên có thể xảy ra trường hợp tính trùng hoặc bỏ sót các phần tử của nó.

- Do quy mô nghiên cứu lớn mà trình độ tổ chức nghiên cứu lại hạn chế dẫn đến các sai sót trong quá trình thu thập thông tin ban đầu, hạn chế độ chính xác của kết quả phân tích.

- Trong nhiều trường hợp không thể nắm được toàn bộ các phần tử của tập hợp cần nghiên cứu, do đó không thể tiến hành nghiên cứu toàn bộ được.

- Nếu các phần tử của tập hợp lại bị phá hủy trong quá trình nghiên cứu thì phương pháp nghiên cứu toàn bộ trở thành vô nghĩa.

Vì thế trong thực tế phương pháp nghiên cứu toàn bộ thường chỉ được áp dụng đối với các tập hợp có quy mô nhỏ, còn chủ yếu người ta áp dụng phương pháp nghiên cứu không toàn bộ, đặc biệt là phương pháp nghiên cứu chọn mẫu. Phương pháp này chủ trương từ tập hợp cần nghiên cứu chọn ra một số phần tử (gọi là mẫu), phân tích các phần tử này và dựa vào đó mà suy ra các kết luận về tập hợp cần nghiên cứu. Nếu mẫu được chọn ra một cách ngẫu nhiên và xử lý bằng các phương pháp xác suất thì vừa thu được các kết luận một cách nhanh chóng, đỡ tốn kém mà vẫn đảm bảo độ chính xác cần thiết.

Việc thu thập, sắp xếp và trình bày các số liệu của tổng thể hoặc một mẫu gọi là thống kê mô tả. Còn việc sử dụng thông tin của mẫu để tiến hành các suy đoán, kết luận về tổng thể gọi là thống kê suy diễn.

§2. TỔNG THỂ NGHIÊN CỨU

2.1. Định nghĩa. *Toàn bộ tập hợp các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hoặc định lượng nào đó được gọi là tổng thể nghiên cứu hay tổng thể.*

Số lượng các phần tử của tổng thể được gọi là *kích thước* của tổng thể, ký hiệu là N. Thường thì kích thước N của tổng thể là hữu hạn, song nếu tổng thể quá lớn hoặc không thể nắm được toàn bộ các phần tử của tổng thể ta có thể giả thiết rằng kích thước của tổng thể là vô hạn. Điều giả thiết này dựa trên cơ sở là khi tăng kích thước của tổng thể lên khá lớn thì thực tế không ảnh hưởng gì đến kết quả tính toán trên số liệu của từng bộ phận rút ra từ tổng thể đó.

Với mỗi tổng thể ta không nghiên cứu trực tiếp tổng thể đó mà thông qua một hay nhiều dấu hiệu đặc trưng cho tổng thể đó. Chúng được gọi là dấu hiệu nghiên cứu, ký hiệu là χ. Các dấu hiệu này có thể là định tính hoặc định lượng.

2.2. Các phương pháp mô tả tổng thể

1. Giả sử trong tổng thể dấu hiệu nghiên cứu định lượng χ nhận các giá trị $x_1, x_2, ..., x_n$ với các tần số tương ứng $N_1, N_2, ..., N_k$ (N_i là số phần tử của tổng thể có chung giá trị x_i). Lúc đó tổng thể có thể mô tả bằng bảng phân phối tần số sau:

Giá trị của χ	x_1	x_2	...	x_i	...	x_k
Tần số	N_1	N_2	...	N_i	...	N_k

Hiển nhiên

$$\begin{cases} 0 \le N_i \le N \quad \forall i \\ \sum_{i=1}^{k} N_i = N \end{cases} \tag{6.1}$$

2. Nếu ký hiệu p_i $(i = \overline{1,k})$ là tần suất của x_i, tức là tỷ số giữa tần số của x_i và kích thước của tổng thể thì :

$$P_i = \frac{N_i}{N} \qquad i = \overline{1,n} \tag{6.2}$$

và lúc đó tổng thể còn có thể mô tả bằng bảng phân phối tần suất sau:

Giá trị của χ	x_1	x_2	...	x_i	...	x_k
Tần suất	p_1	p_2	...	p_i	...	p_k

Từ (6.1) và (6.2) suy ra

$$\begin{cases} 0 \le P_i \le 1 \quad \forall i \\ \sum_{i=1}^{k} P_i = 1 \end{cases}$$

Về mặt hình thức, bảng phân phối tần suất của tổng thể tương tự như bảng phân phối xác suất của biến ngẫu nhiên rời rạc. Nó phản ánh cơ cấu của tổng thể theo dấu hiệu χ.

3. Nếu ký hiệu w_i $(i = \overline{1,k})$ là tần số tích lũy của x_i, tức là tổng số các phần tử có giá trị nhỏ hơn x_i, thì

$$w_i = \sum_{x_j < x_i} N_j \tag{6.3}$$

và $F(x_i)$ $(i = \overline{1,k})$ là tần suất tích lũy của x_i, tức là tỷ số giữa tần số tích lũy của nó và kích thước của tổng thể, thì

$$F(x_i) = \frac{w_i}{N} = \sum_{x_j < x_i} \frac{N_j}{N} \tag{6.4}$$

Tần suất tích lũy là một hàm của x_i, có tính chất giống như hàm phân bố xác suất của biến ngẫu nhiên rời rạc.

Việc mô tả tổng thể theo dấu hiệu nghiên cứu χ bằng bảng phân phối tần số, bảng phân phối tần suất và tần suất tích lũy cho thấy dấu hiệu định lượng χ hoàn toàn có thể mô hình hóa bằng một biến ngẫu nhiên rời rạc X. Điều này cũng đúng đối với các tổng thể mang dấu hiệu χ phân phối liên tục. Biến ngẫu nhiên X dùng để mô hình hóa dấu hiệu nghiên cứu χ được gọi là *biến ngẫu nhiên gốc*, còn quy luật phân phối xác suất của nó gọi là *quy luật phân phối gốc*. Việc mô hình hóa dấu hiệu nghiên cứu χ bằng biến ngẫu nhiên X cho phép áp dụng các công cụ xác suất đã xét ở phần trước trong việc nghiên cứu tổng thể.

2.3. Các tham số đặc trưng của tổng thể

Việc mô tả tổng thể theo dấu hiệu χ bằng bảng phân phối tần số, bảng phân phối tần suất và tần suất tích lũy mới chỉ cung cấp những thông tin chung nhất về tổng thể đó. Trong thực tế nhiều khi người nghiên cứu cần quan tâm đến những thông tin tổng hợp phản ánh những khía cạnh quan trọng nhất của tổng thể theo dấu hiệu nghiên cứu đó. Những thông tin này được biểu hiện qua các tham số đặc trưng chủ yếu sau đây của tổng thể.

1. Trung bình tổng thể

Giả sử trong tổng thể kích thước N dấu hiệu định lượng χ nhận các giá trị $x_1, x_2, ..., x_n$. Trung bình tổng thể, ký hiệu là m là trung bình số học của các giá trị của dấu hiệu trong tổng thể.

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (6.5)$$

Nếu trong tổng thể dấu hiệu χ chỉ nhận các giá trị x_1, x_2,..., x_k với các tần số tương ứng N_1, N_2, ..., N_k thì trung bình tổng thể được xác định bằng biểu thức:

$$m = \frac{1}{N}\sum_{i=1}^{k} x_i N_i \qquad (6.6)$$

Bản chất của trung bình tổng thể có thể làm rõ như sau: Giả sử tổng thể kích thước N bao gồm các phần tử mang các giá trị khác nhau của dấu hiệu nghiên cứu χ là x_1, x_2, ..., x_N. Giả sử từ tập hợp này lấy ngẫu nhiên ra một phần tử thì xác suất để lấy được phần tử mang giá trị x_i hiển nhiên là $\frac{1}{N}(i = \overline{1,N})$. Như vậy giá trị của dấu hiệu χ có thể xem như một biến ngẫu nhiên X với các giá trị có thể có là x_1, x_2, ..., x_n và các xác suất tương ứng đều bằng $\frac{1}{N}$. Từ đó

$$E(X) = x_1\frac{1}{N} + x_2\frac{1}{N} + \ldots + x_N\frac{1}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i = m$$

Như vậy m = E(X).

Mở rộng kết quả thu được cho tổng thể với dấu hiệu nghiên cứu liên tục ta thu được kết quả là nếu xem dấu hiệu nghiên cứu như biến ngẫu nhiên X thì trung bình tổng thể chính là kỳ vọng toán của biến ngẫu nhiên đó.

Thí dụ 1. Tổng thể nghiên cứu là một xí nghiệp có N = 40 công nhân với dấu hiệu nghiên cứu là năng suất lao động (sản phẩm/đơn vị thời gian). Số liệu của tổng thể theo dấu hiệu nghiên cứu được cho trong bảng sau (bảng 6.1).

Bảng 6.1

Năng suất lao động x_i	Số công nhân N_i	$N_i x_i$
50	3	150
55	5	275
60	10	600
65	12	780
70	7	490
75	3	225
	N = 40	$\sum N_i x_i = 2520$

Tìm năng suất lao động trung bình của mỗi công nhân.

Giải. Theo công thức (6.6) ta có:

$$m = \frac{1}{N}\sum_{i=1}^{k} N_i x_i = \frac{2520}{40} = 63$$

Ngoài trung bình tổng thể m (mà thực chất là trung bình số học) là loại trung bình được sử dụng nhiều nhất, trong thực tế, tùy thuộc vào từng trường hợp người ta còn tính các loại trung bình sau.

a. Trung bình điều hòa: Tương tự như trung bình số học, trong đó thay cho các giá trị của dấu hiệu nghiên cứu người ta dùng giá trị nghịch đảo của chúng. Như vậy nếu ký hiệu trung bình điều hòa là m_h thì ta có công thức:

$$m_h = \frac{N}{\sum_{i=1}^{n} \frac{1}{x_i}} \qquad (6.7)$$

Nếu dấu hiệu của tổng thể nhận các giá trị $x_1, x_2 \dots x_k$ với tần số tương ứng $N_1\ N_2 \dots N_k$ thì

$$m_h = \frac{N}{\sum_{i=1}^{k} \frac{N_i}{x_i}} \tag{6.8}$$

Thí dụ 2. Một xí nghiệp có hai phân xưởng cùng lắp ráp 1 loại sản phẩm. Phân xưởng thứ nhất lắp ráp một sản phẩm hết 15 phút, phân xưởng thứ hai hết 20 phút. Nếu trong một ngày mỗi phân xưởng làm việc 8 giờ thì hãy tìm thời gian trung bình để lắp ráp 1 sản phẩm.

Giải. Trong trường hợp này muốn tính thời gian trung bình để lắp ráp 1 sản phẩm không thể dùng công thức trung bình số học vì thời gian trung bình để lắp ráp 1 sản phẩm phải bằng tổng số thời gian sản xuất chia cho tổng số sản phẩm được lắp ráp trong thời gian đó. Vì vậy để tính thời gian trung bình phải tìm được tổng số sản phẩm mà hai phân xưởng lắp ráp được trong một ngày làm việc. Theo công thức (6.8) ta có:

$$m_h = \frac{N}{\sum_{i=1}^{k} \frac{N_i}{x_i}} = \frac{60\times 8 + 60\times 8}{\frac{60\times 8}{15} + \frac{60\times 8}{20}} = \frac{960}{56} = 17{,}14 \text{ phút}$$

Chú ý rằng cũng giống như trung bình số học, trung bình điều hòa chỉ được áp dụng khi các giá trị của dấu hiệu nghiên cứu có quan hệ tổng.

b. Trung bình nhân: Là căn bậc N của tích các giá trị của dấu hiệu trong tổng thể. Vậy nếu ký hiệu trung bình nhân là m_g thì

$$m_g = \sqrt[N]{x_1.x_2...x_N} \qquad (6.9)$$

hoặc

$$m_g = \sqrt[N]{x_{x_1}^{N_1}.x_2^{N_2}...x_k^{N_k}} \qquad (6.10)$$

nếu dấu hiệu chỉ nhận các giá trị x_1, ..., x_k với các tần số tương ứng N_1, ..., N_k.

Thí dụ 3. Trong khoảng thời gian 10 năm, tốc độ tăng giá trị sản lượng của một xí nghiệp như sau: Có 5 năm tốc độ tăng so với năm trước là 110%; có 2 năm tốc độ tăng là 125% và có 3 năm tốc độ tăng là 115%. Tìm tốc độ tăng trưởng trung bình hàng năm của xí nghiệp trong 10 năm đó.

Giải. Ở đây tốc độ tăng trưởng giá trị sản lượng không có quan hệ tổng (vì gốc so sánh khác nhau) mà có quan hệ tích. Do đó để tính tốc độ tăng trưởng bình quân hàng năm so với năm trước đó ta phải tính số trung bình nhân. Theo công thức (6.10) ta có:

$$m_g = \sqrt[N]{x_1^{N_1}.x_2^{N_2}...x_k^{N_k}} = \sqrt[10]{(1,1)^5.(1,25)^2.(1,15)^3}$$

$$\rightarrow \lg m_g = \frac{1}{10}\,(5.\lg 1,1 + 2.\lg 1,25 + 3.\lg 1,15) = 0,0583$$

$$\rightarrow m_g = 1,144 \text{ hay } 114,4\%.$$

Trong kinh tế và xã hội, trung bình nhân thường chỉ dùng để tính tốc độ tăng trưởng bình quân.

2. Phương sai tổng thể

Phương sai tổng thể, ký hiệu là σ^2, là trung bình số học của bình phương các sai lệch giữa các giá trị của dấu hiệu trong tổng thể và trung bình tổng thể.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - m)^2 \qquad (6.11)$$

Nếu các giá trị $x_1, x_2, ..., x_k$ của dấu hiệu có các tần số tương ứng là $N_1, N_2, ..., N_k$ với $N_1 + N_2 + ... + N_k = N$ thì

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}N_i(x_i - m)^2 \qquad (6.12)$$

Do có thể viết

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}N_i(x_i - m)^2 = \sum_{i=1}^{k}\frac{N_i}{N}(x_i - m)^2$$

$$= \sum_{i=1}^{k}(x_i - m)^2 P_i$$

nên về thực chất phương sai tổng thể chính là phương sai của biến ngẫu nhiên trong tổng thể đó. Nó phản ánh mức độ phân tán các giá trị của dấu hiệu χ xung quanh giá trị trung bình tổng thể.

Trong thực tế, để tiện cho việc tính toán, phương sai tổng thể thường được tính bằng công thức:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}N_i x_i^2 - m^2 \qquad (6.13)$$

Thật vậy theo định nghĩa

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}N_i(x_i - m)^2 = \frac{1}{N}\sum_{i=1}^{k}N_i(x_i^2 - 2mx_i + m^2) =$$

$$= \frac{1}{N}\sum_{i=1}^{k}N_i x_i^2 - 2m.\frac{1}{N}\sum_{i=1}^{k}N_i x_i + \frac{1}{N}\sum_{i=1}^{k}N_i m^2$$

song do $\frac{1}{N}\sum_{i=1}^{k}N_i x_i = m$

và $$\frac{1}{N}\sum_{i=1}^{k} N_i m^2 = \frac{Nm^2}{N} = m^2$$

nên $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} N_i x_i^2 - 2m^2 + m^2 = \frac{1}{N}\sum_{i=1}^{k} N_i x_i^2 - m^2$$

Nếu lấy căn bậc hai của phương sai ta sẽ thu được *độ lệch chuẩn tổng thể*:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$$

Thí dụ 4. Với các số liệu cho trong bảng (6.1) hãy tìm phương sai và độ lệch chuẩn của năng suất lao động của công nhân xí nghiệp đó.

Giải. Theo các số liệu của bảng (6.1) ta tìm ngay được

$$\sum_{i=1}^{k} N_i x_i^2 = 160500$$

Theo công thức (6.13) .

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} N_i\, x_i^2 - m^2 = \frac{160500}{40} - 63^2 = 43{,}5$$

$$\sigma = 6{,}6$$

3. Một loại tổng thể thường được nghiên cứu trong thực tế là tổng thể kích thước N, trong đó M phần tử mang dấu hiệu nghiên cứu, còn N – M phần tử còn lại không mang dấu hiệu đó. Lúc đó *tần suất của tổng thể* là tỷ số giữa số phần tử mang dấu hiệu nghiên cứu và kích thước của tổng thể:

$$p = \frac{M}{N} \qquad (6.14)$$

Ta thấy rằng p chính là xác suất để lấy ngẫu nhiên một phần tử thì phần tử đó mang dấu hiệu nghiên cứu. Như vậy, nếu xem dấu hiệu nghiên cứu là biến ngẫu nhiên X phân phối theo quy luật không - một thì p chính là kỳ vọng toán, tức là thực chất tần suất p là trường hợp riêng của trung bình tổng thể m và phản ánh cơ cấu của tổng thể theo dấu hiệu nghiên cứu χ.

Vì có thể đặc trưng dấu hiệu nghiên cứu trong tổng thể bằng một biến ngẫu nhiên X, vì thế các tham số đặc trưng khác của X như Mốt, Trung vị, Hệ số biến thiên, Hệ số bất đối xứng, Hệ số nhọn v.v... cũng đều là các tham số đặc trưng của tổng thể. Biểu thức tính các tham số này giống như đã trình bày ở mục 3 chương III.

§3. MẪU NGẪU NHIÊN

3.1. Định nghĩa mẫu ngẫu nhiên

Các tham số đặc trưng của tổng thể có thể xác định được một cách trực tiếp nếu áp dụng phương pháp nghiên cứu toàn bộ tổng thể. Song do những hạn chế như đã xét ở mục §1, chẳng hạn quy mô quá lớn của tổng thể hay mức độ kém tin cậy của số liệu điều tra nên việc tính toán vừa khó khăn, tốn kém mà vẫn không thu được kết quả chính xác. Đặc biệt, khi không thể nắm được kích thước của tổng thể (và phải coi N là vô hạn) thì thực tế là không thể nghiên cứu trực tiếp tổng thể được. Vì vậy, thường người ta áp dụng phương pháp

mẫu bằng cách chọn ra từ tổng thể n phần tử và chỉ tập trung nghiên cứu các phần tử đó mà thôi. Tập hợp n phần tử này được gọi là *mẫu kích thước n.*

Để có thể căn cứ vào thông tin của mẫu đưa ra những kết luận đủ chính xác về dấu hiệu nghiên cứu trong tổng thể thì trước hết mẫu được chọn phải mang tính đại diện cho tổng thể, tức là phản ánh đúng đặc điểm của tổng thể theo dấu hiệu nghiên cứu đó. Để đảm bảo tính đại diện của mẫu và tiện cho việc mô hình hóa, mẫu được tạo lập với những giả thiết sau:

- Lấy lần lượt từng phần tử vào mẫu. Phương pháp này gọi là phương pháp đơn giản để phân biệt với cách lấy cùng một lúc nhiều phần tử vào mẫu.

- Mỗi phần tử được lấy vào mẫu một cách hoàn toàn ngẫu nhiên, tức là mọi phần tử của tổng thể đều có thể được lấy vào mẫu với khả năng như nhau.

- Các phần tử được lấy vào mẫu theo phương thức hoàn lại, tức là trước khi lấy phần tử thứ k thì trả lại tổng thể phần tử thứ (k – 1) mà ta đã nghiên cứu xong ($k = \overline{2,n}$).

Trong thực tế nếu kích thước của tổng thể khá lớn còn mẫu chỉ chiếm một phần rất nhỏ của tổng thể thì phương thức lấy mẫu hoàn lại và không hoàn lại cho ta các kết quả sai lệch không đáng kể. Đặc biệt khi kích thước của tổng thể là vô hạn còn kích thước của mẫu lại hữu hạn thì không còn sự khác biệt giữa hai phương thức lấy mẫu nói trên nữa. Lúc đó có thể chọn mẫu theo phương thức không hoàn lại và vẫn có thể giả thiết mẫu được chọn theo phương thức hoàn lại.

Giả sử theo một phương pháp nào đó từ tổng thể lấy ra n phần tử tạo nên mẫu kích thước n. Vì mẫu được lấy ra theo

nguyên tắc đơn giản, ngẫu nhiên và hoàn lại nên có thể mô hình hóa mẫu được chọn như sau:

Gọi X_i $(i = \overline{1,n})$ là giá trị của dấu hiệu χ đo lường được trên phần tử thứ i của mẫu. Vì có thể mô hình hóa dấu hiệu χ bằng một biến ngẫu nhiên X với một quy luật phân phối xác suất nào đó nên việc chọn mẫu kích thước n theo nguyên tắc trên có thể xem như tiến hành n phép thử độc lập đối với X và lúc đó các giá trị X_i của dấu hiệu thu được trên mẫu có thể xem như các biến ngẫu nhiên thu được qua việc tiến hành n phép thử độc lập đối với biến ngẫu nhiên X. Từ đó ta có định nghĩa sau:

Định nghĩa. Mẫu ngẫu nhiên kích thước n là tập hợp của n biến ngẫu nhiên độc lập $X_1, X_2, ..., X_n$ được thành lập từ biến ngẫu nhiên X trong tổng thể nghiên cứu và có cùng quy luật phân phối xác suất với X.

Mẫu ngẫu nhiên thường được ký hiệu là:

$$W = (X_1, X_2, ..., X_n)$$

Chú ý rằng với cách xây dựng mẫu ngẫu nhiên như vậy thì các biến ngẫu nhiên $X_1, X_2, ..., X_n$ của mẫu không những có cùng dạng phân phối xác suất với biến ngẫu nhiên gốc X, tức là có cùng hàm phân bố xác suất F(x) mà các tham số đặc trưng của chúng cũng bằng các tham số đặc trưng của X, tức là:

$$E(X_1) = E(X_2) = ... = E(X_n) = E(X) = m \quad (6.15)$$

$$V(X_1) = V(X_2) = ... = V(X_n) = V(X) = \sigma^2 \quad (6.16)$$

Chẳng hạn gọi X là số chấm thu được khi tung một con xúc xắc, X là biến ngẫu nhiên với bảng phân phối xác suất như sau:

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Nếu tung con xúc xắc 3 lần và gọi X_i (i = 1,3) là số chấm xuất hiện ở lần tung thứ i thì ta có 3 biến ngẫu nhiên độc lập tạo nên mẫu ngẫu nhiên kích thước n = 3.

$$W = (X_1, X_2, X_3)$$

Hơn nữa, mỗi biến ngẫu nhiên X_i trong mẫu đều có bảng phân phối xác suất giống như bảng phân phối xác suất của biến ngẫu nhiên X, do đó:

$$E(X_i) = \frac{1}{6}(1 + 2 + ... + 6) = \frac{21}{6} = E(X) \quad \forall i$$

$$V(X_i) = \frac{1}{6}(1^2 + 2^2 + ... + 6^2) - \left(\frac{21}{6}\right)^2 = \frac{35}{12} = V(X) \quad \forall i$$

Lúc đó việc thực hiện một phép thử đối với mẫu ngẫu nhiên W chính là thực hiện một phép thử đối với mỗi thành phần của mẫu. Giả sử X_1 nhận giá trị x_1; X_2 nhận giá trị x_2;..., X_n nhận giá trị x_n. Tập hợp n giá trị $x_1, x_2, ..., x_n$ tạo thành một giá trị của mẫu ngẫu nhiên, hay còn gọi là một mẫu cụ thể, ký hiệu:

$$w = (x_1, x_2, ..., x_n)$$

Chẳng hạn, trong thí dụ trên nếu tiến hành một phép thử đối với mẫu ngẫu nhiên bằng cách tung cụ thể ba lần, lần đầu được 6 chấm, lần thứ hai được 5 chấm, lần thứ ba được 1 chấm thì ta thu được một mẫu cụ thể w = (6, 5, 1). Nếu thực hiện một phép thử khác đối với W lại thu được một mẫu cụ thể khác, chẳng hạn w = (2, 1, 4)...

Như vậy, mẫu ngẫu nhiên là tập hợp của n biến ngẫu nhiên, còn mẫu cụ thể lại là tập hợp của n giá trị cụ thể quan sát được khi thực hiện một phép thử đối với mẫu ngẫu nhiên.

3.2. Các phương pháp chọn mẫu

Như đã trình bày ở trên, với những cách thức tiến hành phép thử khác nhau, ta sẽ thu được những mẫu cụ thể khác nhau từ cùng một mẫu ngẫu nhiên song phải đảm bảo yêu cầu là mẫu phải đại diện cho tổng thể nghiên cứu.

Tùy thuộc vào đặc điểm của từng tổng thể nghiên cứu mà mẫu có thể được chọn theo nhiều phương pháp khác nhau để đảm bảo yêu cầu về tính đại diện của mẫu. Sau đây là một số phương pháp chọn mẫu chủ yếu thường được sử dụng để nghiên cứu các tổng thể kinh tế - xã hội.

1. Mẫu giản đơn: Là loại mẫu được chọn trực tiếp từ danh sách đã được đánh số của tổng thể. Từ tổng thể kích thước N người ta dùng cách rút thăm đơn giản ra n phần tử của mẫu theo một bảng số ngẫu nhiên nào đó.

Các bảng số ngẫu nhiên có thể sử dụng là:

- Các bảng của Tippett gồm các số có 4 chữ số;
- Các bảng của Fisher và Yates;
- Các bảng của Kendall và Babington Smith gồm các số có 5 chữ số;
- Các bảng của Burke Haton;
- Các bảng của công ty Rand... (xem phụ lục 10)

Phương pháp này có ưu điểm là cho phép thu được một mẫu có tính đại diện cao, cho phép suy rộng các kết quả của

mẫu cho tổng thể với một sai số xác định, song để vận dụng phải có được toàn bộ danh sách của tổng thể nghiên cứu. Mặt khác chi phí chọn mẫu sẽ khá lớn.

2. Mẫu hệ thống: Là loại mẫu đã được đơn giản hóa trong cách chọn, trong đó chỉ có phần tử đầu tiên được chọn một cách ngẫu nhiên, sau đó dựa trên danh sách đã được đánh số của tổng thể để chọn ra các phần tử tiếp theo vào mẫu theo một thủ tục nào đó. Chẳng hạn trên một danh sách N khách hàng cần chọn ra một mẫu kích thước n thì ta chia danh sách đó ra n phần bằng nhau, ở phần thứ nhất gồm $\frac{N}{n}$ phần tử chọn ngẫu nhiên ra một phần tử, sau đó theo danh sách cứ cách $\frac{N}{n}$ phần tử ta lấy ra một phần tử vào mẫu cho đến khi đủ n phần tử.

Nhược điểm chính của phương pháp này là dễ mắc sai số hệ thống khi danh sách của tổng thể không được sắp xếp một cách ngẫu nhiên mà lại theo một trật tự chủ quan nào đó. Tuy vậy, do cách thức đơn giản của nó, mẫu ngẫu nhiên hệ thống hay được dùng ở cấp chọn mẫu cuối cùng và khi tổng thể tương đối thuần nhất.

3. Mẫu chùm: Trong một số trường hợp, để tiện cho việc nghiên cứu người ta muốn quy diện nghiên cứu gọn về một khu vực nhất định chứ không để cho các phần tử của mẫu phân tán quá rộng, chẳng hạn tập trung nghiên cứu khách hàng tại một địa phương nào đó. Lúc đó mẫu được chọn theo chùm. Chẳng hạn chùm có thể là một hộ gia đình có nhiều người, một làng có nhiều hộ gia đình... Theo phương pháp này, trước tiên tổng thể điều tra được phân chia ra thành nhiều chùm theo nguyên tắc:

- Mỗi phần tử của tổng thể chỉ được phân vào một chùm.

- Mỗi chùm cố gắng chứa nhiều phần tử khác nhau về dấu hiệu nghiên cứu, sao cho nó có độ phân tán cao như của tổng thể.

- Phân chia sao cho các chùm tương đối đồng đều nhau về quy mô.

Các chùm được chọn một cách ngẫu nhiên và tất cả các phần tử của chùm đó đều được chọn vào mẫu.

Trong phương pháp chọn mẫu chùm do sự đồng đều cao hơn giữa các chùm nên dễ dẫn đến khả năng lặp lại thông tin. Như vậy sai số chọn mẫu có thể cao hơn phương pháp chọn ngẫu nhiên đơn với cùng kích thước mẫu, song nó vẫn được sử dụng cho đỡ tốn kém chi phí và thích hợp với việc nghiên cứu theo nhiều dấu hiệu cùng một lúc.

4. Mẫu phân tổ: Trong chọn mẫu phân tổ, trước hết người ta phân chia tổng thể ra thành các tổ có độ thuần nhất cao để chọn ra các phần tử đại diện cho từng tổ. Việc phân tổ có hiệu quả khi tổng thể nghiên cứu không thuần nhất theo dấu hiệu nghiên cứu. Sau khi đã phân tổ thì kích thước mẫu được phân bổ cho mỗi tổ theo một quy tắc nào đó, chẳng hạn tỷ lệ thuận với kích thước mỗi tổ.

5. Mẫu nhiều cấp: Nếu các phần tử của tổng thể phân tán quá rộng và thiếu thông tin về chúng, người ta thường chọn mẫu theo nhiều cấp. Khi chọn nhiều cấp, ta có nhiều loại đơn vị chọn mẫu ở mỗi cấp, thường được gọi là đơn vị chọn mẫu cấp 1, cấp 2, ... Để chọn mẫu ở mỗi cấp chỉ cần có thông tin về phân bố của dấu hiệu ở cấp ấy là đủ. Chẳng hạn, để

điều tra ý kiến của tổng thể khách hàng trong cả nước về sản phẩm của doanh nghiệp có thể chọn mẫu nhiều cấp như sau:

- Đơn vị mẫu cấp 1: Chọn ra các tỉnh, thành phố đại diện.

- Đơn vị mẫu cấp 2: Trong các tỉnh, thành phố đã chọn, chọn ra một số quận, huyện đại diện.

- Đơn vị mẫu cấp 3: Trong các quận, huyện đã chọn, chọn ra một số phường, xã đại diện...

Việc chọn mẫu ở mỗi cấp có thể tiến hành theo phương pháp mẫu ngẫu nhiên đơn, mẫu ngẫu nhiên hệ thống, mẫu chùm hay mẫu phân tổ.

3.3. Thang đo các giá trị mẫu

Biến ngẫu nhiên trong tổng thể nghiên cứu có thể là định tính hoặc định lượng, do đó mẫu rút ra từ tổng thể cũng gồm các giá trị định tính hoặc định lượng. Vì vậy, để biểu diễn các giá trị của dấu hiệu nghiên cứu trong tổng thể cũng như của mẫu phải dùng các thang đo khác nhau nhằm mục đích lượng hóa dấu hiệu nghiên cứu đó.

Trong kinh tế xã hội các thang đo được sử dụng là:

1. Thang định danh: Là việc đánh số những tính chất hoặc phạm trù cùng loại. Chẳng hạn, thang giới tính gồm hai phạm trù là nam [0] và nữ [1]. Màu sắc sản phẩm có thể có nhiều hơn hai phạm trù là xanh [1] đỏ [2] trắng [3] vàng [4]... Giữa các con số ở đây không có quan hệ hơn kém do đó không thể thực hiện các phép tính số học đối với chúng. Thang định danh thường chỉ dùng để đếm tần số của các hiện tượng xảy ra.

2. Thang thứ bậc: Là loại thang định danh mà giữa các phạm trù đã có quan hệ thứ bậc hơn kém. Chẳng hạn, để đặc trưng học vấn có thể dùng thang thứ bậc thất học [0], tiểu học [1], trung học [2], đại học trở lên [3]; để đặc trưng thái độ của khách hàng đối với giá sản phẩm có thể dùng thang đo: rẻ [1], vừa phải [2], đắt [3]. Đương nhiên sự sai khác giữa các phạm trù không bắt buộc phải đều nhau. Xét về mặt toán học, tập hợp các phạm trù đó đã được sắp xếp nhưng chưa có một metric.

3. Thang đo khoảng: Là thang đo thứ bậc có các khoảng cách đều nhau giữa các bậc. Một thang đo như vậy đã có kết cấu metric, có thể đánh giá sự khác biệt giữa các phạm trù bằng loại thang đo này mặc dù điểm gốc ở đây chỉ là tương đối. Với thang đo khoảng việc cộng và trừ các số đo mới bắt đầu có ý nghĩa, trên cơ sở đó có thể tính được các tham số đặc trưng như trung bình, phương sai v.v... Yêu cầu có khoảng cách đều nhau là đặt ra đối với thang đo, còn đối với lớp các hiện tượng được đo bằng thang này thì không bắt buộc phải đều nhau. Chẳng hạn, để đặc trưng lứa tuổi có thể dùng thang đo khoảng: Trẻ (dưới 35 tuổi) [30], trung niên (từ 36 tuổi đến 60 tuổi) [50], già (từ 60 tuổi trở lên) [70]. Để thu được thang đo khoảng có thể bắt đầu bằng thang thứ bậc sau đó chuẩn hóa sao cho các quãng cách đều nhau để được một thang đo khoảng, sao cho việc tính toán các trị số đo trở nên có ý nghĩa.

Các thang đo định danh, thứ bậc và thang đo khoảng dùng để đặc trưng các giá trị của dấu hiệu nghiên cứu định tính.

4. Thang đo tỷ lệ: Là thang đo khoảng với một điểm gốc tuyệt đối. Chỉ với thang đo tỷ lệ ta mới có thể đo lường các hiện tượng như các đơn vị đo lường vật lý thông thường và mới có thể thực hiện được tất cả các phép toán với các trị số đo theo nghĩa là lượng thông tin sẽ tăng lên cùng với trị số đo.

Thang đo tỷ lệ được dùng để đặc trưng các giá trị của dấu hiệu nghiên cứu định lượng.

Mỗi thang đo cấp cao hơn có thể chuyển xuống một thang đo cấp thấp hơn, chẳng hạn chuyển thang đo tỷ lệ thành thang đo thứ bậc nhưng ngược lại thì không được.

Khi nghiên cứu một dấu hiệu của tổng thể hay rút ra một mẫu từ tổng thể thì việc đầu tiên là phải xác định đúng thang đo cho các giá trị điều tra của tổng thể hay của mẫu làm cơ sở cho quá trình xử lý tiếp theo.

3.4. Các phương pháp mô tả số liệu mẫu

1. Giả sử từ tổng thể với biến ngẫu nhiên gốc X rút ra một mẫu cụ thể kích thước n, trong đó giá trị x_1 xuất hiện với tần số n_1, x_2 xuất hiện với tần số n_2, ..., x_k xuất hiện với tần số n_k (trong đó các giá trị của X có thể được đo bằng các thang khác nhau tùy thuộc vào việc χ là định tính hay định lượng). Lúc đó, sau khi các x_i đã được sắp xếp theo trình tự tăng dần giá trị cụ thể của mẫu ω có thể mô tả bằng bảng phân phối tần số thực nghiệm sau:

x_i	x_1	x_2	...	x_i	...	x_k
n_i	n_1	n_2	...	n_i	...	n_k

Hiển nhiên là $n_1 + n_2 + ... + n_k = n$

2. Nếu ký hiệu $f_i = \frac{n_i}{n}$ (6.17) là tần suất xuất hiện giá trị x_i trong mẫu thì lúc đó giá trị của mẫu w còn có thể mô tả bằng bảng phân phối tần suất thực nghiệm sau:

x_i	x_1	x_2	...	x_i	...	x_k
f_i	f_1	f_2	...	f_i	...	f_k

Từ (6.17) suy ra $f_1 + f_2 + ... + f_k = 1$

Thí dụ 1. Gặt ngẫu nhiên 365 điểm trồng lúa của một huyện thu được các số liệu được sắp xếp thành bảng sau:

Năng suất (tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm gặt tương ứng	6	13	38	74	106	85	30	10	3

Như vậy, giá trị của mẫu đã được mô tả dưới dạng bảng phân phối tần số thực nghiệm. Còn bảng phân phối tần suất thực nghiệm có dạng:

x_i	25	30	33	34	35	36	37	39	40
f_i	0,016	0,036	0,104	0,203	0,290	0,233	0,082	0,027	0,009

3. Nếu ký hiệu $\omega_i (i = \overline{1,k})$ là tần số tích lũy của x_i

$$\omega_i = \sum_{x_j < x_i} n_j \qquad (6.18)$$

và $F^*(x_i)$ là tần suất tích lũy của x_i

$$F^*(x_i) = \frac{\omega_i}{n} = \sum_{x_j < x_i} \frac{n_j}{n} \qquad (6.19)$$

thì $F^*(x_i)$ là một hàm của x_i và gọi là hàm phân bố thực nghiệm của mẫu.

Khác với hàm phân bố xác suất F(x) của biến ngẫu nhiên gốc X trong tổng thể xác định xác suất của biến cố $X < x$, hàm phân bố thực nghiệm $F^*(x)$ xác định tần suất của biến cố đó. Theo định lý Bernoulli, khi kích thước mẫu đủ lớn tần suất của biến cố $X < x$ tức là $F^*(x)$ sẽ hội tụ theo xác suất về xác suất F(x) của biến cố đó, tức là $F^*(x)$ và F(x) sai khác nhau không đáng kể. Từ đó có thể dùng hàm phân bố thực nghiệm của mẫu để biểu diễn một cách gần đúng quy luật phân phối gốc F(x) của tổng thể.

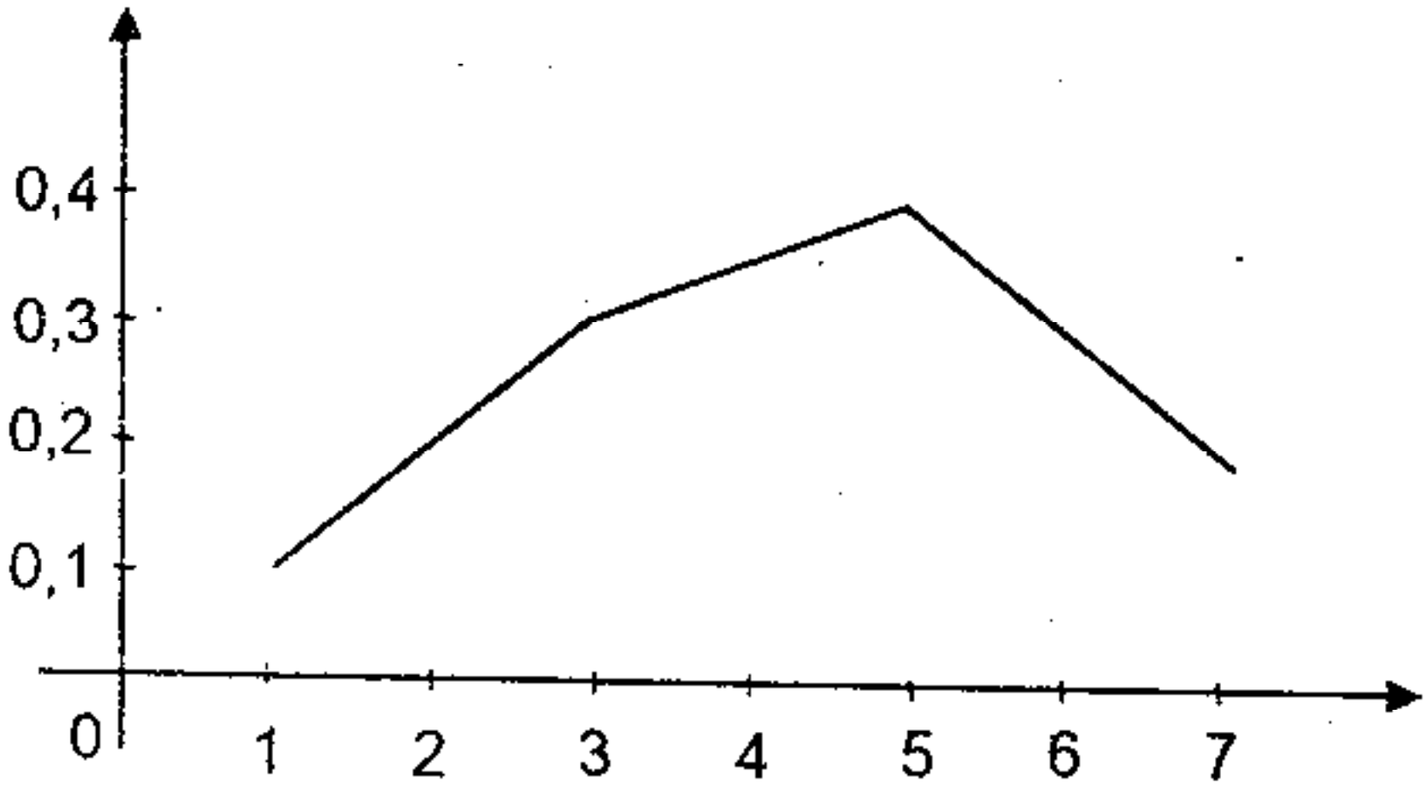
4. Để mô tả số liệu mẫu một cách rõ ràng hơn cho phép đưa ra những nhận xét sơ bộ ban đầu về tổng thể, người ta còn xây dựng các loại đồ thị khác nhau của phân phối thực nghiệm.

- *Đa giác tần số* là một đường gãy khúc mà các đoạn thẳng của nó nối các điểm (x_1, n_1), (x_2, n_2) ... (x_k, n_k) trên mặt phẳng.

- *Đa giác tần suất* là một đường gãy khúc mà các đoạn thẳng của nó nối các điểm (x_1, f_1) (x_2, f_2) ... (x_k, f_k) trên mặt phẳng.

Thí dụ 2. Vẽ đa giác tần suất của phân phối thực nghiệm sau:

x_i	1	3	5	7
f_i	0,1	0,3	0,4	0,2



***Hình 6.1.* Đa giác tần suất**

Đa giác tần suất thường được dùng để mô tả các số liệu mẫu theo thời gian.

- Khi dấu hiệu nghiên cứu có phân phối liên tục thì nên xây dựng *biểu đồ tần số* hoặc *biểu đồ tần suất*. Để làm điều đó khoảng chứa tất cả các giá trị quan sát được của mẫu được chia ra thành một số đoạn có chiều dài bằng h và tại mỗi đoạn đưa vào các tần số hoặc tần suất tương ứng với đoạn đó. Như vậy, biểu đồ tần số sẽ là một hình bậc thang tạo nên bởi nhiều hình chữ nhật có đáy bằng h và chiều cao bằng $\frac{n_i}{h}$.

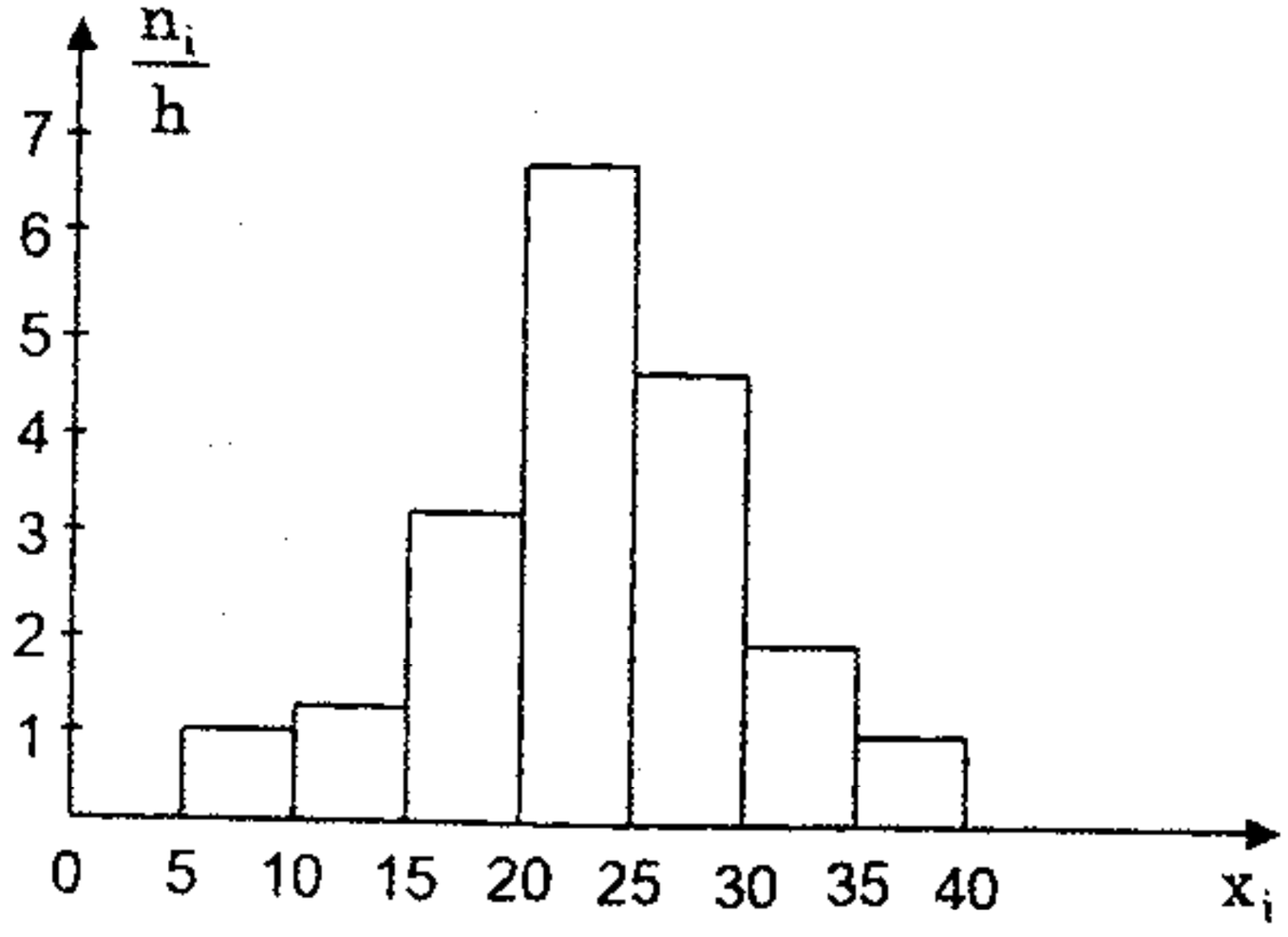
Lúc đó diện tích của hình chữ nhật thứ i bằng $h.\frac{n_i}{h} = n_i$ là tổng tần số ứng với đoạn thứ i, vì vậy diện tích của tất cả các hình chữ nhật sẽ bằng kích thước mẫu n.

Tương tự biểu đồ tần suất là một hình bậc thang tạo nên bởi nhiều hình chữ nhật có đáy bằng h và chiều cao bằng $\frac{f_i}{h}$. Lúc đó diện tích của hình chữ nhật thứ i bằng $h.\frac{f_i}{h} = f_i$ và diện tích của toàn bộ hình bậc thang đó sẽ bằng 1.

Thí dụ 3. Vẽ biểu đồ tần số của phân phối thực nghiệm cho ở bảng sau

Bảng 6.2

Đoạn giá trị chiều dài h = 5	Tổng các tần số tương ứng n_i	$\frac{n_i}{h}$
5 - 10	4	0,8
10 - 15	6	1,2
15 - 20	16	3,2
20 - 25	36	7,2
25 - 30	24	4,8
30 - 35	10	2,0
35 - 40	4	0,8



Hình 6.2. **Biểu đồ tần số**

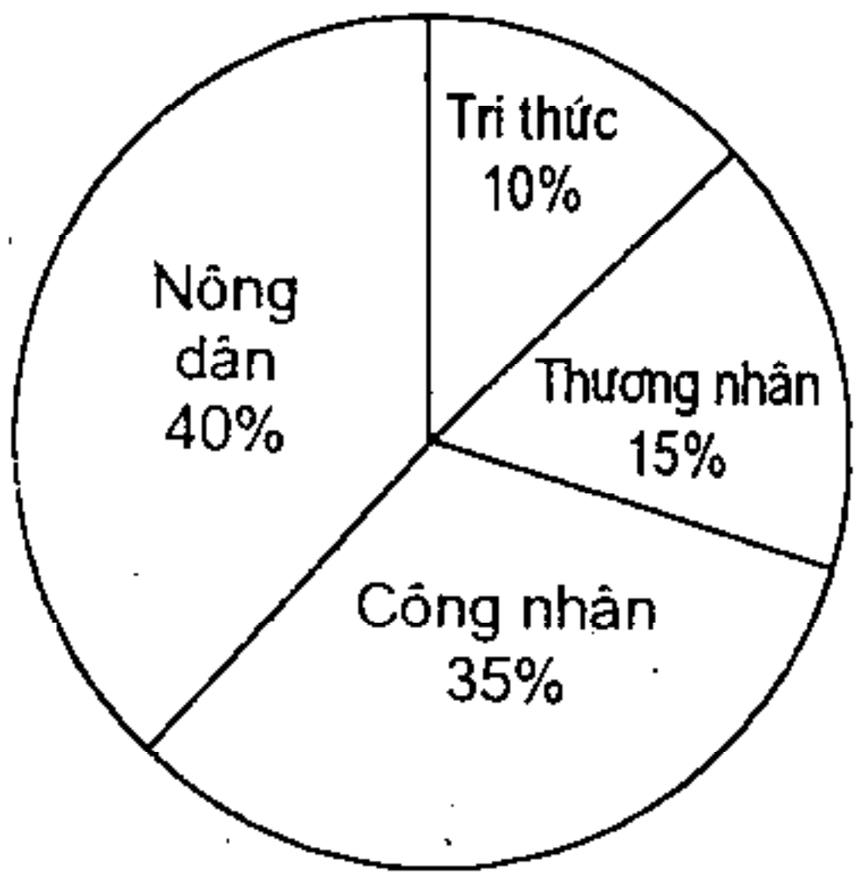
Với các dấu hiệu nghiên cứu là định tính thì người ta thường mô tả các số liệu mẫu bằng biểu đồ hình bánh xe. Đó là một hình tròn được chia ra thành nhiều bộ phận tương ứng với cơ cấu của các phạm trù xuất hiện trong mẫu.

Thí dụ 4. Điều tra ngẫu nhiên 100 khách hàng của doanh nghiệp thì thấy các khách hàng được phân theo tỷ lệ sau đây về tầng lớp xã hội (bảng 6.3).

Bảng 6.3

Tầng lớp xã hội	Số khách hàng	Tỷ lệ
Công nhân	35	0,35
Nông dân	40	0,40
Thương nhân	15	0,15
Trí thức	10	0,10
Tổng số	100	1

Vẽ biểu đồ hình bánh xe về cơ cấu của 100 khách hàng được điều tra (hình 6.3).



Hình 6.3. **Biểu đồ hình bánh xe**

Đồ thị của phân phối mẫu có thể vẽ dễ dàng nhờ việc sử dụng các phần mềm cho máy vi tính. Mọi phần mềm thống kê như Excel, SPSS, MFIT, Stata đều có chương trình cho phép vẽ các loại đồ thị trên.

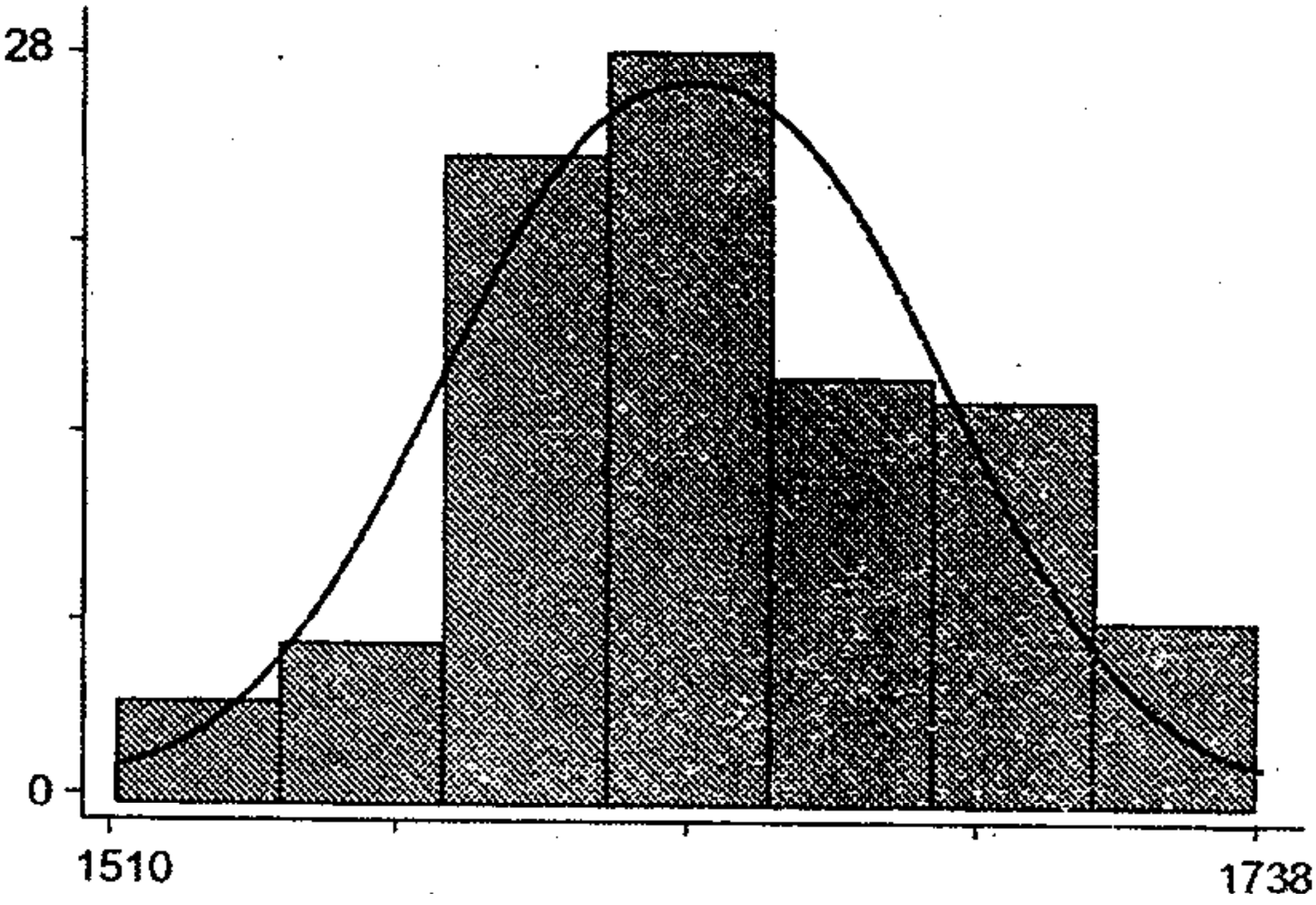
Thí dụ A. Từ ba vùng người ta điều tra ngẫu nhiên thu nhập hàng năm (tính bằng USD) của 300 người (mỗi vùng 100 người) đang làm việc tại các công ty tư nhân và thu được các số liệu sau:

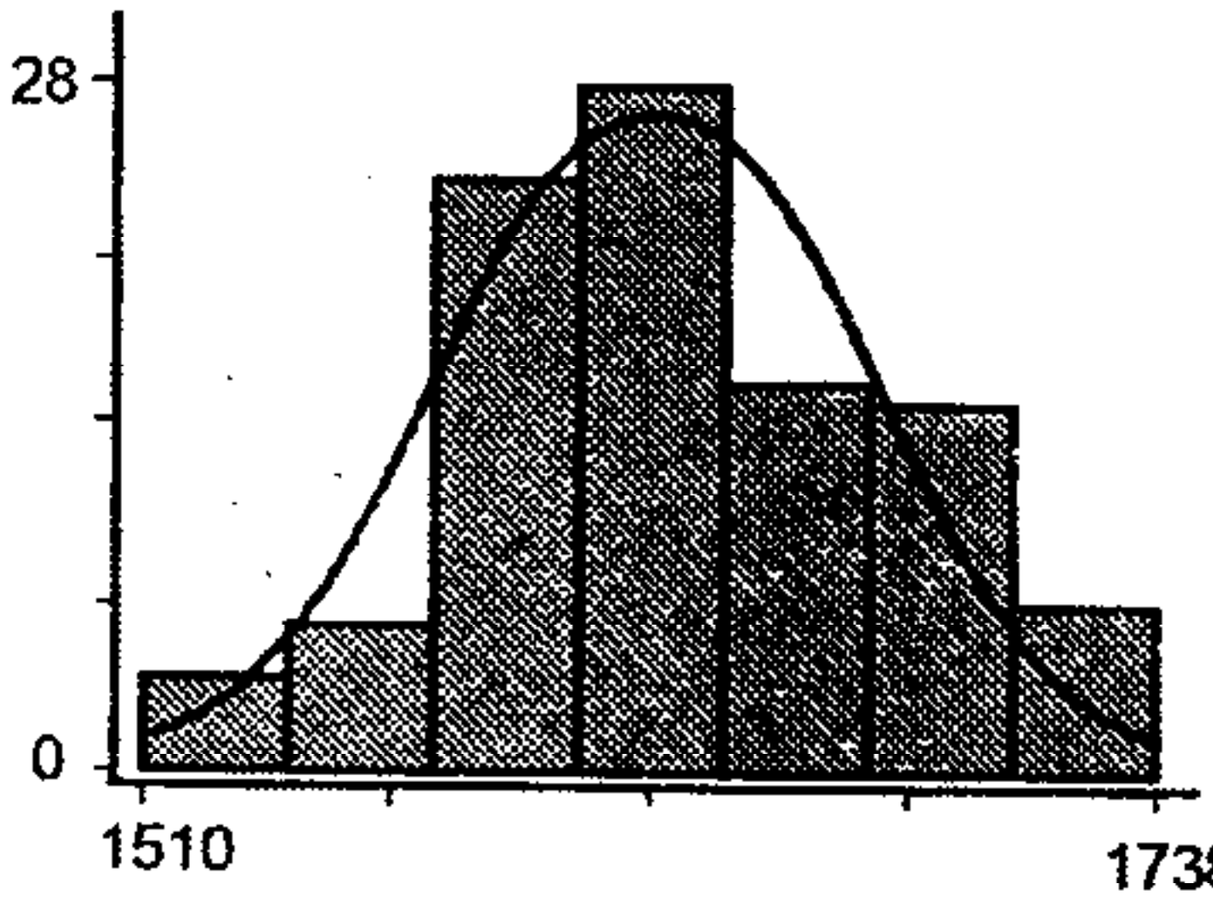
Vùng 1		Vùng 2		Vùng 3	
Thu nhập	Số người	Thu nhập	Số người	Thu nhập	Số người
1547	1	1557	1	1510	1
1553	1	1563	1	1525	1
1573	2	1576	1	1527	1
1575	1	1582	1	1530	1
1580	1	1586	2	1544	1
1582	1	1588	1	1547	1
1586	1	1590	2	1550	1
1592	1	1595	1	1560	1
1595	3	1596	1	1573	2
1597	4	1597	1	1578	1
1598	1	1598	2	1582	2
1599	2	1600	1	1583	2
1602	1	1603	1	1584	1
1607	2	1605	2	1585	1
1608	2	1610	2	1586	1
1610	1	1612	2	1588	2
1617	1	1613	1	1590	2
1620	3	1620	1	1592	1
1621	1	1624	1	1593	2

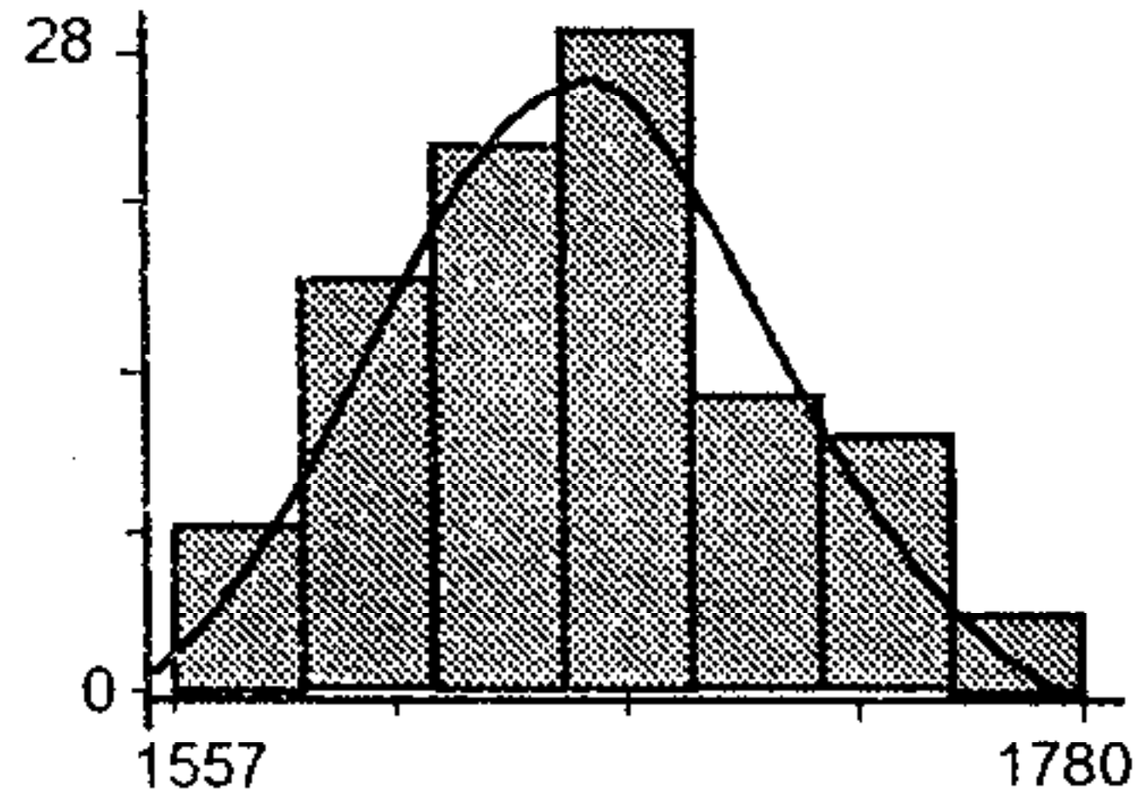
Vùng 1		Vùng 2		Vùng 3	
Thu nhập	Số người	Thu nhập	Số người	Thu nhập	Số người
1623	2	1625	2	1594	1
1624	1	1628	3	1597	2
1625	1	1629	1	1601	1
1626	1	1630	1	1602	2
1627	2	1632	1	1603	2
1630	3	1633	1	1607	1
1632	1	1637	2	1608	1
1633	1	1638	1	1610	3
1636	1	1640	2	1613	1
1637	3	1642	1	1614	1
1638	1	1643	1	1615	1
1640	2	1645	1	1616	2
1642	1	1648	2	1620	3
1644	1	1650	2	1622	1
1645	1	1652	1	1623	3
		1656	1	1625	4
1654	1	1657	2	1627	1
1655	3	1658	4	1630	1
1658	1	1660	2	1632	3
1660	1	1662	3	1633	2
1662	1	1663	1	1640	1
1667	1	1665	1	1642	2
1670	3	1666	2	1645	2
1672	1	1670	3	1648	1
1674	1	1672	1	1650	4
1675	2	1673	1	1652	1

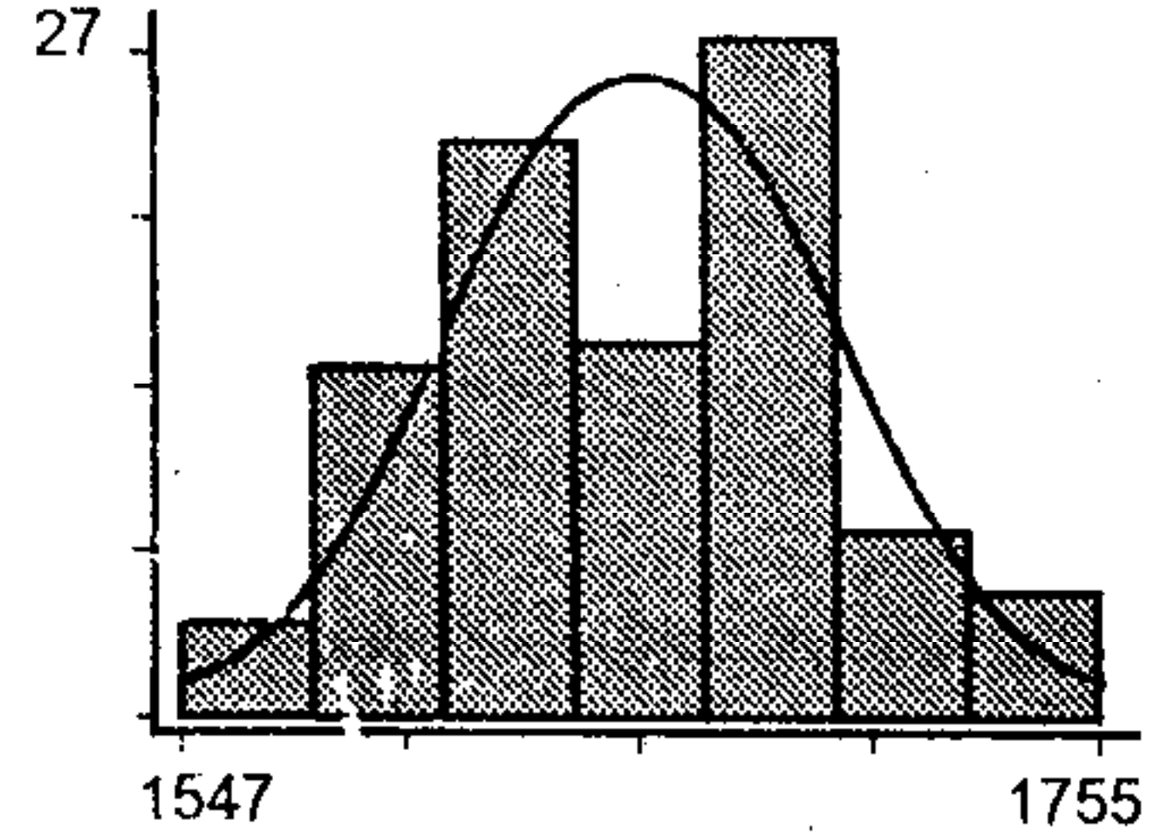
Vùng 1		Vùng 2		Vùng 3	
Thu nhập	Số người	Thu nhập	Số người	Thu nhập	Số người
1678	1	1675	2	1653	1
1680	3	1677	1	1654	1
1682	3	1678	1	1655	1
1683	2	1680	2	1660	2
1685	4	1682	1	1670	1
1687	1	1692	1	1674	1
1688	2	1694	2	1675	3
1690	1	1695	1	1678	1
1692	1	1696	1	1680	4
1693	1	1697	1	1690	1
1706	1	1700	3	1693	2
1708	1	1703	1	1698	1
1710	1	1707	2	1700	1
1712	1	1717	1	1704	1
1718	1	1720	1	1707	1
1720	2	1724	1	1710	1
1724	1	1730	1	1715	1
1737	1	1732	1	1720	1
1740	1	1734	1	1721	1
1747	2	1738	2	1736	1
1750	1	1740	2	1738	1
1755	1	1750	1		

Dùng phần mềm Stata ta vẽ được biểu đồ hình cột sau đây cho các số liệu của vùng 3 (được ghép lại thành 7 lớp thu nhập). Các vùng khác cũng vẽ tương tự.


28
0
1510
1738


28
0
1510
1738



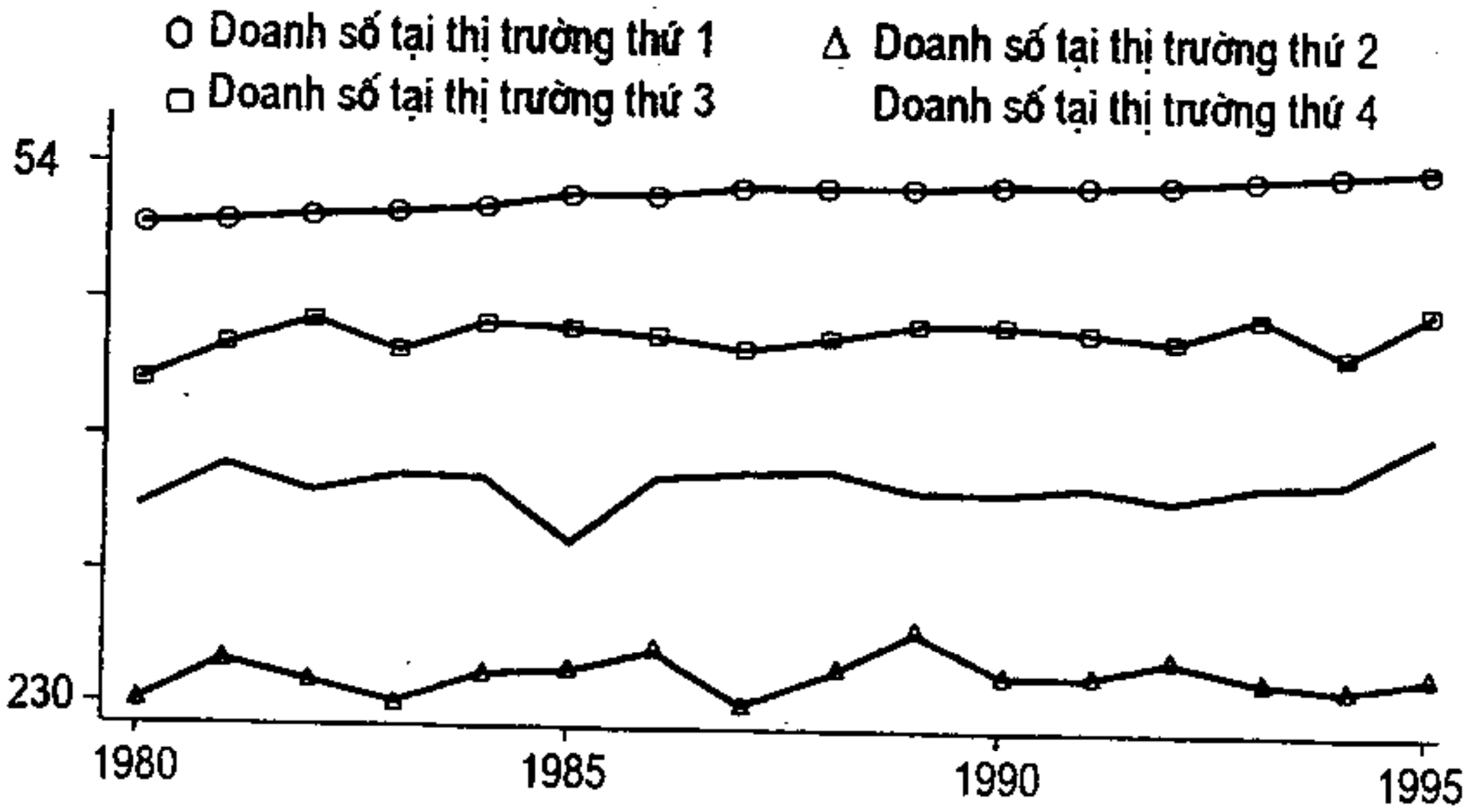


Để có thể so sánh sơ bộ phân phối thu nhập của 3 vùng ta có thể đặt các biểu đồ hình cột cạnh nhau.

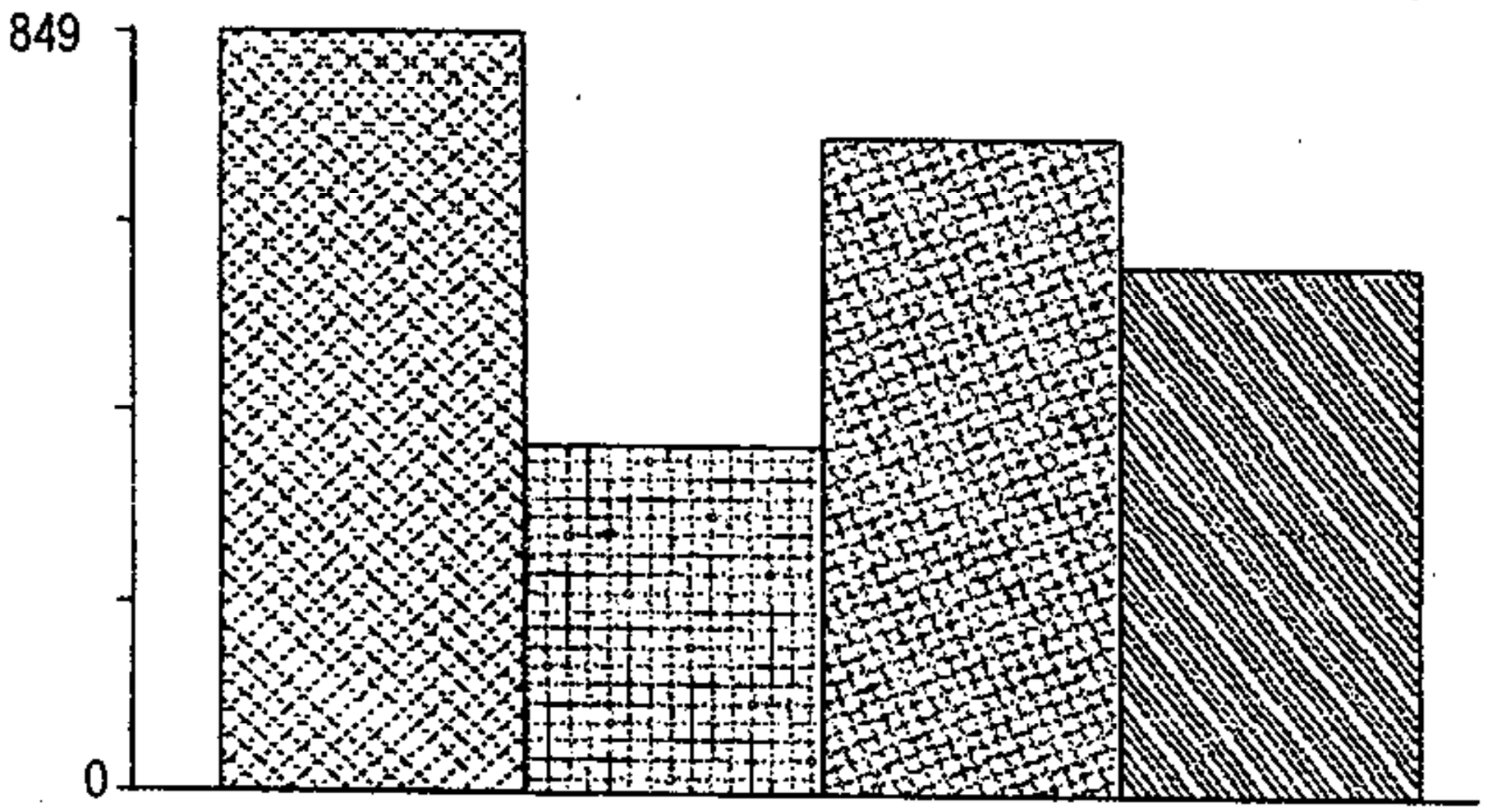
Thí dụ B. Một công ty hoạt động ở 4 thị trường. Thống kê doanh số X_1, X_2, X_3, X_4 của công ty tại 4 thị trường nói trên trong 16 năm thu được kết quả sau:

	x_1	x_2	x_3	x_4	Năm
1	510	240	428	345	1980
2	515	250	445	365	1981
3	520	240	460	355	1982
4	520	230	440	370	1983
5	522	245	460	370	1984
6	530	245	455	342	1985
7	530	255	455	370	1986
8	530	230	450	375	1987
9	535	245	460	375	1988
10	535	270	465	365	1989
11	540	245	465	365	1990
12	540	245	463	370	1991
13	540	255	460	360	1992
14	542	240	470	370	1993
15	543	235	455	373	1994
16	545	240	470	405	1995

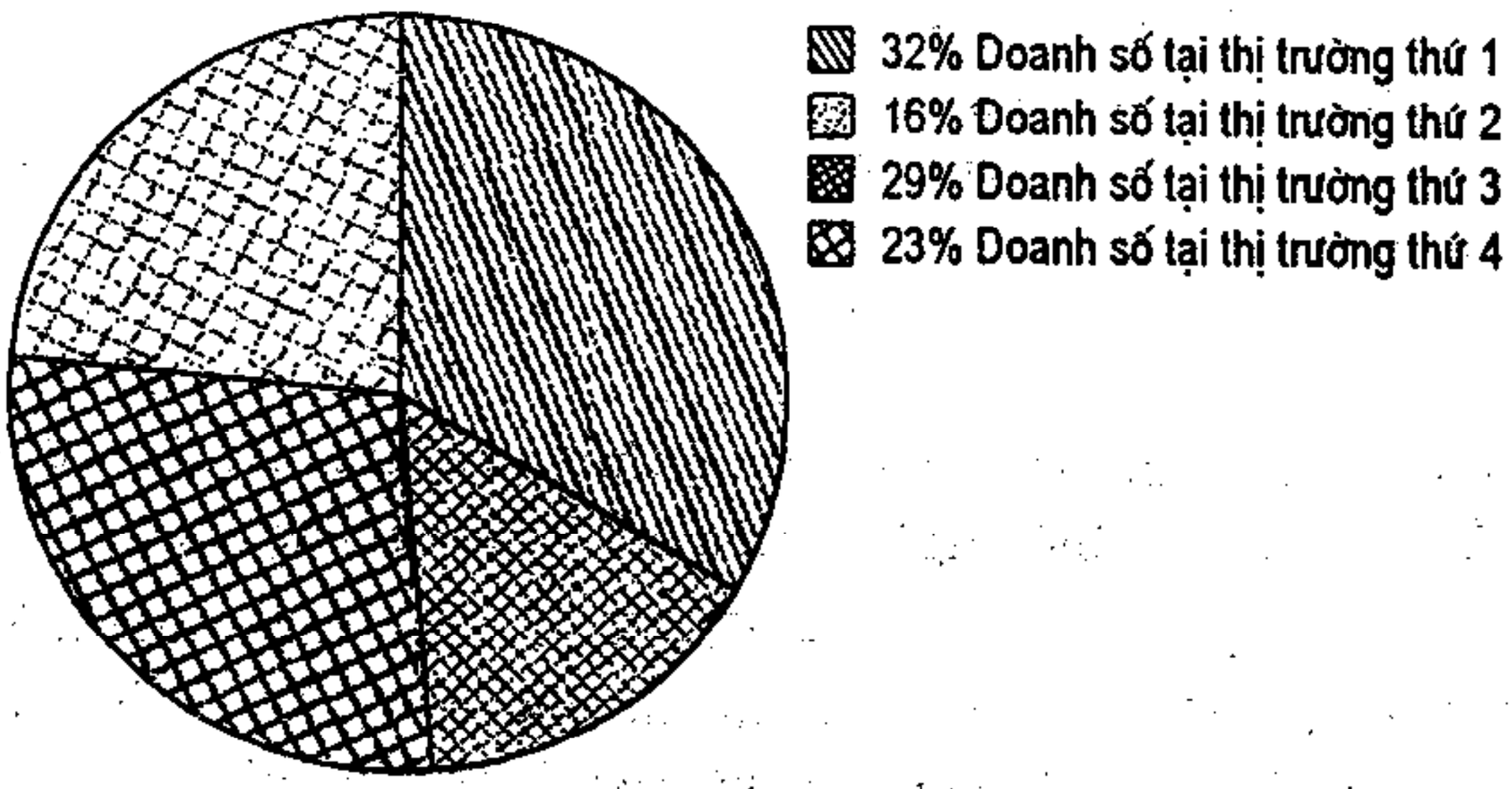
Dùng Stata vẽ được các đường đa giác mô tả doanh số của công ty tại 4 thị trường theo thời gian như sau:



Dùng Stata ta thu được biểu đồ hình cột mô tả cơ cấu của 4 thị trường trong tổng doanh số của công ty như sau:



Biểu đồ hình cột doanh số tại các thị trường



Biểu đồ hình bánh xe tỷ trọng doanh số tại các thị trường

Những phương pháp mô tả nói trên phản ánh từng giá trị của mẫu ngẫu nhiên. Chúng có thể được sử dụng nhằm các mục đích khác nhau. Trước hết là để sơ bộ phân tích bản thân các số liệu mẫu, chẳng hạn để lọc đi những giá trị quá "xa" xu hướng chung của dấu hiệu nghiên cứu. Đó là biện pháp tiếp theo để loại trừ các sai sót trong quá trình chọn mẫu và tăng thêm tính đại diện của nó. Sau nữa phân phối thực nghiệm của mẫu cho phép đưa ra những nhận xét sơ bộ về dấu hiệu nghiên cứu trong tổng thể làm cơ sở để tiếp tục khái quát hóa chúng.

§4. THỐNG KÊ

4.1. Định nghĩa

Để nghiên cứu biến ngẫu nhiên gốc X trong tổng thể, nếu chỉ rút ra một mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ thì mới chỉ có được một vài kết luận sơ bộ và rời rạc về X, vì các giá trị X_i của mẫu có cùng quy luật phân phối xác suất với X song quy luật này lại thường chưa được xác định hoàn toàn. Song nếu tổng hợp các biến ngẫu nhiên $X_1, X_2, ..., X_n$ này lại thì theo luật số lớn chúng sẽ bộc lộ những tính quy luật mới làm cơ sở để nhận định về biến ngẫu nhiên gốc X trong tổng thể.

Việc tổng hợp mẫu $W = (X_1, X_2, ..., X_n)$ được thực hiện dưới dạng một hàm nào đó của các giá trị $X_1, X_2, ..., X_n$ của mẫu. Nó được gọi là thống kê, ký hiệu là G. Như vậy

$$G = f(X_1, X_2, ..., X_n)$$

Như vậy về thực chất thống kê là một hàm của các biến ngẫu nhiên do đó bản thân nó cũng sẽ là một biến ngẫu nhiên tuân theo một quy luật phân phối xác suất nhất định và có các tham số đặc trưng như E(G), V(G)... Mặt khác, khi mẫu ngẫu nhiên nhận một giá trị cụ thể $\omega = (x_1, x_2, ..., x_n)$ thì G cũng nhận một giá trị cụ thể là:

$$g = f(x_1, x_2, ..., x_n)$$

Các thống kê cùng với quy luật phân phối xác suất của chúng là cơ sở để suy rộng các thông tin của mẫu cho dấu hiệu nghiên cứu của tổng thể.

Sau đây ta sẽ nghiên cứu một số thống kê thông dụng nhất.

4.2. Một số thống kê đặc trưng của mẫu ngẫu nhiên

Các thống kê đặc trưng của mẫu ngẫu nhiên cũng được chia thành ba loại:

- Các thống kê đặc trưng xu hướng trung tâm của phân phối của mẫu như trung bình mẫu, trung vị, mốt v.v...

- Các thống kê đặc trưng độ phân tán của phân phối của mẫu như khoảng biến thiên, phương sai, độ lệch chuẩn v.v...

- Các thống kê đặc trưng dạng phân phối.

Sau đây ta sẽ xem xét một số thống kê đặc trưng mẫu quan trọng nhất.

1. Trung bình mẫu: Giả sử từ biến ngẫu nhiên gốc X trong tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Trung bình mẫu là một thống kê, ký hiệu là $\overline{X}$ và là trung bình số học của các giá trị mẫu:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (6.20)$$

Ta chú ý rằng trung bình mẫu là một thống kê do đó nó là một biến ngẫu nhiên, tuân theo một quy luật phân phối xác suất nào đó với các tham số đặc trưng tương ứng. Khi mẫu ngẫu nhiên nhận một giá trị cụ thể $\omega = (x_1, x_2, ..., x_n)$ thì trung bình mẫu cũng nhận giá trị cụ thể bằng

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (6.21)$$

hoặc

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \qquad (6.22)$$

Trung bình mẫu $\overline{X}$ có tính chất sau: Nếu biến ngẫu nhiên gốc có kỳ vọng toán $E(X) = m$ và phương sai $V(X) = \sigma^2$ thì

$$E(\overline{X}) = m \qquad (6.23)$$

và

$$V(\overline{X}) = \frac{\sigma^2}{n} \qquad (6.24)$$

Thật vậy, theo định nghĩa

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

Áp dụng các tính chất của kỳ vọng toán ta có:

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

Song từ (6.15) ta có $E(X_i) = E(X) = m$ $(\forall i)$ do đó:

$$E(\overline{X}) = \frac{1}{n} nm = m$$

Mặt khác do X_i là các biến ngẫu nhiên độc lập, do đó theo tính chất của phương sai, ta có:

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$

và cũng từ (6.16) ta có $V(X_i) = V(X) = \sigma^2$ $(\forall i)$ nên

$$V(\overline{X}) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Vậy bất kể biến ngẫu nhiên gốc phân phối theo quy luật nào, trung bình mẫu $\overline{X}$ cũng có kỳ vọng toán bằng kỳ vọng toán của biến ngẫu nhiên gốc, tức là $E(\overline{X}) = E(X) = m$, còn phương sai $V(\overline{X})$ của nó nhỏ hơn n lần so với phương sai của biến ngẫu nhiên gốc, $V(\overline{X}) = \frac{\sigma^2}{n}$, nghĩa là các giá trị có thể có của $\overline{X}$ ổn định quanh kỳ vọng toán m hơn các giá trị có thể có của X.

Nếu lấy căn bậc hai của $V(\overline{X})$ ta thu được độ lệch chuẩn

$$\sqrt{V(\overline{X})} = \frac{\sigma}{\sqrt{n}}$$

Độ lệch chuẩn này của $\overline{X}$ thường được dùng để phản ánh sai số ước lượng do đó người ta thường gọi là sai số chuẩn Se của trung bình mẫu $\overline{X}$. Vậy

$$Se(\overline{X}) = \frac{\sigma}{\sqrt{n}} \qquad (6.25)$$

Ở trên ta đã luôn giả thiết rằng mẫu được rút ra từ tổng thể theo phương thức có hoàn lại. Nếu kích thước tổng thể là vô hạn hoặc kích thước tổng thể hữu hạn song $n < 0,1N$ thì có thể lấy mẫu không hoàn lại mà không ảnh hưởng đến kết

quả. Song nếu n > 0,1N thì trong các công thức trên phải sử dụng hệ số hiệu chỉnh do mẫu là không lặp (xem mục 4 chương III). Lúc đó:

$$V(\overline{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \qquad (6.26)$$

và

$$Se(\overline{X}) = \sqrt{\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}} \qquad (6.27)$$

2. Trung vị: Trung vị, ký hiệu là X_d là giá trị nằm ở chính giữa tức là giá trị chia các số liệu mẫu thành hai phần bằng nhau.

Đối với một mẫu cụ thể, trung vị được xác định như sau:

a. Nếu các số liệu mẫu gồm n giá trị rời rạc được sắp xếp theo trình tự tăng dần và nếu n là một số lẻ thì trung vị là giá trị thứ $\frac{n+1}{2}$ trong dãy số liệu đó.

Thí dụ 1. Giả sử có các số liệu mẫu sau:

240 220 210 225 235 225 270 250 280

Ta có n = 9 số liệu nên trung vị là giá trị thứ $\frac{9+1}{2} = 5$ trọng dãy số liệu đã được xếp theo thứ tự tăng dần:

210 220 225 225 235 240 250 270 280

↑

X_d

Còn nếu n là một số chẵn thì trung vị là hai giá trị nằm chính giữa của dãy số liệu đó. Nó được gọi là khoảng trung vị.

Thí dụ 2. Giả sử mẫu có thêm số liệu 200 tức là n = 10 do đó hai giá trị nằm chính giữa là giá trị thứ 5 và thứ 6. Tức là

200 210 220 225 225 235 240 250 270 280

$$X_d = [225 - 235]$$

b. Nếu các số liệu mẫu được ghép lớp theo phân phối tần số thì giá trị trung vị có thể tính gần đúng bởi công thức sau:

$$X_d \approx L + \frac{\left(\frac{n}{2} - S\right)}{n_{X_d}} h \qquad (6.28)$$

trong đó: L - Giới hạn dưới của lớp chứa trung vị

n - Kích thước mẫu

S - Tổng tần số của các lớp đứng trước lớp chứa trung vị

n_{X_d} - Tần số của lớp chứa trung vị

h - Độ dài của lớp chứa trung vị

Thí dụ 3. Tìm trung vị của mẫu được cho bởi phân phối thực nghiệm trong bảng 6.4.

Bảng 6.4

Đoạn giá trị chiều dài h = 5	Tần số n_i	Tần số tích lũy w_i
5 - 10	4	4
10 - 15	6	10
15 - 20	16	26
20 - 25	36	62
25 - 30	24	86
30 - 35	10	96
35 - 40	4	100
Tổng số	n = 100	

Ta có n/2 = 100/2 = 50 vậy trung vị nằm ở lớp thứ tư căn cứ vào cột tần số tích lũy.

Từ đó

$$X_d \approx 20 + \frac{50-26}{36}\,5 = 23{,}33$$

Trung vị, cũng như trung bình mẫu, phản ánh xu hướng trung tâm của phân phối mẫu song nó có đặc điểm là không san bằng các chênh lệch giữa các giá trị của mẫu do đó nó thường được dùng để bổ sung hoặc thay thế trung bình mẫu khi không có đủ số liệu để tính.

Người ta đã chứng minh được rằng nếu biến ngẫu nhiên X trong tổng thể là liên tục và mẫu rút ra có kích thước khá lớn thì thống kê trung vị X_d sẽ có kỳ vọng toán chính bằng trung bình tổng thể.

$$E(X_d) = m$$

và phương sai được xác định bằng biểu thức

$$V(X_d) = \frac{1}{4n[f(x_d)]^2}$$

trong đó n là kích thước mẫu, còn $f(X_d)$ là hàm mật độ xác suất tại điểm trung vị X_d. Chẳng hạn nếu X trong tổng thể phân phối chuẩn thì trung vị trùng với trung bình và

$$f(X_d) = f(\mu)\frac{1}{\sigma\sqrt{2\pi}}$$

do đó:

$$V(X_d) = \frac{1}{4n[f(x_d)]^2} = \frac{2\pi\sigma^2}{4n} = 1{,}57\frac{\sigma^2}{n} = 1{,}57V(\overline{X})$$

Ta có n/2 = 100/2 = 50 vậy trung vị nằm ở lớp thứ tư căn cứ vào cột tần số tích lũy.

Từ đó

$$X_d \approx 20 + \frac{50-26}{36}\,5 = 23{,}33$$

Trung vị, cũng như trung bình mẫu, phản ánh xu hướng trung tâm của phân phối mẫu song nó có đặc điểm là không san bằng các chênh lệch giữa các giá trị của mẫu do đó nó thường được dùng để bổ sung hoặc thay thế trung bình mẫu khi không có đủ số liệu để tính.

Người ta đã chứng minh được rằng nếu biến ngẫu nhiên X trong tổng thể là liên tục và mẫu rút ra có kích thước khá lớn thì thống kê trùng vị X_d sẽ có kỳ vọng toán chính bằng trung bình tổng thể.

$$E(X_d) = m$$

và phương sai được xác định bằng biểu thức

$$V(X_d) = \frac{1}{4n[f(x_d)]^2}$$

trong đó n là kích thước mẫu, còn $f(X_d)$ là hàm mật độ xác suất tại điểm trung vị X_d. Chẳng hạn nếu X trong tổng thể phân phối chuẩn thì trung vị trùng với trung bình và

$$f(X_d) = f(\mu)\,\frac{1}{\sigma\sqrt{2\pi}}$$

do đó:

$$V(X_d) = \frac{1}{4n[f(x_d)]^2} = \frac{2\pi\sigma^2}{4n} = 1{,}57\frac{\sigma^2}{n} = 1{,}57V(\bar{X})$$

d_2 - Hiệu số giữa tần số của lớp chứa mốt và tần số của lớp đứng sau;

h - Độ dài của lớp chứa mốt.

Thí dụ 5. Với các số liệu mẫu cho trong bảng 6.2 hãy tìm giá trị mốt.

Từ bảng 6.2 ta tìm thấy ngay lớp chứa mốt là lớp thứ 4.

Từ đó $X_0 \approx 20 + \left(\frac{20}{20+12}\right) \cdot 5 = 23{,}125$

Cũng như trung vị, mốt không san bằng, bù trừ chênh lệch giữa các giá trị của mẫu, do đó nó bổ sung hoặc thay thế trung bình mẫu khi việc tính trung bình mẫu gặp khó khăn.

Trung bình, Trung vị và Mốt là các tham số chủ yếu đặc trưng cho xu hướng trung tâm của mẫu. Sau đây ta sẽ nghiên cứu một vài tham số đặc trưng cho độ phân tán của các giá trị của mẫu.

4. Khoảng biến thiên. Khoảng biến thiên, ký hiệu là R là sai lệch giữa giá trị lớn nhất và nhỏ nhất của mẫu.

$$R = X_{max} - X_{min} \tag{6.30}$$

Nếu các số liệu mẫu được ghép lớp thì khoảng biến thiên là hiệu số giữa cận trên của lớp cuối cùng với cận dưới của lớp đầu tiên trong dãy phân phối các giá trị của mẫu.

Việc tính khoảng biến thiên khá đơn giản song không mang lại nhiều thông tin về độ phân tán của các giá trị mẫu.

5. Khoảng tứ phân vị: Trong phân tích kinh tế, xã hội nhiều khi phải tính đến thứ bậc của các đơn vị nghĩa là chia các đơn vị của số liệu mẫu trong bảng phân phối thành các phần bằng nhau. Nếu mẫu được chia thành 4 phần bằng

nhau thì có tứ phân vị. Tứ phân vị đầu là giá trị của mẫu đứng ở vị trí cách đơn vị đầu tiên 1/4 số đơn vị của mẫu. Tứ phân vị thứ hai chính là trung vị. Tứ phân vị thứ ba là giá trị của mẫu đứng ở vị trí cách đơn vị đầu tiên 3/4 số đơn vị của mẫu.

Chẳng hạn với các số liệu mẫu

200 210 220 225 225 235 240 250 270 280

thì tứ phân vị đầu là giá trị nằm ở vị trí thứ n/4 = 10/4 = 2,5. Song do thứ tự vị trí phải là nguyên do đó nó là giá trị nằm ở vị trí thứ ba.

$$Q_1 = 220$$

Tứ phân vị thứ ba là giá trị nằm ở vị trí thứ 3n/4 = 3.10/4 = 7,5 do đó nó là giá trị nằm ở vị trí thứ 8.

$$Q_3 = 250$$

Nếu các số liệu mẫu được ghép lớp thì các tứ phân vị được tính gần đúng theo các công thức sau:

$$Q_1 = L_{Q_1} + \frac{\frac{n}{4} - S_{Q_1}}{n_{Q_1}} h_{Q_1} \qquad (6.31)$$

$$Q_3 = L_{Q_3} + \frac{\frac{3n}{4} - S_{Q_3}}{n_{Q_3}} h_{Q_3} \qquad (6.32)$$

trong đó:

L_{Q_1} và L_{Q_3} - Các giới hạn dưới của các lớp chứa Q_1 và Q_3.

S_{Q_1} và S_{Q_3} - Tổng các tần số của các lớp đứng trước lớp chứa Q_1 và Q_3.

n_{Q_1} và n_{Q_3} - Tần số của các lớp chứa Q_1 và Q_3.

h_{Q_1} và h_{Q_3} - Độ dài của các lớp chứa Q_1 và Q_3.

Chẳng hạn với các số liệu của bảng 6.3 ta tìm được lớp chứa Q_1 là lớp thứ ba (100/4 = 25) do đó:

$$Q_1 = 15 + \frac{25-10}{16}5 = 19,6875$$

Lớp chứa Q_3 là lớp thứ năm (3.100/4 = 75) do đó:

$$Q_3 = 25 + \frac{75-62}{24}5 = 27,7083$$

Lúc đó khoảng tứ phân vị, ký hiệu là IQR được xác định bằng biểu thức:

$$IQR = Q_3 - Q_1 \qquad (6.33)$$

Khoảng tứ phân vị tuy đã nhạy cảm hơn đối với số liệu mẫu so với khoảng biến thiên song nó chỉ có nhiều ý nghĩa khi so sánh hai mẫu, còn với từng mẫu thì cũng không có nhiều ý nghĩa đặc trưng.

6. Tổng bình phương các sai lệch và độ lệch bình phương trung bình

Cho mẫu ngẫu nhiên được xây dựng từ biến ngẫu nhiên gốc X

$$W = (X_1, X_2, ..., X_n)$$

Lúc đó tổng bình phương các sai lệch giữa các giá trị của mẫu và trung bình mẫu được ký hiệu là SS và bằng

$$SS = \sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad (6.34)$$

Giá trị SS thường được sử dụng trong phân tích phương sai.

Nếu đem chia SS cho kích thước mẫu ta thu được trung bình số học của tổng bình phương sai lệch giữa các giá trị của mẫu và trung bình mẫu gọi tắt là *độ lệch bình phương trung bình*, ký hiệu là MS:

$$MS = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad (6.35)$$

Ta chú ý rằng, cũng giống như trung bình mẫu, SS và MS là các biến ngẫu nhiên với quy luật phân phối xác suất xác định và các tham số đặc trưng tương ứng. Còn trên một mẫu cụ thể chúng sẽ nhận những giá trị cụ thể.

Trong thực tế để tiện cho việc tính toán MS thường được tính bằng công thức sau:

$$MS = \frac{1}{n}\sum_{i=1}^{k} n_i X_i^2 - \overline{X}^2 \qquad (6.36)$$

Bạn đọc có thể tự chứng minh kết quả trên.

MS có tính chất sau: Nếu biến ngẫu nhiên gốc có $E(X) = m$ và $V(X) = \sigma^2$ thì

$$E(MS) = \frac{n-1}{n}\sigma^2 \qquad (6.37)$$

Để chứng minh, trước hết ta biến đổi công thức của MS như sau:

$$MS = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i - m) - (\overline{X} - m)]^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}[(X_i - m)^2 - 2(\overline{X} - m)(X_i - m) + (\overline{X} - m)^2] =$$

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

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

Do đó

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

Do $E(X_i - m)^2 = V(X_i) = V(X) = \sigma^2$

và $E(\overline{X} - m)^2 = V(\overline{X}) = \frac{\sigma^2}{n}$

$\Rightarrow$ $E(MS) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$

7. Phương sai mẫu S^2 và phương sai S^{*2}

- *Phương sai mẫu*, ký hiệu là S^2 được xác định bằng công thức:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right] \qquad (6.38)$$

- *Phương sai* S^{*2} được xác định bằng biểu thức:

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 \qquad (6.39)$$

trong đó m là trung bình tổng thể.

Dễ dàng thấy rằng giữa phương sai mẫu và MS có mối liên hệ sau:

$$S^2 = \frac{n}{n-1}MS \qquad (6.40)$$

Cũng giống như MS, phương sai mẫu S^2 và phương sai S^{*2} thực chất là các biến ngẫu nhiên với các quy luật phân phối xác suất tương ứng, còn giá trị của chúng trên một giá trị cụ thể của mẫu là những số xác định, ký hiệu là s^2 và s^{*2}.

Từ tính chất của MS có thể suy ra tính chất sau đây của phương sai mẫu:

$$E(S^2) = \sigma^2 \tag{6.41}$$

Thật vậy, theo (6.40) ta có:

$$E(S^2) = E\left(\frac{n}{n-1} MS\right) = \frac{n}{n-1} E(MS)$$

Thay giá trị của E(MS) từ (6.37) vào ta có:

$$E(S^2) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

Đối với phương sai S^{*2} ta cũng có tính chất tương tự sau:

$$E(S^{*2}) = \sigma^2 \tag{6.42}$$

Thật vậy

$$E(S^{*2}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2\right] = \frac{1}{n}\sum_{i=1}^{n}E(X_i - m)^2$$

Do $E(X_i - m)^2 = V(X_i) = V(X) = \sigma^2$

nên $E(S^{*2}) = \sigma^2$

Nếu lấy căn bậc hai của phương sai mẫu S^2, ta thu được thống kê gọi là *độ lệch chuẩn mẫu*, ký hiệu là S. Như vậy:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{6.43}$$

Còn giá trị của nó trên một giá trị cụ thể của mẫu là một số xác định, ký hiệu là s.

8. *Hệ số biến thiên*

Hệ số biến thiên, ký hiệu là CV là biểu thức:

$$CV = \left|\frac{S}{\overline{X}}\right| 100 \qquad (6.44)$$

Hệ số biến thiên được đo bằng phần trăm và được dùng để nhận xét về độ thuần nhất của phân phối mẫu và qua đó đo mức độ đại diện của trung bình mẫu cho xu hướng trung tâm của phân phối. Nếu CV < 15% thì mẫu được xem là khá thuần nhất.

9. *Hệ số bất đối xứng*

Trên mẫu, hệ số bất đối xứng được xác định bằng biểu thức:

$$a_3 = \frac{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^3}{n}}{S^3} \qquad (6.45)$$

Trong nhiều trường hợp hệ số bất đối xứng chỉ cần xác định một cách gần đúng bằng công thức:

$$a_3 \approx \frac{3(\overline{X} - X_d)}{S}$$

Giá trị của a_3 càng gần 0 thì phân phối thực nghiệm của các giá trị của mẫu càng đối xứng qua giá trị trung bình mẫu.

10. *Hệ số nhọn*

Trên mẫu hệ số nhọn được xác định bằng biểu thức:

$$a_4 = \frac{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^4}{n}}{S^4} \qquad (6.46)$$

11. Tần suất mẫu

Giả sử từ tổng thể kích thước N, trong đó M phần tử mang dấu hiệu nghiên cứu, lấy ra một mẫu ngẫu nhiên kích thước n và trong đó thấy có X phần tử mang dấu hiệu nghiên cứu. Lúc đó tần suất mẫu là một thống kê, ký hiệu là f và là tỷ số giữa số phần tử mang dấu hiệu nghiên cứu trong mẫu và kích thước mẫu:

$$f = \frac{X}{n} \tag{6.47}$$

Về thực chất, thống kê f cũng là một biến ngẫu nhiên vì nó là hàm của biến ngẫu nhiên X - số lần xuất hiện dấu hiệu trong mẫu, tức là trong n phép thử độc lập. Còn giá trị của nó trên một giá trị cụ thể của mẫu là một số xác định.

Ta chú ý rằng, cũng giống như tần suất của tổng thể là trường hợp riêng của trung bình tổng thể, tần suất mẫu là một trường hợp riêng của trung bình mẫu $\overline{X}$ khi xem dấu hiệu nghiên cứu trong tổng thể như biến ngẫu nhiên tuân theo quy luật không - một.

Với nhận xét đó có thể chứng minh tính chất sau đây của tần suất mẫu: Nếu biến ngẫu nhiên gốc tuân theo quy luật A(p) với E(X) = p và V(X) = p(1 – p) thì:

$$E(f) = p \tag{6.48}$$

$$V(f) = \frac{p(1-p)}{n} \tag{6.49}$$

Thật vậy, theo định nghĩa của tần suất mẫu:

$$E(f) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X)$$

Song X là số lần xuất hiện dấu hiệu nghiên cứu trong mẫu, tức là số lần xuất hiện biến cố "dấu hiệu nghiên cứu" trong n phép thử độc lập nên X là biến ngẫu nhiên phân phối theo quy luật nhị thức với các tham số là E(X) = np và V(X) = np(1 – p) vì vậy:

$$E(f) = \frac{1}{n} np = p$$

Tương tự:

$$V(f) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

Từ đó sai số chuẩn của tần suất mẫu là:

$$Se(f) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Nếu mẫu lấy ra theo phương pháp không hoàn lại thì sai số chuẩn của mẫu là:

$$Se(f) = \sqrt{\frac{N-n}{N-1} \cdot \frac{p(1-p)}{n}} \qquad (6.50)$$

Việc tính giá trị của tần suất mẫu f trên một mẫu cụ thể nói chung không có gì khó khăn, còn để tính trung bình mẫu $\overline{X}$, MS và các phương sai S^2, S^{*2} có thể gặp khó khăn khi có nhiều số liệu, do đó người ta thường lập bảng để tính toán cho tiện lợi hơn.

Thí dụ 6. Lấy ngẫu nhiên 100 sản phẩm do một máy sản xuất ra thấy có 5 phế phẩm. Tìm tỷ lệ phế phẩm của mẫu nói trên. Như vậy, dấu hiệu nghiên cứu trong tổng thể là "tính chất phế phẩm" của sản phẩm. Dấu hiệu này xuất hiên 5 lần

trong mẫu kích thước bằng 100. Vậy ta có tần suất mẫu, theo (6.47) bằng

$$f = \frac{5}{100} = 0{,}05$$

Thí dụ 7. Trở lại thí dụ về điều tra năng suất lúa. Gặt ngẫu nhiên 365 điểm trồng lúa của huyện thu được bảng số liệu sau:

Năng suất (tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm gặt tương ứng	6	13	38	74	106	85	30	10	3

Dấu hiệu nghiên cứu trong tổng thể là năng suất lúa trên một ha canh tác. Vậy các số liệu của mẫu nói trên cho phép xác định các thống kê đặc trưng của mẫu như trung bình mẫu (năng suất trung bình), phương sai mẫu (độ phân tán của năng suất) v.v...

Để tiện cho việc tính toán có thể lập bảng tính toán sau (bảng 6.6).

Từ đó, theo (6.22), ta có:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{12700}{365} = 34{,}795 \text{ tạ/ha}$$

và theo (6.36)

$$MS = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 = \frac{443466}{365} - (34{,}795)^2$$

$$= 1214{,}975 - 1210{,}692 = 4{,}283$$

Từ đó, theo (6.40) ta có:

$$s^2 = \frac{n}{n-1} MS = \frac{365}{364}.4{,}283 = 4{,}295$$

và

$$s = \sqrt{4{,}295} = 2{,}072$$

Bảng 6.6

x_i	n_i	$x_i n_i$	$n_i x_i^2$
25	6	150	3750
30	13	390	11700
33	38	1254	41382
34	74	2516	85544
35	106	3710	129850
36	85	3060	110160
38	30	1110	41070
39	10	390	15210
40	3	120	4800
	$\Sigma n_i = n = 365$	$\Sigma n_i x_i = 12700$	$\Sigma n_i x_i^2 = 443466$

Nếu các giá trị thu được của mẫu sai khác nhau rất ít hoặc khi có quá nhiều giá trị, người ta thường xây dựng bảng phân phối tần số ghép lớp.

Lúc đó, để tính các thống kê đặc trưng của mẫu, người ta lấy giá trị giữa của mỗi lớp đại diện cho lớp đó để tính toán.

Thí dụ 8. Lấy ngẫu nhiên 100 thanh niên ở một tỉnh đem đo chiều cao và thu được các số liệu sau:

Chiều cao (cm)	154-158	158-162	162-166	166-170	170-174	174-178	178-182
Số thanh niên có chiều cao tương ứng	10	14	26	28	12	8	2

Ở đây dấu hiệu nghiên cứu là chiều cao thanh niên. Để xác định các thống kê mẫu như trung bình mẫu, phương sai mẫu... ta lập bảng tính, trong đó các lớp giá trị x_i được thay bằng giá trị giữa của mỗi lớp (bảng 6.7).

Bảng 6.7

x_i	n_i	$x_i n_i$	$n_i x_i^2$
156	10	1560	243360
160	14	2240	358400
164	26	4264	699296
168	28	4704	790272
172	12	2064	355008
176	8	1408	247808
180	2	360	64800
	$\Sigma n_i = n = 100$	$\Sigma n_i x_i = 16600$	$\Sigma n_i x_i^2 = 2758944$

Vậy
$$\bar{x} = \frac{16600}{100} = 166\text{cm}$$

$$MS = \frac{2758944}{100} - 166^2 = 33{,}44$$

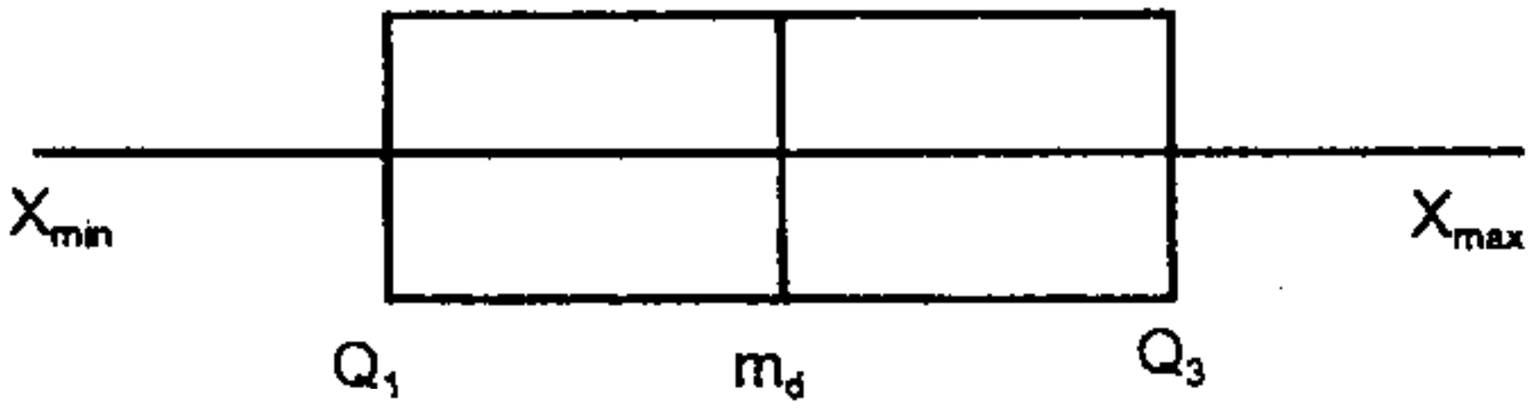
$$s = \sqrt{\frac{100}{99}} 33{,}44 = 5{,}812$$

Khi kích thước mẫu khá lớn thì việc tính toán thủ công các thống kê đặc trưng mẫu sẽ gặp khó khăn. Lúc đó nên sử dụng các phần mềm như Excel, SPSS, Stata, MFIT v.v... để tính các thống kê nói trên một cách nhanh chóng.

4.3. Đồ thị hình hộp (Box-Plot)

Đồ thị hình hộp là phương pháp mô tả và tổng hợp các số liệu mẫu bằng đồ thị, trên đó phản ánh được cùng một lúc cả các đặc trưng về xu hướng trung tâm cũng như độ phân tán của các giá trị của mẫu.

Để xây dựng đồ thị hình hộp người ta sử dụng các thống kê đặc trưng mẫu là trung vị, các tứ phân vị Q_1, Q_3 và các giá trị x_{max} và x_{min} của phân phối mẫu. Nó có dạng như sau (hình 6.4).



***Hình 6.4.* Đồ thị hình hộp**

Thí dụ 15. Vẽ đồ thị hình hộp với các số liệu mẫu cho trong bảng 6.3.

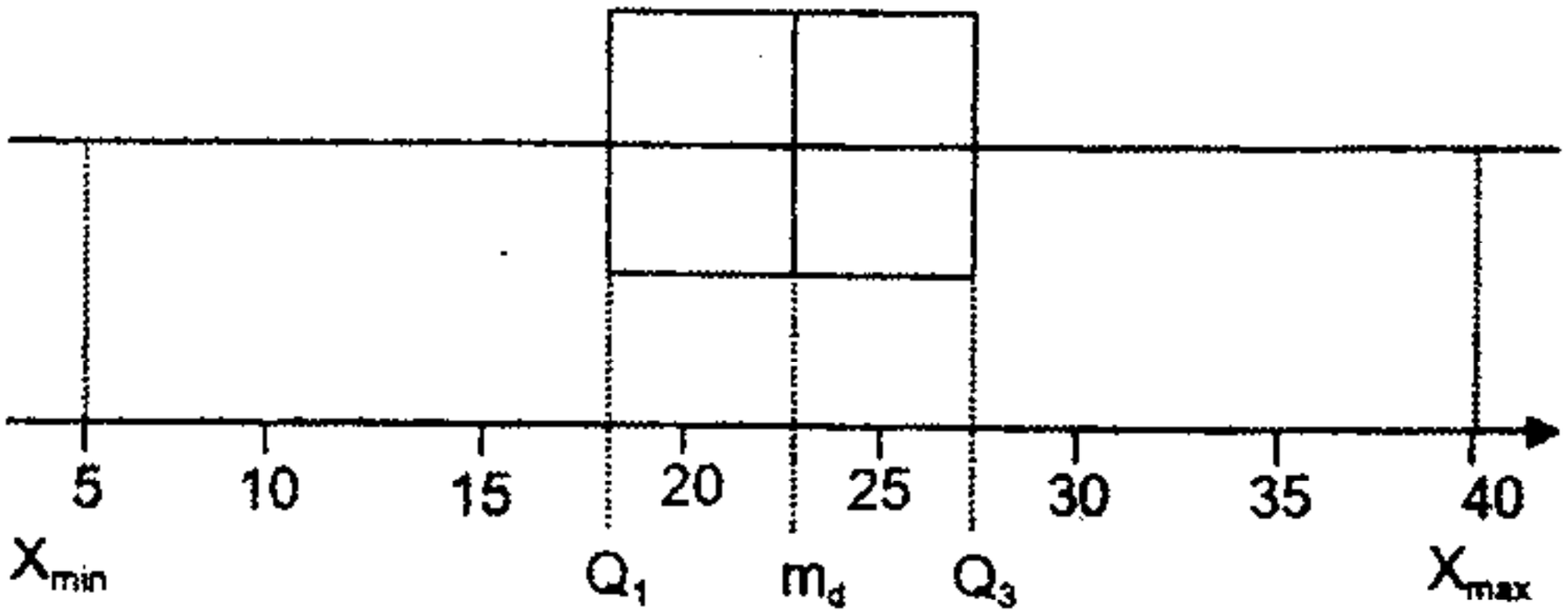
Từ các kết quả tính toán ở trên ta có:

$$m_d = 23,33$$

$$Q_1 = 19,6875 \ , \ Q_3 = 27,7083$$

$$X_{min} = 5 \ ; \ X_{max} = 40.$$

Vậy đồ thị hình hộp có dạng sau (hình 6.5).



Hình 6.5

Đặc biệt khi có nhiều mẫu rút ra từ các tổng thể nghiên cứu mà chúng lại có các thống kê đặc trưng khác nhau thì việc vẽ đồng thời đồ thị hình hộp của các mẫu đó lên cùng một mặt phẳng sẽ cho phép so sánh trực quan các mẫu, từ đó có được những nhận xét sơ bộ về sự khác biệt của các tổng thể nghiên cứu tương ứng.

Trong một số phần mềm thống kê chuyên dùng như SPSS, Stata... luôn có chương trình để vẽ đồ thị hình hộp một cách dễ dàng và nhanh chóng.

Thí dụ. Sau đây là các tham số đặc trưng chính của mẫu được tìm bằng phần mềm Stata theo các số liệu của thí dụ A.

sum x_1 x_2 x_3, detail

Thu nhập cá nhân (USD/năm) vùng 1

	Percentiles	Smallest		
1%	1550	1547		
5%	1577.5	1553		
10%	1595	1573	Obs	100
25%	1613.5	1573	Sum of Wgt.	100
50%	1643		Mean	1649.73
		Largest	Std. Dev.	47.3749
75%	1683	1747		
90%	1715	1747	Variance	2244.381
95%	1738.5	1750	Skewness	1886731
99%	1752.5	1755	Kurtosis	2.404846

Thu nhập cá nhân (USD/năm) vùng 2

	Percentiles	Smallest		
1%	1560	1557		
5%	1586	1563		
10%	1595.5	1576	Obs	100
25%	1624.5	1582	Sum of Wgt.	100
50%	1657.5		Mean	1655.99
		Largest	Std. Dev.	47.40828
75%	1687	1740		
90%	1727	1750	Variance	2247.545
95%	1739	1757	Skewness	249057
99%	1768.5	1780	Kurtosis	2.578889

Thu nhập cá nhân (USD/năm) vùng 3

	Percentiles	Smallest		
1%	1517.5	1510		
5%	1545.5	1525		
10%	1575.5	1527	Obs	100
25%	1593.5	1530	Sum of Wgt.	100
50%	1624		Mean	1628.12
		Largest	Std. Dev.	48.3997
75%	1657.5	1720		
90%	1695.5	1721	Variance	2342.531
95%	1712.5	1736	Skewness	.0724307
99%	1737	1738	Kurtosis	2.777361

Các đồ thị hình hộp thu được bằng phần mềm Stata được đặt trong cùng một mặt phẳng để tiện so sánh.

§5. MẪU NGẪU NHIÊN HAI CHIỀU

5.1. Khái niệm

Giả sử trên cùng một tổng thể phải nghiên cứu đồng thời hai dấu hiệu định tính hoặc định lượng, trong đó dấu hiệu nghiên cứu thứ nhất có thể xem như biến ngẫu nhiên X, còn dấu hiệu nghiên cứu thứ hai có thể xem như biến ngẫu nhiên Y. Lúc đó, việc nghiên cứu đồng thời hai dấu hiệu trong tổng thể tương tự như việc nghiên cứu một biến ngẫu nhiên hai chiều.

Từ tổng thể lấy ra một mẫu kích thước n, tức là thực hiện n phép thử đối với biến ngẫu nhiên (X, Y). Gọi X_i và Y_i tương ứng là giá trị của biến (X, Y) đo được trên phần tử thứ i của mẫu ($i = \overline{1,n}$), ta thu được n biến ngẫu nhiên hai chiều độc lập. Từ đó ta có định nghĩa sau.

Định nghĩa. *Mẫu ngẫu nhiên hai chiều kích thước n là tập hợp của n biến ngẫu nhiên độc lập (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) được thành lập từ biến ngẫu nhiên hai chiều (X, Y) và có cùng quy luật phân phối xác suất với (X, Y).*

Mẫu ngẫu nhiên hai chiều được ký hiệu là:

$$W = [(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)]$$

Việc thực hiện một phép thử đối với mẫu ngẫu nhiên W là thực hiện một phép thử đối với mỗi thành phần của mẫu.

Giả sử thành phần (X_i, Y_i) nhận giá trị (x_i, y_i) ($i = \overline{1,n}$) ta thu được một mẫu cụ thể là:

$$w = [(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)]$$

Các giá trị X_i ($i = \overline{1,n}$) được gọi là thành phần X_i của mẫu, còn các giá trị Y_i ($i = \overline{1,n}$) được gọi là thành phần Y_i của mẫu. (Cũng giống như mẫu ngẫu nhiên 1 chiều X_i và Y_i có thể được đo bằng những thang khác nhau tùy thuộc vào dấu hiệu nghiên cứu (X, Y) là định tính hay định lượng).

5.2. Phương pháp mô tả mẫu ngẫu nhiên hai chiều

Giả sử từ tổng thể rút ra một mẫu kích thước n, trong đó thành phần X nhận các giá trị $x_1, x_2, ..., x_i, ..., x_h$, còn thành phần Y nhận các giá trị $y_1, y_2, ..., y_j, ..., y_k$ trong đó giá trị (x_i, y_j) xuất hiện với tần số n_{ij} ($i = \overline{1,h}$; $j = \overline{1,k}$). Lúc đó, sau khi các giá trị x_i và y_j được sắp xếp theo thứ tự tăng dần thì giá trị cụ thể của mẫu w được mô tả bằng bảng phân phối tần số thực nghiệm sau:

X \ Y	y_1	y_2	...	y_j	...	y_k	n_i
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_1
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_2
⋮							
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_i
⋮							
x_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	n_h
m_j	m_1	m_2	...	m_j	...	m_k	$\Sigma = n$

trong đó n_i ký hiệu tổng các tần số của mẫu mang giá trị x_i của thành phần X, m_j ký hiệu tổng các tần số của mẫu mang giá trị y_j của thành phần Y.

Thí dụ 1. Bảng dưới đây chỉ kết quả thu hoạch Y (tạ/ha) và lượng phân bón X (kg/ha) của một loại hoa màu tại 100 thửa ruộng gieo trồng loại hoa màu đó (bảng 6.8).

Bảng 6.8

X \ Y	14	15	16	17	18	n_i
1	10					10
2	8	12				20
3		7	28			35
4			6	8		14
5				9	12	21
m_j	18	19	34	17	12	$\Sigma = 100$

Các điểm trống trên bảng thay cho tần số bằng 0.

Các số liệu của mẫu hai chiều cũng có thể mô tả thông qua đồ thị. Nếu biểu diễn các cặp giá trị (x_i, y_i) $(i = \overline{1,n})$ lên mặt phẳng thì ta sẽ thu được đồ thị rải điểm của mẫu, còn nếu nối các điểm trên bằng các đoạn thẳng thì thu được đồ thị liền nét của các giá trị của mẫu. Các đồ thị nói trên đều có thể vẽ bằng cách sử dụng các phần mềm thống kê thông thường.

5.3. Một số thống kê đặc trưng của mẫu ngẫu nhiên hai chiều

Từ bảng phân phối tần số thực nghiệm của mẫu ngẫu nhiên hai chiều ta có thể rút ra:

1. Bảng phân phối thực nghiệm của thành phần X

X	x_1	x_2	...	x_i	...	x_h
n_i	n_1	n_2	...	n_i	...	n_h

và tính được các thống kê đặc trưng của X:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{h} n_i X_i;\ S_X^2 = \frac{1}{n-1}\sum_{i=1}^{h} n_i (X_i - \overline{X})^2 \qquad (6.51)$$

2. Bảng phân phối thực nghiệm của thành phần Y

Y	Y_1	Y_2	...	Y_j	...	Y_k
m_j	m_1	m_2	...	m_j	...	m_k

và tính được các thống kê đặc trưng của Y:

$$\overline{Y} = \frac{1}{n}\sum_{j=1}^{k} m_j Y_j;\ S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{k} m_j (Y_j - \overline{Y})^2 \qquad (6.52)$$

3. Bảng phân phối có điều kiện của Y khi $X = X_i$

$Y_{X=Xi}$	Y_1	Y_2	...	Y_j	...	Y_k
n_{ij}	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}

trong đó $\sum_{j=1}^{k} n_{ij} = n_i$

Từ bảng phân phối có điều kiện của Y tính được trung bình có điều kiện của Y:

$$\overline{Y}_{X=X_i} = \frac{1}{n_i}\sum_{j=1}^{k} n_{ij} Y_j \qquad (6.53)$$

4. Tương tự ta có bảng phân phối có điều kiện của X khi $Y = Y_j$

$X_{Y=Yj}$	X_1	X_2	...	X_i	...	X_h
n_{ij}	n_{1j}	n_{2j}	...	n_{ij}	...	n_{hj}

trong đó $\sum_{i=1}^{h} n_{ij} = m_j$

và tính được trung bình có điều kiện của X

$$\overline{X}_{Y=Y_j} = \frac{1}{m_j}\sum_{i=1}^{h} n_{ij}X_i \qquad (6.54)$$

Thí dụ 2. Từ bảng (6.8) ta có bảng phân phối có điều kiện của Y khi X = 3.

$Y_{X=3}$	15	16
n_{ij}	7	28

Từ đó $\overline{Y}_{X=3} = \frac{15.7 + 16.28}{35} = 15{,}8$

Tương tự, ta có bảng phân phối có điều kiện của X khi Y = 15

$X_{Y=15}$	2	3
n_{ij}	12	7

Từ đó $\overline{X}_{Y=15} = \frac{2.12 + 3.7}{19} = 2{,}368p$

5. Hệ số tương quan mẫu, ký hiệu là r và được xác định bằng biểu thức:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{MS_X}\sqrt{MS_Y}} \qquad (6.55)$$

Đối với mẫu cụ thể mà các số liệu mẫu thường được trình bày dưới dạng bảng thì hệ số tương quan mẫu thường được tính bằng công thức sau:

$$r = \frac{n\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij}x_i y_j - \sum_{i=1}^{h} n_i x_i \cdot \sum_{j=1}^{k} m_j y_j}{\sqrt{n\sum_{i=1}^{h} n_i x_i^2 - \left(\sum_{i=1}^{h} n_i x_i\right)^2}\sqrt{n\sum_{j=1}^{k} m_j y_j^2 - \left(\sum_{j=1}^{k} m_j y_j\right)^2}}$$

Hệ số tương quan mẫu phản ánh mức độ kết hợp giữa X và Y. Các thống kê đặc trưng khác của mẫu hai chiều sẽ được đề cập tiếp ở các chương sau (xem chương X).

Trên đây là một số khái niệm liên quan đến mẫu ngẫu nhiên hai chiều. Khi nghiên cứu cùng một lúc n dấu hiệu trong tổng thể bằng phương pháp mẫu thì ta có khái niệm mẫu ngẫu nhiên n chiều và là sự mở rộng tương ứng các khái niệm đã xét ở trên.

§6. QUY LUẬT PHÂN PHỐI XÁC SUẤT CỦA MỘT SỐ THỐNG KÊ ĐẶC TRƯNG MẪU

Vì bản chất của các thống kê đặc trưng mẫu đã xét ở trên là các biến ngẫu nhiên, do đó để nắm được đầy đủ thông tin về các thống kê này cần khảo sát quy luật phân phối xác suất của chúng. Mặt khác, các quy luật phân phối xác suất này cũng phản ánh mối liên hệ mật thiết giữa các thống kê của mẫu với các tham số đặc trưng của dấu hiệu nghiên cứu trong tổng thể. Đó là căn cứ cho việc suy rộng các kết quả thu được trên mẫu cho toàn bộ tổng thể nghiên cứu.

Nói chung quy luật phân phối xác suất của các thống kê đặc trưng mẫu phụ thuộc chặt chẽ vào quy luật phân phối

xác suất của biến ngẫu nhiên gốc X. Để tiện cho việc sử dụng ở các chương tiếp theo trong một số trường hợp ta sẽ không khảo sát trực tiếp quy luật phân phối xác suất của các thống kê mẫu mà là quy luật phân phối xác suất của các thống kê là hàm của các thống kê nói trên. Mặt khác, ta cũng chỉ dừng lại ở việc thừa nhận các kết luận. Bạn đọc có thể tìm thấy chứng minh các kết luận này ở các tài liệu đầy đủ hơn về thống kê toán.

Sau đây ta xét một số thống kê thông dụng nhất.

6.1. Trường hợp biến ngẫu nhiên gốc phân phối theo quy luật chuẩn

Giả sử dấu hiệu nghiên cứu trong tổng thể có thể xem như một biến ngẫu nhiên tuân theo quy luật chuẩn với $E(X) = \mu$ và $V(X) = \sigma^2$. Các tham số này có thể đã biết hoặc chưa biết. Từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Lúc đó, để xác định quy luật phân phối xác suất của các thống kê đặc trưng mẫu có thể sử dụng định lý sau:

Nếu các biến ngẫu nhiên $X_1, X_2, ..., X_n$ độc lập và cùng phân phối theo quy luật chuẩn thì mọi tổ hợp tuyến tính của các biến ngẫu nhiên đó cũng phân phối theo quy luật chuẩn (xem mục 7 chương III).

Như vậy, theo định lý vừa trình bày thì các thành phần $X_1, X_2, ..., X_n$ của mẫu ngẫu nhiên thỏa mãn các điều kiện của định lý, do đó có thể sử dụng định lý để xét quy luật phân phối xác suất của các đặc trưng mẫu với tư cách như những tổ hợp tuyến tính của $X_1, X_2, ..., X_n$.

1. Thống kê trung bình mẫu $\overline{X}$ là một tổ hợp tuyến tính của $X_1, X_2, ..., X_n$, do đó nó cũng phân phối theo quy luật chuẩn với các tham số [xem (6.23) và (6.24) là:

$$E(\overline{X}) = \mu \text{ và } V(\overline{X}) = \frac{\sigma^2}{n}$$

Do đó, nếu xây dựng tiếp thống kê

$$G = U = \frac{\overline{X} - \mu}{Se(\overline{X})} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

thì (xem mục 7 chương III) thống kê U nói trên sẽ phân phối theo quy luật chuẩn hóa: N(0, 1).

2. Từ định nghĩa của phương sai S^{*2}

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

suy ra

$$nS^{*2} = \sum_{i=1}^{n}(X_i - \mu)^2$$

Đem chia cả hai vế cho σ^2 ta thu được thống kê sau:

$$G = \chi^2 = \frac{nS^{*2}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

Do đó, thống kê χ^2 (xem mục 8 chương III) sẽ phân phối theo quy luật "Khi bình phương" với n bậc tự do: $\chi^2(n)$.

3. Cũng trên cơ sở các thống kê mẫu nói trên, ta xây dựng thống kê

$$G = \chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

Có thể chứng minh được rằng thống kê χ^2 nói trên phân phối theo quy luật "Khi bình phương" với (n – 1) bậc tự do: $\chi^2(n-1)$.

4. Từ các kết quả thu được là:

$$U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \quad \text{phân phối N(0, 1) và}$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad \text{phân phối } \chi^2(n-1)$$

nên nếu ta xây dựng tiếp thống kê:

$$G = T = \frac{U}{\sqrt{\frac{\chi^2}{n-1}}} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} : \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{(\overline{X} - \mu)\sqrt{n}}{S}$$

thì thống kê T (xem mục 9 chương III) sẽ phân phối theo quy luật Student với (n – 1) bậc tự do: T(n – 1).

Chú ý rằng khi n khá lớn thì quy luật Student hội tụ khá nhanh về quy luật chuẩn hóa, do đó với n > 30 thực tế có thể xem thống kê T phân phối xấp xỉ N(0, 1).

6.2. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật chuẩn

Giả sử ta xét cùng một lúc hai tổng thể. Ở tổng thể thứ nhất dấu hiệu nghiên cứu được xem như biến ngẫu nhiên X_1 phân phối chuẩn với $E(X_1) = \mu_1$ và $V(X_1) = \sigma_1^2$; Ở tổng thể thứ hai dấu hiệu nghiên cứu được xem như biến ngẫu nhiên X_2 phân phối chuẩn với $E(X_2) = \mu_2$ và $V(X_2) = \sigma_2^2$.

Từ hai tổng thể nói trên rút ra hai mẫu ngẫu nhiên độc lập có kích thước tương ứng là n_1 và n_2:

$$W_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$$

1. Xét thống kê $(\overline{X}_1 - \overline{X}_2)$ là hiệu của hai trung bình mẫu. Đây là một tổ hợp tuyến tính của hai biến ngẫu nhiên độc lập cùng phân phối theo quy luật chuẩn, do đó bản thân nó cũng là biến ngẫu nhiên phân phối chuẩn với kỳ vọng toán là:

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2$$

và phương sai là:

$$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) - V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

do đó nếu xây dựng thống kê

$$G = U = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

thì U sẽ phân phối chuẩn hóa N(0, 1).

2. Mặt khác ta đã thấy các thống kê $\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2}$ và $\chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$ cùng phân phối theo quy luật khi bình phương với các bậc tự do tương ứng là $(n_1 - 1)$ và $(n_2 - 1)$, do đó theo tính chất cộng của χ^2 ta có thống kê

$$\chi^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

cũng sẽ phân phối theo quy luật khi bình phương với $(n_1 + n_2 - 2)$ bậc tự do.

Ở trên ta đã có thống kê

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

phân phối chuẩn hóa và độc lập với χ^2, vì vậy nếu $\sigma_1^2 = \sigma_2^2 = \sigma^2$ thì thống kê

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} : \frac{\sqrt{\frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}}}{\sqrt{n_1 + n_2 - 2}}$$

$$= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

sẽ phân phối theo quy luật Student với $(n_1 + n_2 - 2)$ bậc tự do và nếu $n_1 > 30$ và $n_2 > 30$ thì nó phân phối xấp xỉ chuẩn hóa N(0, 1).

Mặt khác nếu xét thống kê

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

thì người ta đã chứng minh được rằng nó phân phối Student với số bậc tự do là:

$$k = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)C^2 + (1 - C)^2(n_1 - 1)}$$

trong đó:
$$C = \frac{S_1^2/n_1}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

song nếu $n_1 > 30$ và $n_2 > 30$ thì có thể xem như T phân phối xấp xỉ chuẩn hóa N(0, 1).

3. Do

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ và } \chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

phân phối χ^2 $(n_1 - 1)$ và $\chi^2(n_2 - 1)$. Do đó nếu ta xây dựng thống kê

$$G = F = \frac{\frac{n_1 - 1}{\chi_1^2}}{\frac{\chi_2^2}{n_2 - 2}} = \frac{S_1^2.\sigma_2^2}{S_2^2.\sigma_1^2}$$

thì thống kê F (Xem mục 10 chương III) sẽ phân phối theo quy luật Fisher – Snedecor với $(n_1 - 1)$ và $(n_2 - 1)$ bậc tự do:

$$F(n_1 - 1, n_2 - 1).$$

6.3. Trường hợp biến ngẫu nhiên gốc X phân phối theo quy luật không - một

Giả sử trong tổng thể dấu hiệu nghiên cứu có thể xem như biến ngẫu nhiên X phân phối theo một quy luật nào đó khác với quy luật chuẩn. Lúc đó để cho đơn giản người ta thường yêu cầu có thể rút ra một mẫu ngẫu nhiên có kích thước n khá lớn

$$W = (X_1, X_2, ..., X_n)$$

Lúc đó, để xác định quy luật phân phối xác suất của các thống kê đặc trưng mẫu, theo định lý giới hạn Lindenberg – Lewi ta có:

$$U = \frac{(\overline{X} - m)\sqrt{n}}{\sigma} \text{ và } U = \frac{(\overline{X} - m)\sqrt{n}}{S}$$

sẽ phân phối xấp xỉ chuẩn hóa N(0, 1) khi n khá lớn.

Ta sẽ vận dụng định lý trên để xét một vài trường hợp thông dụng nhất trong thực tế.

1. Giả sử trong tổng thể dấu hiệu nghiên cứu có thể xem như biến ngẫu nhiên phân phối theo quy luật không – một. Từ tổng thể lập mẫu ngẫu nhiên kích thước n

$$W = (X_1, X_2, ..., X_n)$$

Ở mục 2 chương III ta đã biết rằng lúc đó tần suất mẫu f sẽ phân phối theo quy luật nhị thức với các tham số đặc trưng là:

$$E(f) = p$$

$$V(f) = \frac{p(1-p)}{n}$$

với p là xác suất (cơ cấu) của tổng thể.

2. Nếu kích thước mẫu n lớn mà p lại rất nhỏ và $np \approx np(1-p)$ thì tần suất mẫu sẽ phân phối xấp xỉ quy luật Poisson với tham số là p.

3. Nếu kích thước mẫu n lớn mà p lại không nhỏ song thỏa mãn điều kiện

$$n > 5 \text{ và } \frac{\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right|}{\sqrt{n}} < 0,3$$

thì tần suất mẫu sẽ phân phối xấp xỉ chuẩn với $E(f) = p$ và $V(f) = \frac{p(1-p)}{n}$. Do đó biến ngẫu nhiên

$$U = \frac{f-p}{Se(f)} = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}}$$

sẽ phân phối xấp xỉ chuẩn hóa N(0, 1).

6.4. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật không - một

Giả sử có hai tổng thể trong đó dấu hiệu nghiên cứu trong hai tổng thể có thể xem như các biến ngẫu nhiên phân phối không – một với các tham số tương ứng là p_1 và p_2. Từ hai tổng thể nói trên rút ra hai mẫu ngẫu nhiên độc lập kích thước tương ứng là n_1 và n_2.

$$W_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$$

Xét thống kê $f_1 - f_2$ là hiệu của hai tần suất mẫu. Lúc đó nếu $n_1 > 30$ và $n_2 > 30$ thì $f_1 - f_2$ sẽ phân phối xấp xỉ chuẩn theo định lý giới hạn trung tâm với các tham số đặc trưng là:

$$E(f_1 - f_2) = p_1 - p_2$$

và

$$V(f_1 - f_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

do đó thống kê

$$U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

sẽ phân phối xấp xỉ chuẩn hóa N(0, 1).

Trên đây là quy luật phân phối xác suất của một số thống kê thông dụng nhất. Còn quy luật phân phối xác suất của một số thống kê đặc biệt khác sẽ được đề cập trong từng trường hợp cụ thể.

§7. SUY DIỄN THỐNG KÊ

Như phần trên đã thấy, quy luật phân phối xác suất của các thống kê đặc trưng mẫu phản ánh mối liên hệ chặt chẽ giữa các tham số của mẫu với các tham số tương ứng của tổng thể nghiên cứu. Trong thực tế, các kết luận thu được ở trên về phân phối xác suất của các đặc trưng mẫu được sử dụng trong suy đoán thống kê theo hai cách thức trái ngược nhau:

- Suy diễn thống kê: Nếu đã biết quy luật phân phối xác suất cũng như các tham số đặc trưng của tổng thể thì có thể sử dụng các kết luận trên để suy đoán về tính chất của một mẫu ngẫu nhiên rút ra từ tổng thể đó. Nói cách khác đây chính là việc sử dụng thông tin đã biết của tổng thể để suy đoán về một bộ phận của tổng thể đó.

- Quy nạp thống kê: Nếu đã biết các tham số đặc trưng của mẫu thì căn cứ vào các kết luận trên để suy đoán về tính chất của tổng thể nghiên cứu mà từ đó mẫu được rút ra. Nói cách khác đây là việc sử dụng thông tin đã biết của một bộ phận của tổng thể để suy đoán về toàn bộ tổng thể đó.

Ở phần này ta sẽ nghiên cứu một số bài toán suy diễn thống kê quan trọng hơn cả trong thực tiễn. Còn các vấn đề quy nạp thống kê sẽ được xét ở các chương tiếp theo.

7.1. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối chuẩn

Giả sử trong tổng thể nghiên cứu biến ngẫu nhiên X phân phối chuẩn với kỳ vọng toán μ và phương sai σ^2 đã biết. Lúc đó nếu từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n thì có thể căn cứ vào thông tin đã biết về tổng thể để suy đoán về mẫu như sau:

1. Suy đoán về giá trị của trung bình mẫu

Ta có thống kê

$$U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0,1)$$

Vì vậy với xác suất $1 - \alpha$ có thể tìm được các giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tìm được các giá trị tới hạn $u_{1-\alpha_1}$ và u_{α_2} tương ứng sao cho thỏa mãn điều kiện:

$$p[u_{1-\alpha_1} < U < u_{\alpha_2}] = 1 - \alpha$$

Từ đó

$$P\left[u_{1-\alpha_1} < \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} < u_{\alpha_2}\right] = 1 - \alpha$$

Sau khi chuyển vế ta thu được:

$$P\left[\mu - \frac{\alpha}{\sqrt{n}} u_{\alpha_1} < \overline{X} < \mu + \frac{\alpha}{\sqrt{n}} u_{\alpha_2}\right] = 1 - \alpha \qquad (6.56)$$

Với những cặp giá trị α_1 và α_2 khác nhau ta thu được khoảng giá trị khác nhau của $\overline{X}$.

Thí dụ 1. Một công ty sản xuất mỳ ăn liền với trọng lượng đóng gói là biến ngẫu nhiên phân phối chuẩn có trọng

lượng đóng gói trung bình là 340 gam và độ lệch chuẩn là 10 gam. Lấy ngẫu nhiên 16 gói mỳ để kiểm tra.

a. Tìm xác suất để trọng lượng trung bình của các gói mỳ đó nằm trong khoảng từ 335 gam đến 345 gam.

b. Tìm xác suất để trọng lượng trung bình của các gói mỳ đó lớn hơn 340 gam.

c. Với xác suất 0,99 thì trọng lượng trung bình của các gói mỳ đó nằm trong khoảng nào xung quanh giá trị trung bình.

Giải. Gọi X là trọng lượng mỳ đóng gói. Theo giả thiết X phân phối chuẩn với $\mu = 340$ và $\sigma = 10$. Lấy ngẫu nhiên một mẫu $n = 16$.

a. Ta phải tìm xác suất để trung bình mẫu $\overline{X}$ nằm trong khoảng (335; 345). Theo công thức (6.56) ta có:

$$\mu - \frac{\sigma}{\sqrt{n}} u_{\alpha_1} = 340 - \frac{10}{\sqrt{15}} u_{\alpha_1} = 335$$

Từ đó $u_{\alpha_1} = 2$

và
$$\mu + \frac{\sigma}{\sqrt{n}} u_{\alpha_2} = 340 + \frac{10}{\sqrt{16}} u_{\alpha_2} = 345$$

Từ đó $u_{\alpha_2} = 2$

Tra bảng giá trị tới hạn u_α ta có $u_{\alpha_1} = u_{\alpha_2} = 2 = u_{0,0228}$ vậy $\alpha_1 = \alpha_2 = 0,0228$ hay $\alpha = \alpha_1 + \alpha_2 = 0,0456$.

Từ đó $1 - \alpha = 1 - 0,0456 = 0,9544$.

b. Để tìm xác suất sao cho trung bình mẫu lớn hơn 340 ta lấy $\alpha_2 = 0$; $\alpha_1 = \alpha$ từ đó $u_{\alpha_2} = +\infty$ và công thức (6.56) trở thành

$$P\left[\overline{X} > \mu - \frac{\sigma}{\sqrt{n}} u_\alpha\right] = 1 - \alpha$$

Từ đó

$$\mu - \frac{\sigma}{\sqrt{n}} u_\alpha = 340 - \frac{10}{\sqrt{16}} u_\alpha = 340$$

Từ đó

$u_\alpha = 0 = u_{0,5}$. Vậy $\alpha = 0,5 \Rightarrow 1 - \alpha = 0,5$

c. Ta phải tìm giá trị ε sao cho

$$P[a < \overline{X} < b] = 0,99 = 1 - \alpha$$

vậy $\alpha = 0,01 \Rightarrow \alpha_1 = \alpha_2 = \frac{\alpha}{2} = 0,005$

$\Rightarrow u_{0,005} = 2,58$

Do đó

$$a = \mu - \frac{\sigma}{\sqrt{n}} U_{\alpha_1} = 340 - \frac{10}{\sqrt{16}}.2,58 = 333,55 \text{ (gam)}$$

và $$b = \mu + \frac{\sigma}{\sqrt{n}} U_{\alpha_2} = 340 + \frac{10}{\sqrt{16}}.2,58 = 346,45 \text{ (gam)}$$

Vậy với xác suất 0,99 trọng lượng trung bình của các gói mỳ được đem kiểm tra sẽ nằm trong khoảng (333,55 ; 346,45) gam.

2. Suy đoán về giá trị của phương sai mẫu

Ta có thống kê

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Vì vậy với xác suất $1-\alpha$ có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tìm được các giá trị tới hạn $\chi^{2(n-1)}_{1-\alpha_1}$ và $\chi^{2(n-1)}_{\alpha_2}$ tương ứng sao cho thỏa mãn điều kiện

$$P\left[\chi^{2(n-1)}_{1-\alpha_1} < \chi^2 < \chi^{2(n-1)}_{\alpha_2}\right] = 1-\alpha$$

Thay giá trị của χ^2 vào biểu thức và sau khi chuyển vế ta thu được

$$P\left[\frac{\sigma^2}{n-1}\chi^{2(n-1)}_{1-\alpha_1} < S^2 < \frac{\sigma^2}{n-1}\chi^{2(n-1)}_{\alpha_2}\right] = 1-\alpha \qquad (6.57)$$

với mỗi cặp giá trị α_1 và α_2 ta sẽ có một khoảng giá trị tương ứng của S^2.

Thí dụ 2. Xí nghiệp sử dụng một loại nguyên liệu với lượng tạp chất là biến ngẫu nhiên phân phối chuẩn với phương sai là 20 $(gam)^2$ trong 1kg nguyên liệu. Từ một lô nguyên liệu mới nhập về người ta lấy ngẫu nhiên ra 16kg. Tìm xác suất để độ phân tán (phương sai) của lượng tạp chất trong mẫu hàng nằm trong khoảng [9,68 ; 33,33] $(gam)^2$.

Giải. Gọi X là lượng tạp chất trong nguyên liệu thì X phân phối chuẩn với $\sigma^2 = 20$. Nếu lấy một mẫu ngẫu nhiên $n = 16$ thì theo (6.57) ta có:

$$\frac{\sigma^2}{n-1}\chi^{2(n-1)}_{1-\alpha_1} = \frac{20}{15}\chi^{2(15)}_{1-\alpha_1} = 9,68$$

$$\Rightarrow \chi^{2(15)}_{1-\alpha_1} = 7,26 = \chi^{2(15)}_{0,95}$$

$$\Rightarrow 1-\alpha_1 = 0,95 \Rightarrow \alpha_1 = 0,05$$

và $$\frac{\sigma^2}{n-1}\chi^{2(n-1)}_{\alpha_2} = 24{,}9975 = \chi^{2(15)}_{0{,}05}$$

$$\Rightarrow \alpha_2 = 0{,}05$$

Từ đó $\alpha = \alpha_1 + \alpha_2 = 0{,}1 \Rightarrow 1 - \alpha = 0{,}9$.

7.2. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối không - một

Giả sử biến ngẫu nhiên X trong tổng thể phân phối không - một với tham số là p. Lúc đó nếu từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n thì có thể suy đoán về số lần xuất hiện biến cố trong mẫu đó hoặc tần suất mẫu.

Ta có thống kê

$$U = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim \text{xấp xỉ } N(0, 1)$$

nếu thỏa mãn điều kiện

$$n > 5 \text{ và } \left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \Big/ \sqrt{n} < 0{,}3$$

Lúc đó với xác suất $1 - \alpha$ có thể tìm được cặp giá trị α_1 và α_2 ($\alpha_1 + \alpha_2 = \alpha$) và hai giá trị tới hạn $u_{1-\alpha_1}$ và u_{α_2} tương ứng thỏa mãn điều kiện.

$$P\left[u_{1-\alpha_1} < U < u_{\alpha_2}\right] = 1 - \alpha$$

Thay biểu thức của U vào và chuyển vế ta thu được công thức:

$$P\left[p - \frac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\alpha_1} < f < p + \frac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\alpha_2}\right] = 1 - \alpha \qquad (6.58)$$

Từ đó tiến hành các suy đoán tương ứng đối với f.

Thí dụ 3. Một lô hàng đủ tiêu chuẩn xuất khẩu nếu tỷ lệ sản phẩm loại một của lô hàng đó là 90%. Từ một lô hàng dự định xuất khẩu người ta lấy ngẫu nhiên ra 100 sản phẩm thì trong đó phải có ít nhất bao nhiêu sản phẩm loại một thì đủ tiêu chuẩn xuất khẩu. Hãy kết luận với xác suất 0,9.

Giải. Gọi p là tỷ lệ sản phẩm loại một của lô hàng đó, $p = 0,9$. Nếu lấy ra một mẫu sản phẩm $n = 100$ và gọi X là số sản phẩm loại một trong đó thì ta có tần suất mẫu

$$f = \frac{X}{n} = \frac{X}{100}$$

Do $n = 100 > 5$ và $\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \Big/ \sqrt{n}$

$$= \left|\sqrt{\frac{0,9}{0,1}} - \sqrt{\frac{0,1}{0,9}}\right| \Big/ \sqrt{100} = 0,267 < 0,3$$

nên có thể dùng công thức (6.58). Trong biểu thức này nếu lấy $\alpha_2 = 0$ thì ta có:

$$P\left[f > p - \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_\alpha\right] = 1 - \alpha = 0,9$$

từ đó $\alpha = 0,1$ và $u_\alpha = 1,282$

vậy $$f > 0,9 - \frac{\sqrt{0,9.0,1}}{\sqrt{100}} 1,282 = 0,86154$$

Từ đó $X > 100.0,86154 = 86,154$.

Song vì X phải là số nguyên nên $X \geq 87$ sản phẩm. Vậy trong mẫu sản phẩm đó phải có ít nhất 87 sản phẩm loại một thì lô hàng mới có thể xuất khẩu được.

Chú ý rằng việc suy diễn bằng công thức (6.58) chỉ tiến hành được nếu mẫu đủ lớn để quy luật nhị thức có thể xấp xỉ chuẩn. Nếu điều kiện đó không được thỏa mãn thì phải sử dụng trực tiếp quy luật nhị thức hoặc quy luật Poisson.

Trên đây là một số bài toán suy diễn thống kê đơn giản nhất. Việc xây dựng công thức suy diễn đối với hiệu hai trung bình mẫu, hiệu hai tần suất mẫu và tỷ số của hai phương sai mẫu của hai mẫu rút ra từ hai tổng thể nghiên cứu cũng được thực hiện tương tự bằng cách sử dụng các kết luận đã thu được ở §6.

Các ký hiệu và công thức cơ bản

* Tổng thể nghiên cứu kích thước N

* Dấu hiệu nghiên cứu χ được mô hình hóa bằng biến ngẫu nhiên X.

* Tần suất tổng thể

$p_i = \frac{N_i}{N}$ trong đó N_i – số phần tử của tổng thể mang giá trị x_i

* Tần suất tích lũy của tổng thể

$$F(x_i) = \frac{W_i}{N} = \sum_{x_j < x_i} \frac{N_j}{N}$$

* Các tham số cơ bản của tổng thể

+ Trung bình tổng thể (trung bình số học)

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i$$

+ Trung bình điều hòa

$$m_h = \frac{N}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

+ Trung bình nhân

$$m_g = \sqrt[N]{x_1.x_2...x_N}$$

+ Phương sai tổng thể

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - m^2$$

* Mẫu ngẫu nhiên kích thước n: $W = (X_1, X_2, ..., X_n)$

* Mẫu cụ thể $w = (x_1, x_2, ..., x_n)$

* Tần suất mẫu

$f_i = \frac{n_i}{n}$ trong đó n_i là số phần tử của mẫu mang giá trị x_i

* Tần suất tích lũy của mẫu

$$F^*(x_i) = \frac{w_i}{n} = \sum_{x_j<x_i}\frac{n_j}{n}$$

* Các thống kê đặc trưng mẫu

+ Trung bình mẫu

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

- Kỳ vọng toán của trung bình mẫu

$$E(\overline{X}) = m$$

- Sai số chuẩn của trung bình mẫu

$$Se(\overline{X}) = \sqrt{V(\overline{X})} = \frac{\sigma}{\sqrt{n}}$$

+ Trung vị mẫu (số liệu ghép lớp)

$$X_d \approx L + \frac{(n/2 - S)}{n_{x_d}} h$$

+ Mốt (số liệu ghép lớp)

$$X_0 \approx L + \left(\frac{d_1}{d_1 + d_2}\right) h$$

+ Khoảng biến thiên

$$R = x_{max} - x_{min}$$

+ Khoảng tứ phân vị

$$IQR = Q_3 - Q_1$$

trong đó:

$$Q_1 = L_{Q_1} + \frac{n/4 - S_{Q_1}}{n_{Q_1}} h_{Q_1}$$

$$Q_3 = L_{Q_3} + \frac{3n/4 - S_{Q_3}}{n_{Q_3}} h_{Q_3}$$

+ Độ lệch bình phương trung bình

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

+ Phương sai mẫu

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

+ Phương sai S^{*2}

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n} (X_i - m)^2$$

+ Độ lệch chuẩn mẫu S

+ Tần suất mẫu (với số liệu định danh)

$f = \frac{X}{n}$ trong đó X – số phần tử mang dấu hiệu nghiên cứu

• Kỳ vọng toán của tần suất mẫu $E(f) = p$

• Sai số chuẩn của tần suất mẫu

$$Se(f) = \sqrt{V(f)} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

* Mẫu ngẫu nhiên hai chiều

$$W = [(X_1, Y_1)(X_2, Y_2) \dots (X_n, Y_n)]$$

* Hệ số tương quan mẫu

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{MS_X}\sqrt{MS_Y}}$$

* Phân phối xác suất của các thống kê đặc trưng mẫu

+ $U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0, 1)$

+ $T = \frac{(\overline{X} - \mu)\sqrt{n}}{S} \sim T(n-1)$

+ $\chi^2 = \frac{nS^{*2}}{\sigma^2} \sim \chi^2(n)$

+ $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

+ $U = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0, 1)$

+ $U = \frac{(f-p)\sqrt{n}}{\sqrt{f(1-f)}} \sim N(0, 1)$ nếu $n \geq 100$

+ $U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$.

+ $T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2)$

nếu $\sigma_1^2 = \sigma_2^2$

+ $T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T(k)$

với $k = \frac{(n_1-1)(n_2-1)}{(n_2-1)C^2 + (1-C)^2(n_1-1)}$; $C = \frac{S_1^2/n_1}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

+ $F = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1)$

+ $$U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1)$$

* Công thức suy diễn về trung bình mẫu

$$P\left(\mu - \frac{\sigma}{\sqrt{n}} u_{\alpha_1} < \overline{X} < \mu + \frac{\sigma}{\sqrt{n}} u_{\alpha_2}\right) = 1 - \alpha$$

* Công thức suy diễn về phương sai mẫu

$$P\left(\frac{\sigma^2}{n-1} \chi^{2(n-1)}_{1-\alpha_1} < S^2 < \frac{\sigma^2}{n-1} \chi^{2(n-1)}_{\alpha_2}\right) = 1 - \alpha$$

* Công thức suy diễn về tần suất mẫu

$$P\left(p - \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{\alpha_1} < f < p + \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{\alpha_2}\right) = 1 - \alpha$$

Câu hỏi ôn tập

1. Hãy phân biệt sự khác nhau giữa tổng thể và mẫu, giữa các tham số và các thống kê. Cho ví dụ.

2. Hãy phân biệt sự khác nhau giữa dấu hiệu nghiên cứu định lượng và định tính. Cho ví dụ. Các dấu hiệu định lượng và định tính có thể mô hình hóa bằng cách nào?

3. Thống kê mô tả và thống kê suy diễn khác nhau như thế nào? Trong nghiên cứu thống kê thì loại thống kê nào nằm ở dạng thức cao hơn? Vì sao?

4. Một giám đốc xí nghiệp muốn tìm hiểu xem trong một năm bình quân mỗi công nhân của xí nghiệp nghỉ việc bao nhiêu ngày. Song do xí nghiệp có tới hơn 2000 công nhân nên ông ta không thể đủ thời gian để xem sổ chấm công của từng người. Trong trường hợp này giám đốc có thể làm như thế nào để có được thông tin cần thiết?

5. Cho thí dụ về một mẫu ngẫu nhiên kích thước n = 10 và một mẫu cụ thể của mẫu ngẫu nhiên ấy.

6. Cho ví dụ về một mẫu ngẫu nhiên kích thước n được xây dựng từ biến ngẫu nhiên tuân theo quy luật không - một, quy luật nhị thức, quy luật chuẩn.

7. Khi nào thì mẫu lặp (có hoàn lại) và mẫu không lặp (không hoàn lại) có thể coi là như nhau?

8. Một mẫu ngẫu nhiên phải được chọn theo những nguyên tắc nào? Cho ví dụ.

9. Hãy liệt kê các đặc trưng về xu hướng trung tâm, độ phân tán và dạng phân phối của tổng thể và của mẫu.

10. Tại sao sau khi đã thu thập được số liệu về tổng thể hoặc mẫu lại phải sắp xếp chúng lại và mô tả theo một hình thức nào đó? Hãy phân tích ưu điểm của các hình thức mô tả sau:

a - Bảng phân phối tần số

b - Bảng phân phối tần số tích lũy

c - Bảng phân phối tần suất

d - Đồ thị hình cột

e - Đồ thị hình bánh xe

f - Đường đa giác

11. Với những điều kiện nào thì dùng trung bình điều hòa và trung bình nhân tốt hơn trung bình số học.

12. Với những điều kiện nào thì dùng trung vị tốt hơn trung bình để đặc trưng cho xu hướng trung tâm.

13. Đồ thị hình hộp chứa đựng những thông tin nào về dãy số liệu mà nó mô tả?

14. Nếu rút ra một mẫu kích thước n từ tổng thể thì kích thước mẫu sẽ ảnh hưởng như thế nào đến độ phân tán của trung bình mẫu? Độ phân tán của trung bình mẫu sẽ thay đổi như thế nào khi n tăng lên?

15. Tiến sĩ A được thuyên chuyển từ vụ thống kê sang vụ kinh tế của một Bộ. Người phụ trách vụ thống kê cho rằng như vậy chỉ số IQ của vụ thống kê sẽ giảm đi còn chỉ số IQ của vụ kinh tế sẽ tăng lên. Với những điều kiện nào thì lời nhận xét trên là đúng?

Chương VII

ƯỚC LƯỢNG CÁC THAM SỐ CỦA BIẾN NGẪU NHIÊN

Giả sử cần phải nghiên cứu dấu hiệu χ trong tổng thể. Như đã phân tích ở chương VI, một trong những mục tiêu cơ bản của việc nghiên cứu là xác định các tham số đặc trưng của tổng thể như trung bình, phương sai, cơ cấu của tổng thể theo dấu hiệu nghiên cứu. Đó là những chỉ tiêu tổng hợp để phân tích tổng thể cần nghiên cứu.

Nếu dấu hiệu nghiên cứu trong tổng thể có thể xem như một biến ngẫu nhiên X và giả sử bằng phân tích lý thuyết đã xác định được dạng phân phối xác suất của nó thì vấn đề xác định các tham số đặc trưng của tổng thể sẽ được quy về bài toán xác định các tham số đặc trưng của quy luật phân phối xác suất xác định biến ngẫu nhiên gốc X. Chẳng hạn, nếu đã biết được rằng dấu hiệu nghiên cứu trong tổng thể có thể xem như biến ngẫu nhiên phân phối theo quy luật chuẩn thì bài toán đặt ra là phải ước lượng (tức là xác định một cách gần đúng) các tham số kỳ vọng toán μ và phương sai σ^2 của nó vì các tham số trên hoàn toàn xác định quy luật phân phối chuẩn và thực chất chúng chính là trung bình và phương sai của tổng thể.

Như vậy, bài toán ước lượng tham số có thể phát biểu như sau: Cho biến ngẫu nhiên X với quy luật phân phối xác suất đã biết song chưa biết tham số θ nào đó của nó. Phải ước lượng (xác định một cách gần đúng) giá trị θ.

Phương pháp mẫu cho phép giải quyết bài toán trên bằng quy nạp thống kê như sau: Từ tổng thể nghiên cứu rút ra một mẫu ngẫu nhiên kích thước n và dựa vào đó mà xây dựng một thống kê $\hat{\theta}$ dùng để ước lượng θ bằng cách này hay cách khác. Có hai phương pháp sử dụng $\hat{\theta}$ để ước lượng θ là phương pháp ước lượng điểm và phương pháp ước lượng bằng khoảng tin cậy. Sau đây ta sẽ xem xét những nội dung chính của hai phương pháp đó.

§1. PHƯƠNG PHÁP ƯỚC LƯỢNG ĐIỂM

Phương pháp ước lượng điểm chủ trương dùng một giá trị để thay thế cho tham số θ chưa biết của tổng thể, vì bản thân $\hat{\theta}$ là một số xác định. Thông thường giá trị được chọn là một thống kê $\hat{\theta}$ nào đó của mẫu ngẫu nhiên. Có nhiều cách chọn thống kê $\hat{\theta}$ khác nhau tạo nên những phương pháp ước lượng điểm khác nhau.

1.1. Phương pháp hàm ước lượng (phương pháp mômen)

1. Khái niệm: Giả sử cần ước lượng tham số θ của biến ngẫu nhiên gốc X. Từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Chọn lập thống kê $\hat{\theta} = f(X_1, X_2, ..., X_n)$ mà thực chất là một thống kê đặc trưng mẫu tương ứng với tham số θ cần ước lượng. Chẳng hạn, để ước lượng kỳ vọng toán m của biến ngẫu nhiên gốc thì chọn thống kê trung bình mẫu $\overline{X}$, để ước lượng phương sai σ^2 của biến ngẫu nhiên gốc thì chọn thống kê phương sai mẫu S^2 ... Nếu lập một mẫu cụ thể và tính được giá trị $\hat{\theta} = f(x_1, x_2, ..., x_n)$ của thống kê $\hat{\theta}$ trên mẫu cụ thể đó thì ước lượng của θ là giá trị $\hat{\theta}$ vừa tính.

Vì thống kê $\hat{\theta} = f(X_1, X_2, ..., X_n)$ thực chất là hàm của các biến ngẫu nhiên nên nó được gọi là *hàm ước lượng* của θ.

Chất lượng của ước lượng không thể đánh giá qua một giá trị cụ thể của $\hat{\theta}$. Vì như vậy chỉ có cách so sánh trực tiếp $\hat{\theta}$ và θ mà θ lại chưa biết. Do đó chỉ có thể đánh giá chất lượng của ước lượng thông qua bản thân thống kê $\hat{\theta} = f(X_1, X_2, ..., X_n)$. Rõ ràng là có vô số cách chọn hàm f, tức là có vô số thống kê $\hat{\theta}$ có thể dùng làm ước lượng của θ. Vì vậy, cần đưa ra các tiêu chuẩn để đánh giá chất lượng của các thống kê $\hat{\theta}$, từ đó lựa chọn được thống kê "xấp xỉ một cách tốt nhất" tham số cần ước lượng. Dưới đây ta xét các tiêu chuẩn đó.

2. Các tiêu chuẩn lựa chọn hàm ước lượng

a. Ước lượng không chệch

Giả sử thống kê $\hat{\theta}$ là ước lượng của tham số θ của biến ngẫu nhiên gốc. Với k mẫu cụ thể rút ra từ tổng thể, thống kê $\hat{\theta}$ sẽ nhận k giá trị cụ thể tương ứng là $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$. Nếu thống kê $\hat{\theta}$ là một ước lượng có dư của θ thì mọi giá trị $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ cũng đều sẽ lớn hơn θ và giá trị trung bình của

chúng (tức là kỳ vọng toán của $\hat{\theta}$) cũng sẽ lớn hơn θ: $E(\hat{\theta}) > \theta$. Ngược lại, nếu θ là một ước lượng thiếu của θ thì mọi giá trị của $\hat{\theta}_i (i = \overline{1,k})$ cũng đều sẽ nhỏ hơn θ nên kỳ vọng toán của $\hat{\theta}$ cũng sẽ nhỏ hơn θ: $E(\theta) < \theta$.

Như vậy, việc sử dụng một thống kê mà kỳ vọng toán của nó khác với tham số cần ước lượng có thể dẫn đến các sai số hệ thống (tất cả các giá trị của $\hat{\theta}$ đều lớn hơn hoặc nhỏ hơn θ). Dĩ nhiên yêu cầu trên không loại trừ được hoàn toàn các sai số, song như vậy các sai số khác dấu sẽ xuất hiện tương đối đều nhau, do đó các giá trị của $\hat{\theta}$ sẽ không bị lệch hẳn về một phía so với θ. Từ đó ta có định nghĩa sau:

Định nghĩa. Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng không chệch của tham số θ của biến ngẫu nhiên gốc X nếu:

$$E(\hat{\theta}) = \theta$$

Ngược lại, nếu $E(\hat{\theta}) \neq \theta$ thì $\hat{\theta}$ được gọi là *ước lượng chệch* của θ.

Rõ ràng từ phân tích ở trên ta nên dùng loại ước lượng không chệch.

Chú ý rằng $\hat{\theta}$ là ước lượng không chệch của θ không có nghĩa là mọi giá trị của $\hat{\theta}$ đều trùng khít với θ mà chỉ có nghĩa: Trung bình các giá trị của $\hat{\theta}$ bằng θ. Từng giá trị của $\hat{\theta}$ có thể sai lệch rất lớn so với θ.

Dựa vào các kết quả đã chứng minh được ở mục §4 chương VI có thể rút ra một số kết luận sau:

- Trung bình mẫu $\overline{X}$ là ước lượng không chệch của kỳ vọng toán m của biến ngẫu nhiên gốc $[E(\overline{X}) = m]$.

- Tần suất mẫu f là ước lượng không chệch của xác suất p của biến ngẫu nhiên gốc $[E(f) = p]$.

- Phương sai mẫu S^2 và phương sai S^{*2} đều là các ước lượng không chệch của phương sai σ^2 của biến ngẫu nhiên gốc:

$$E(S^2) = \sigma^2 \text{ và } E(S^{*2}) = \sigma^2 \qquad (7.1)$$

Nếu $\tilde{\theta}$ là một ước lượng chệch của θ song thoả mãn điều kiện

$$\lim_{n\to\infty} E(\tilde{\theta}) = 0 \qquad (7.2)$$

thì $\tilde{\theta}$ được gọi là ước lượng tiệm cận không chệch của θ. Độ chệch của một ước lượng được đo bằng giá trị:

$$BS = |E(\tilde{\theta}) - \theta| \qquad (7.3)$$

Hiển nhiên là trong số các ước lượng chệch thì nên chọn ước lượng nào có BS nhỏ nhất.

b. Ước lượng hiệu quả

Như đã phân tích ở trên, dù $\hat{\theta}$ là ước lượng không chệch của θ thì từng giá trị cụ thể của $\hat{\theta}$ vẫn có thể sai lệch rất lớn so với θ, tức là phương sai $V(\theta)$ vẫn có thể lớn. Lúc đó, nếu lấy một giá trị của $\hat{\theta}$ tìm được trên một mẫu cụ thể, chẳng hạn $\hat{\theta}_1$ để ước lượng θ thì nó có thể sai lệch rất nhiều so với giá trị trung bình $\hat{\theta}$ tức là bản thân tham số θ cần ước lượng. Như vậy, nếu lấy một giá trị của $\hat{\theta}$, chẳng hạn $\hat{\theta}_1$ để ước lượng θ, thì có thể mắc sai số rất lớn. Còn nếu như đòi hỏi phương sai của $\hat{\theta}$ phải nhỏ thì có thể hạn chế được loại sai số này trong ước lượng. Từ đó ta có định nghĩa sau:

Định nghĩa: Thống kê của mẫu được gọi là ước lượng hiệu quả nhất của tham số θ của biến ngẫu nhiên gốc X nếu nó là ước lượng không chệch và có phương sai nhỏ nhất so với mọi ước lượng không chệch khác được xây dựng trên cùng mẫu đó.

Như vậy, để xét xem $\hat{\theta}$ có phải là ước lượng hiệu quả nhất của θ hay không, cần phải tìm được giá trị nhỏ nhất có thể có của phương sai các hàm ước lượng. Nếu $\hat{\theta}$ đã là một ước lượng không chệch của θ thì trong nhiều trường hợp giá trị nhỏ nhất của phương sai $V(\hat{\theta})$ có thể tìm được dựa vào *bất đẳng thức Cramer - Rao* được phát biểu như sau:

Cho mẫu ngẫu nhiên $W = (X_1, X_2, \ldots X_n$ được xây dựng từ biến ngẫu nhiên gốc X *có hàm mật độ xác suất (hay biểu thức xác suất)* $f(x, \theta)$ thỏa mãn *một số điều kiện nhất định* (thường được thỏa mãn trong thực tế, ít ra là các phân phối xác suất đã xét ở chương III) và θ^* là một ước lượng không chệch bất kỳ của θ thì

$$V(\theta^*) \geq \frac{1}{nE\left[\frac{\partial \ln f(x,\theta)}{\partial \theta}\right]^2} \qquad (7.4)$$

Chẳng hạn, dựa vào bất đẳng thức trên ta có thể chứng minh được rằng trung bình mẫu $\overline{X}$ là ước lượng hiệu quả nhất của kỳ vọng toán μ của biến ngẫu nhiên gốc X trong tổng thể khi X phân phối chuẩn $N(\mu,\sigma^2)$.

Thật vậy, ở mục §4 chương VI ta đã chứng minh được $V(\overline{X}) = \frac{\sigma^2}{n}$. Mặt khác, do $\overline{X}$ là ước lượng không chệch của μ và do X phân phối $N(\mu,\sigma^2)$ nên ta có:

$$f(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\ln f(x, \theta) = -\ln \sqrt{2\pi}\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \ln f(x,\theta)}{\partial \theta} = \frac{\partial \ln f(x,\mu)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

Vậy $$nE\left[\frac{\partial \ln f(x,\theta)}{\partial \theta}\right]^2 = nE\left(\frac{x-\mu}{\sigma^2}\right)^2 = \frac{n}{\sigma^4}E(x-\mu)^2$$

$$= \frac{n}{\sigma^4}V(X) = \frac{n\sigma^2}{\sigma^4} = \frac{n}{\sigma^2}$$

Như vậy, $V(\overline{X})$ bằng biểu thức ở vế phải của bất đẳng thức Cramer-Rao. Vậy $\overline{X}$ là ước lượng hiệu quả nhất của μ.

Ta cũng có thể chứng minh được rằng tần suất mẫu f là ước lượng hiệu quả nhất của xác suất P của biến ngẫu nhiên gốc X phân phối theo quy luật A(p). Thật vậy, ở mục §4 chương VI ta đã có $V(f) = \frac{p(1-p)}{n}$. Mặt khác, do f là ước lượng không chệch của p và do X phân phối A(p) nên:

$$f(x, \theta) = P(X = x) = p^x (1-p)^{1-x}$$

$$\ln f(x,\theta) = x\ln p + (1-x)\ln(1-p)$$

Lấy đạo hàm riêng của biểu thức trên theo p ta được

$$\frac{\partial \ln f(x,\theta)}{\partial \theta} = \frac{x}{p} - \frac{1-x}{1-p}$$

do đó $$\left(\frac{x}{p} - \frac{1-x}{1-p}\right)^2 = \left[\frac{x-p}{p(1-p)}\right]^2$$

và $$nE\left[\frac{\partial \ln f(x,\theta)}{\partial \theta}\right]^2 = nE\left[\frac{x-p}{p(1-p)}\right]^2 = \frac{n}{p^2(1-p)^2}E(x-p)^2 =$$

$$= \frac{n}{p^2(1-p)^2}V(X) = \frac{n}{p^2(1-p)^2}p(1-p) = \frac{n}{p(1-p)} \quad .$$

Vậy V(f) bằng biểu thức ở vế phải của bất đẳng thức Cramer - Rao hay f là ước lượng hiệu quả nhất của p.

Khi hai ước lượng $\hat{\theta}_1$ và $\hat{\theta}_2$ nào đó đều là các ước lượng không chệch của θ song không phải là ước lượng hiệu quả nhất thì có thể so sánh phương sai của hai ước lượng đó để tìm ra ước lượng hiệu quả hơn. Giả sử $V(\hat{\theta}_1) < V(\hat{\theta}_2)$, lúc đó độ hiệu quả của $\hat{\theta}_1$ so với $\hat{\theta}_2$ được xác định bằng biểu thức:

$$\dot{E}F = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)} \tag{7.5}$$

Thí dụ. Từ một mẫu ngẫu nhiên kích thước n = 2 ta xét hai ước lượng sau đây của trung bình tổng thể m

$$\overline{X} = \frac{1}{2}(X_1 + X_2)$$

và
$$X' = \frac{1}{3}X_1 + \frac{2}{3}X_2$$

a. Xét xem $\overline{X}$ và X' có phải là ước lượng không chệch của m hay không?

b. Ước lượng nào hiệu quả hơn.

Giải:

a. Ta có

$$E(\overline{X}) = E\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) = \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) = \frac{1}{2}m + \frac{1}{2}m = m$$

$$E(X') = E\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) = \frac{1}{3}E(X_1) + \frac{2}{3}E(X_2) = \frac{1}{3}m + \frac{2}{3}m = m$$

Vậy cả $\overline{X}$ và X' đều là các ước lượng không chệch của m.

b. Ta có

$$V(\overline{X}) = V\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) = \frac{1}{4}V(X_1) + \frac{1}{4}V(X_2) = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{\sigma^2}{2}$$

$$V(X') = V\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) = \frac{1}{9}V(X_1) + \frac{4}{9}V(X_2) = \frac{1}{9}\sigma^2 + \frac{4}{9}\sigma^2 = \frac{5}{9}\sigma^2$$

$V(\overline{X}) < V(X')$. Vậy $\overline{X}$ hiệu quả hơn

$\Rightarrow$ EF của $\overline{X}$ so với X' là:

$$EF = \frac{(5/9)\sigma^2}{(1/2)\sigma^2} = \frac{10}{9} = 1{,}11 = 111\%$$

Trong thực tế đôi khi người ta còn xây dựng một tổ hợp tuyến tính của hai ước lượng không chệch $\hat{\theta}_1$ và $\hat{\theta}_2$ dạng $\hat{\theta} = \alpha\hat{\theta}_1 + (1 - \alpha)\hat{\theta}_2$ nhằm thu được một ước lượng hiệu quả hơn cho θ.

c. Ước lượng vững: Khi xét những mẫu có kích thước lớn thì nảy sinh vấn đề là mẫu càng lớn thì thống kê $\hat{\theta}$ của mẫu phải càng gần tham số θ cần ước lượng. Từ đó ta có định nghĩa:

Định nghĩa. *Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng vững của tham số θ của biến ngẫu nhiên gốc X nếu $\hat{\theta}$ hội tụ theo xác suất đến θ khi $n \to \infty$.*

Tức là với mọi ε dương bé tùy ý ta luôn có:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1$$

Như vậy theo luật số lớn của Trêbusép (trường hợp riêng) và luật số lớn của Bernoulli (xem chương V) ta suy ngay ra là trung bình mẫu $\overline{X}$ là ước lượng vững của tham số m của biến ngẫu nhiên gốc X và tần suất mẫu f là ước lượng vững của xác suất p của biến ngẫu nhiên gốc X.

Trong trường hợp $\hat{\theta}$ là ước lượng không chệch của θ thì để tìm ước lượng vững có thể sử dụng định lý sau: *Nếu* $\hat{\theta}$ là ước lượng không chệch của θ và $\lim_{n\to\infty} V(\hat{\theta}) = 0$ thì $\hat{\theta}$ là ước lượng vững của θ.

d. Ước lượng đủ: Một ước lượng $\hat{\theta}$ *được gọi là ước lượng đủ nếu nó chứa đựng toàn bộ các thông tin trong mẫu về tham số* θ *của ước lượng.*

Chẳng hạn, trung bình mẫu và trung vị mẫu đều là các ước lượng không chệch của trung bình tổng thể song trung bình mẫu là ước lượng đủ của trung bình tổng thể vì nó sử dụng toàn bộ thông tin của mẫu, còn trung vị mẫu không phải là ước lượng đủ vì nó chỉ dùng đến giá trị chính giữa của dãy số liệu mẫu mà thôi.

Hiển nhiên là trong ước lượng ta sẽ chủ trương sử dụng các ước lượng đủ nếu điều đó cho phép.

Mặt khác, trong thực tế $\hat{\theta}$ có thể là ước lượng tuyến tính hoặc phi tuyến của các giá trị của mẫu. Nếu $\hat{\theta}$ là ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất thì nó được gọi là ước lượng tuyến tính không chệch tốt nhất (BLUE) của θ. Còn nếu $\hat{\theta}$ là ước lượng phi tuyến song vẫn thỏa mãn điều kiện không chệch và có phương sai nhỏ nhất thì được gọi là ước lượng không chệch tốt nhất (BUE).

3. Một vài kết luận của phương pháp hàm ước lượng

Dùng những tiêu chuẩn nêu trên để đánh giá các thống kê đặc trưng mẫu khác nhau cho phép ta lựa chọn được những thống kê tốt nhất, tức là ước lượng một cách chính xác nhất các tham số đặc trưng của tổng thể. Có thể đưa ra một số kết luận chung sau đây:

- Vì trung bình mẫu $\overline{X}$ là ước lượng không chệch, hiệu quả nhất và vững của trung bình tổng thể m và đồng thời là ước lượng tuyến tính không chệch tốt nhất, do đó nếu chưa biết m có thể dùng $\overline{X}$ để ước lượng nó.

- Vì tần suất mẫu f là ước lượng không chệch, hiệu quả nhất và vững của tần suất tổng thể p và đồng thời là ước lượng tuyến tính không chệch tốt nhất do đó nếu chưa biết p có thể dùng f để ước lượng nó.

- Vì phương sai mẫu S^2 và phương sai S^{*2} đều là các ước lượng không chệch của phương sai tổng thể σ^2 do đó nếu chưa biết phương sai σ^2 có thể dùng S^2 hoặc S^{*2} để ước lượng nó.

Chú ý rằng MS và phương sai mẫu S^2 chỉ khác nhau chút ít bởi hệ số $\frac{n-1}{n}$. Khi n lớn thì sự sai khác này là không đáng kể. Trong thực tế phương sai mẫu S^2 được sử dụng khi $n < 30$.

1.2. Phương pháp ước lượng hợp lý tối đa

Giả sử đã biết quy luật phân phối xác suất tổng quát của biến ngẫu nhiên gốc X dưới dạng hàm mật độ $f(x,\theta)$. Đó cũng có thể là biểu thức xác suất nếu X là biến ngẫu nhiên rời rạc. Cần phải ước lượng tham số θ nào đó của X. Lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

và xây dựng hàm của đối số θ tại một giá trị cụ thể của mẫu:

$$L(x_1, x_2, ..., x_n, \theta) = f(x_1,\theta).f(x_2,\theta)...f(x_n,\theta).$$

Hàm L được gọi là *hàm hợp lý* của tham số θ. Giá trị của

hàm hợp lý chính là xác suất hay mật độ xác suất tại điểm $(x_1, x_2,..., x_n)$, còn giá trị của thống kê θ tại điểm đó

$$\hat{\theta} = f(x_1, ..., x_n)$$

được gọi là ước lượng hợp lý tối đa của θ nếu ứng với giá trị này của θ, hàm hợp lý đạt cực đại.

Do hàm L và lnL đạt cực đại tại cùng một giá trị của θ nên có thể tìm giá trị của θ để lnL đạt cực đại với các bước sau:

a. Tìm đạo hàm bậc nhất của lnL theo θ.

b. Giải phương trình $\frac{d \ln L}{d\theta} = 0$

Giả sử nó có nghiệm $\theta = \hat{\theta} = f(x_1, x_2, ..., x_n)$

c. Tìm đạo hàm bậc hai $\frac{d^2 \ln L}{d\theta^2}$

Nếu tại điểm $\theta = \hat{\theta}$ đạo hàm bậc hai âm thì tại điểm này hàm lnL đạt cực đại, do đó $\hat{\theta} = f(x_1, x_2, ..., x_n)$ là ước lượng điểm hợp lý tối đa cần tìm của θ. Chú ý rằng đối số của hàm hợp lý là θ chứ không phải là $(x_1, x_2, ..., x_n)$ do đó nếu thay giá trị của mẫu bằng bản thân mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ thì kết quả thu được vẫn đúng. Từ đó sẽ thu được kết quả tổng quát hơn, tức là tìm được hàm ước lượng hợp lý tối đa

Thí dụ 1. Tìm ước lượng hợp lý tối đa của tham số p trong quy luật phân phối A(p).

Biểu thức xác suất tổng quát của quy luật không - một là

$$P_x = p^x(1 - p)^{1-x}$$

Lập mẫu ngẫu nhiên kích thước n: $W = (X_1, X_2, ..., X_n)$ và tại giá trị $w = (x_1, x_2, ..., x_n)$ của nó lập hàm hợp lý với mỗi giá trị x_i ($x_i = 0$ hoặc 1)

$$L(x_1, x_2, ..., x_n, p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

Từ đó

$$\ln L = \sum_{i=1}^{n} [x_i \ln p + (1-x_i) \ln(1-p)]$$

$$\frac{d \ln L}{dp} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (1-x_i)$$

$$\frac{d \ln L}{dp} = 0 \Rightarrow p = \bar{x}$$

Dễ thấy rằng $\frac{d^2 \ln L}{dp^2} = -\frac{n}{p(1-p)} < 0$ do đó ước lượng hợp lý tối đa của p là $\bar{x}$.

Thí dụ 2. Tìm ước lượng hợp lý tối đa của các tham số μ và σ^2 của biến ngẫu nhiên X tuân theo quy luật chuẩn.

Hàm hợp lý có dạng

$$L(x_1, x_2, ..., x_n, \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-1/2\sigma^2 \sum_{i=1}^{n} (x_i - \mu)^2}$$

Tìm các đạo hàm riêng của lnL theo μ và σ^2 và cho bằng không ta có:

$$\begin{cases} \frac{\partial \ln L}{\partial \mu} = \frac{n\bar{X} - n\mu}{\sigma^2} = 0 \\ \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^3} = 0 \end{cases}$$

Giải hệ phương trình trên thu được $\mu = \bar{X}$ và $\sigma^2 = MS$. Vậy ước lượng hợp lý tối đa của μ là $\bar{X}$, còn của σ^2 là MS. Vậy ước lượng hợp lý tối đa của σ^2 là ước lượng chệch.

§2. PHƯƠNG PHÁP ƯỚC LƯỢNG BẰNG KHOẢNG TIN CẬY

2.1. Khái niệm

Phương pháp ước lượng điểm nói trên có một nhược điểm cơ bản là khi kích thước mẫu nhỏ thì ước lượng điểm tìm được có thể sai lệch rất nhiều so với giá trị của tham số cần ước lượng, tức là sai số của ước lượng có thể rất lớn. Mặt khác dùng các phương pháp trên không thể đánh giá được khả năng mắc sai lầm khi ước lượng bằng bao nhiêu. Do đó khi kích thước mẫu nhỏ người ta thường sử dụng phương pháp *ước lượng bằng khoảng tin cậy*.

Để ước lượng tham số θ của biến ngẫu nhiên gốc X trong tổng thể, phương pháp này chủ trương từ một thống kê G nào đó của mẫu xây dựng một khoảng giá trị (G_1, G_2) sao cho với một xác suất cho trước tham số θ sẽ rơi vào khoảng (G_1, G_2) đó. Chú ý rằng do G là biến ngẫu nhiên nên khoảng (G_1, G_2) cũng là một khoảng ngẫu nhiên, còn θ lại là một số xác định nên phải nói chính xác hơn là khoảng (G_1, G_2) sẽ chứa đựng giá trị θ với một xác suất cho trước. Từ đó ta có định nghĩa sau:

Định nghĩa. Khoảng (G_1, G_2) của thống kê G được gọi là khoảng tin cậy của tham số θ nếu với xác suất bằng $(1 - \alpha)$ cho trước thỏa mãn điều kiện

$$P(G_1 < \theta < G_2) = 1 - \alpha \qquad (7.6)$$

xác suất $(1 - \alpha)$ được gọi là độ tin cậy của ước lượng, còn $I = (G_2 - G_1)$ được gọi là độ dài khoảng tin cậy.

Như vậy, vấn đề chủ yếu của phương pháp ước lượng bằng khoảng tin cậy là làm thế nào để xác định được khoảng tin cậy (G_1, G_2) thỏa mãn (7.6). Để làm điều đó người ta tiến hành như sau:

Từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

và từ đó xây dựng thống kê $G = f(X_1, X_2, ..., X_n, \theta)$ sao cho quy luật phân phối xác suất của G không phụ thuộc vào các đối số của nó và hoàn toàn xác định. Lúc đó với độ tin cậy bằng $(1 - \alpha)$ cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tương ứng với chúng tìm được cặp giá trị $g\alpha_1$ và $g\alpha_2$ thỏa mãn điều kiện

$$P(G < g\alpha_1) = \alpha_1 \qquad (7.7)$$

và

$$P(G > g\alpha_2) = \alpha_2 \qquad (7.8)$$

Từ (7.7) và (7.8) suy ra

$$P(g\alpha_1 < G < g\alpha_2) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha \qquad (7.9)$$

Như vậy, với độ tin cậy $1 - \alpha$ ta xây dựng được khoảng tin cậy $(g\alpha_1, g\alpha_2)$ cho G. Bằng các phép biến đổi tương đương bao giờ cũng có thể đưa (7.9) về dạng biểu thức tương đương

$$P(G_1 < \theta < G_1) = 1 - \alpha$$

Đó chính là khoảng tin cậy cần tìm.

Trong thực tế thường yêu cầu độ tin cậy $(1 - \alpha)$ khá lớn (chẳng hạn $1 - \alpha = 0,95$) nên theo nguyên lý xác suất lớn biến cố $(G_1 < \theta < G_2)$ hầu như chắc chắn sẽ xảy ra trong một phép thử. Tiến hành một phép thử với mẫu ngẫu nhiên $W = (X_1,$

$X_2, ..., X_n$) ta thu được một mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ do đó tính được giá trị của G_1 và G_2 ứng với mẫu cụ thể này, ký hiệu là g_1 và g_2. Lúc đó có thể kết luận là: Qua mẫu cụ thể với độ tin cậy $(1 - \alpha)$ tham số θ của biến ngẫu nhiên gốc X sẽ nằm trong khoảng (g_1, g_2) tức là $(g_1 < \theta < g_2)$.

Phương pháp ước lượng bằng khoảng tin cậy khắc phục được các nhược điểm của phương pháp ước lượng điểm. Chẳng những nó làm tăng độ chính xác của ước lượng mà còn đánh giá được mức độ tin cậy của ước lượng đó nữa. Tuy nhiên nó cũng chứa đựng khả năng mắc sai lầm bằng α.

Dưới đây sẽ áp dụng phương pháp này để ước lượng một vài tham số đặc trưng cơ bản của biến ngẫu nhiên X trong tổng thể.

2.2. Ước lượng kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật chuẩn

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối chuẩn $N(\mu, \sigma^2)$ nhưng chưa biết tham số μ của nó. Để ước lượng μ từ tổng thể ta lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n).$$

Để chọn thống kê G thích hợp ta xét hai trường hợp sau:

1. Đã biết phương sai σ^2 của biến ngẫu nhiên gốc X trong tổng thể

Lúc đó ta chọn thống kê

$$G = U = \frac{\overline{X} - \mu}{Se(\overline{X})} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \qquad (7.10)$$

trong đó $\overline{X}$ là trung bình mẫu. Từ mục §6 Chương VI ta đã

biết thống kê U phân phối chuẩn hóa N (0,1). Do đó với độ tin cậy $(1 - \alpha)$ cho trước tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ từ đó tìm được hai giá trị tới hạn tương ứng của phân phối chuẩn hóa là $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn các điều kiện

$$P(U < u_{1-\alpha_1}) = \alpha_1$$

và $$P(U > u_{\alpha_2}) = \alpha_2$$

Từ đó $P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha$

Vì u_{α_1} có tính chất $u_{\alpha_1} = - u_{1-\alpha_1}$ nên biểu thức trên có thể viết

$$P(-u_{\alpha_1} < U < u_{\alpha_2}) = 1 - \alpha \qquad (7.11)$$

Thay biểu thức của U từ (7.10) vào (7.11) và giải ra μ, ta thu được biểu thức tương đương:

$$P(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1}) = 1 - \alpha \qquad (7.12)$$

Ý nghĩa của biểu thức thu được là: Với độ tin cậy bằng $(1 - \alpha)$ tham số μ của biến ngẫu nhiên gốc X sẽ nằm trong khoảng

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2}; \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1}) \qquad (7.13)$$

Biểu thức (7.13) mới chỉ cho ta một khoảng tin cậy tổng quát. Với độ tin cậy $(1 - \alpha)$ cho trước hiển nhiên có thể tìm được vô số cặp giá trị α_1 và α_2 thỏa mãn điều kiện $\alpha_1 + \alpha_2 = \alpha$, từ đó sẽ có vô số khoảng tin cậy tương ứng. Trong thực tế từ biểu thức (7.13) người ta thường chỉ sử dụng một số trường hợp đặc biệt sau:

- Khoảng tin cậy đối xứng:

Nếu lấy $\alpha_1 = \alpha_2 = \alpha/2$ từ (7.13) suy ra khoảng tin cậy của μ là

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha/2}\right) \qquad (7.14)$$

Nếu ký hiệu $\varepsilon = \frac{\sigma}{\sqrt{n}} u_{\alpha/2}$ (7.15)

thì biểu thức của khoảng tin cậy sẽ có dạng $(\overline{X} - \varepsilon; \overline{X} + \varepsilon)$

ε được gọi là *độ chính xác* của ước lượng. Nó phản ánh mức độ sai lệch của trung bình mẫu so với trung bình tổng thể với xác suất $(1-\alpha)$ cho trước.

- Khoảng tin cậy bên phải:

Nếu lấy $\alpha_1 = 0$ và $\alpha_2 = \alpha$ thì $u_{\alpha_1} = u_0 = +\infty$ do đó khoảng tin cậy của μ là

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha}; +\infty) \qquad (7.16)$$

Biểu thức (7.16) được dùng để ước lượng giá trị tối thiểu của μ.

- Khoảng tin cậy bên trái:

Nếu lấy $\alpha_2 = 0$ và $\alpha_1 = \alpha$ thì $u_{\alpha_2} = u_o = +\infty$ do đó khoảng tin cậy của μ là

$$(-\infty; \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha}) \qquad (7.17)$$

Biểu thức (7.17) được dùng để ước lượng giá trị tối đa của μ.

Với cùng độ tin cậy (1- α) hiển nhiên khoảng tin cậy nào ngắn hơn sẽ tốt hơn. Trong trường hợp này độ dài khoảng tin cậy I sẽ là ngắn nhất khi khoảng tin cậy là đối xứng. Lúc đó độ dài khoảng tin cậy sẽ bằng hai lần độ chính xác và được xác định bằng biểu thức

$$I = 2\varepsilon = \frac{2\sigma}{\sqrt{n}} u_{\alpha/2} \qquad (7.18)$$

Từ (7.18) ta sẽ thu được công thức xác định kích thước mẫu tối thiểu n sao cho với độ tin cậy bằng (1 – α) cho trước, độ dài khoảng tin cậy không vượt quá giá trị I_o cho trước. Công thức có dạng

$$n \geq \left[\frac{4\sigma^2}{I_0^2} u_{\alpha/2}^2\right] = \left[\frac{\sigma^2}{\varepsilon_0^2} u_{\alpha/2}^2\right] \qquad (7.19)$$

tức là n là số nguyên dương nhỏ nhất lớn hơn hoặc bằng

$$\frac{\sigma^2}{\varepsilon_0^2} u_{\alpha/2}^2$$

Bài toán xác định kích thước mẫu tối thiểu n theo công thức (7.19) thường được đặt ra trước khi chọn mẫu, khi phải xác định kích thước mẫu cần điều tra để đáp ứng những yêu cầu chất lượng cho trước về độ tin cậy và độ chính xác của ước lượng.

Các khoảng tin cậy (7.14) (7.16) và (7.17) vẫn đang còn là khoảng tin cậy ngẫu nhiên. Vì vậy thủ tục ước lượng chưa thể coi là kết thúc. Vì độ tin cậy (1 – α) khá lớn nên áp dụng nguyên lý xác suất lớn có thể coi biến cố

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1}\right)$$

sẽ xảy ra trong một phép thử đối với mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$. Thực hiện một phép thử đối với mẫu này thu được mẫu cụ thể $w = (x_1, x_2, ..., x_n)$; từ đó tìm được giá trị cụ thể $\bar{x}$ của trung bình mẫu. Lúc đó với độ tin cậy $(1 - \alpha)$, qua một mẫu cụ thể, khoảng tin cậy của μ là:

$$(g_1; g_2) = (\bar{x} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2}; \bar{x} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1})$$

Bằng cách tiến hành tương tự ta cũng thu được khoảng tin cậy cụ thể của các công thức (7.14) (7.16) (7.17) cũng như độ dài cụ thể của khoảng tin cậy theo công thức (7.18).

Đến đây ta thấy rõ vai trò của từng loại mẫu. Mẫu ngẫu nhiên cho phép xác định khoảng tin cậy ngẫu nhiên, còn mẫu cụ thể cho phép tìm ra khoảng tin cậy cụ thể (bằng số) của μ.

Thí dụ 1. Trọng lượng một loại sản phẩm là biến ngẫu nhiên phân phối theo quy luật chuẩn với độ lệch chuẩn là 1 gam. Cân thử 25 sản phẩm loại này ta thu được kết quả sau.

Trọng lượng (gam)	18	19	20	21
Số SP tương ứng	3	5	15	2

Với độ tin cậy 0,95 hãy tìm khoảng tin cậy đối xứng của trọng lượng trung bình của loại sản phẩm nói trên.

Giải. Gọi X là "trọng lượng sản phẩm", theo giả thiết X phân phối chuẩn với $\sigma = 1$. Vậy trọng lượng trung bình của sản phẩm chính là tham số μ. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng giá trị của tham số μ của phân phối $N(\mu, \sigma^2)$ khi đã biết phương sai của nó. Vậy theo biểu thức (7.14) ta có khoảng tin cậy là

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2}; \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha/2})$$

Lấy từ tổng thể ra một mẫu ngẫu nhiên kích thước n = 25, gọi X_i là trọng lượng của sản phẩm thứ i ($i = \overline{1,25}$) ta có:

$$W = (X_1, X_2, ..., X_{25})$$

Từ đó: $$\overline{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$$

Với độ tin cậy $1 - \alpha = 0,95$ thì $\alpha/2 = 0,025$

Tra bảng giá trị tới hạn chuẩn có $u_{0,025} = 1,96$

Vậy khoảng tin cậy đối xứng của μ là:

$$(\overline{X} - \frac{1}{\sqrt{25}} 1,96; \overline{X} + \frac{1}{\sqrt{25}} 1,96) = (\overline{X} - 0,392; \overline{X} + 0,392)$$

Kết quả thu được cho biết 95% số mẫu kích thước n = 25 sẽ chứa đựng tham số μ trong khoảng $(\overline{X} - 0,392; \overline{X} + 0,392)$

Từ bảng số liệu tìm được trung bình mẫu cụ thể:

$$\overline{x} = \frac{3.18 + 5.19 + 15.20 + 2.21}{25} = 19,64$$

Vậy với độ tin cậy 0,95 qua mẫu cụ thể này, khoảng tin cậy đối xứng của μ là:

$$(19,64 - 0,392; 19,64 + 0,392)$$

hay $$(19,248 < \mu < 20,032)$$

Chú ý rằng không thể viết $P(19,248 < \mu < 20,032) = 0,95$ vì độ tin cậy gắn với khoảng tin cậy ngẫu nhiên chứ không gắn với mẫu cụ thể. Hơn nữa do μ là một hằng số nên nó chỉ có thể thuộc hoặc không thuộc khoảng $(19,248 < \mu < 20,032)$

tức là với một mẫu cụ thể thì biến cố ($19,248 < \mu < 20,032$) không phải là biến cố ngẫu nhiên. Hoặc nó là biến cố chắc chắn, hoặc nó là biến cố không thể có:

Chú ý: Từ biểu thức (7.18) ta có thể đưa ra các nhận xét sau:

- Khi tăng kích thước mẫu n lên và giữ nguyên độ tin cậy $1 - \alpha$ cho trước thì ε giảm đi tức là độ chính xác của ước lượng tăng lên.

- Khi tăng độ tin cậy $1 - \alpha$ lên mà giữ nguyên kích thước mẫu n thì giá trị tới hạn chuẩn cũng tăng lên theo do đó ε cũng tăng lên làm cho độ chính xác của ước lượng giảm đi.

Thí dụ 2: Trong thí dụ 1, nếu yêu cầu độ chính xác của ước lượng chỉ là 0,1 và giữ nguyên độ tin cậy $1 - \alpha = 0,95$ thì phải điều tra một mẫu kích thước bằng bao nhiêu?

Giải: Với $\varepsilon_0 = 0,1$ theo công thức (7.19) ta có:

$$n \geq \left[\frac{1^2}{(0,1)^2} (1,96)^2 \right] = 385$$

Vậy để làm tăng độ chính xác của ước lượng từ 0,392 lên đến 0,1 thì kích thước mẫu phải tăng từ 25 lên đến 385.

Nếu mẫu được chon theo phương thức không hoàn lại từ tổng thể hữu hạn và $n > 0,1N$ thì $Se(\overline{X})$ sẽ nhân thêm hệ số điều chỉnh, do đó biểu thức (7.10) của thống kê G sẽ trở thành:

$$G = U = \frac{(\overline{X} - \mu)}{Se(\overline{X})} = \frac{\overline{X} - \mu}{\sqrt{\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}}}$$

và việc xây dựng các công thức ước lượng cũng tiến hành tương tự.

2. Chưa biết phương sai σ^2 của biến ngẫu nhiên gốc X trong tổng thể và kích thước mẫu n < 30

Lúc đó ta chọn thống kê:

$$G = T = \frac{(\overline{X} - \mu)\sqrt{n}}{S} \qquad (7.20)$$

trong đó S là độ lệch chuẩn mẫu. Từ mục §6 Chương VI ta đã biết thống kê T phân phối theo quy luật Student với (n – 1) bậc tự do. Vì vậy, với độ tin cậy bằng (1 – α) cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$, từ đó tìm được hai giá trị tới hạn Student tương ứng là $t_{1-\alpha_1}^{(n-1)}$ và $t_{\alpha_2}^{(n-1)}$ thỏa mãn điều kiện:

$$P(T < t_{1-\alpha_1}^{(n-1)}) = \alpha_1 \text{ và}$$

$$P(T > t_{\alpha_2}^{(n-1)}) = \alpha_2$$

Từ đó $$P(t_{1-\alpha_1}^{(n-1)} < T < t_{\alpha_2}^{(n-1)}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha$$

Vì giá trị tới hạn Student có tính chất $t_{\alpha_1}^{(n-1)} = -t_{1-\alpha_1}^{(n-1)}$ nên biểu thức có thể viết:

$$P(-t_{\alpha_1}^{(n-1)} < T < t_{\alpha_2}^{(n-1)}) = 1 - \alpha \qquad (7.21)$$

Thay biểu thức của T từ (7.20) vào (7.21) và giải ra μ ta thu được biểu thức tương đương:

$$P(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha_2}^{(n-1)} < \mu < \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha_1}^{(n-1)}) = 1 - \alpha \qquad (7.22)$$

Như vậy, khoảng tin cậy của μ với độ tin cậy (1 – α) là:

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha_2}^{(n-1)}; \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha_1}^{(n-1)}) \quad (7.23)$$

Từ biểu thức tổng quát (7.23) có thể xây dựng công thức khoảng tin cậy trong những trường hợp đặc biệt sau:

- Khoảng tin cậy đối xứng khi $\alpha_1 = \alpha_2 = \alpha/2$

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}; \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}) \quad (7.24)$$

- Khoảng tin cậy bên phải khi $\alpha_1 = 0; \alpha_2 = \alpha$

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha}^{(n-1)}; +\infty) \quad (7.25)$$

- Khoảng tin cậy bên trái, khi $\alpha_2 = 0; \alpha_1 = \alpha$

$$(-\infty; \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha}^{(n-1)}) \quad (7.26)$$

Trong trường hợp này độ dài khoảng tin cậy I cũng là ngắn nhất khi khoảng tin cậy là đối xứng, do đó nó cũng bằng hai lần độ chính xác và được xác định bằng biểu thức:

$$I = 2\varepsilon = \frac{2S}{\sqrt{n}} t_{\alpha/2}^{(n-1)} \quad (7.27)$$

Lúc đó việc xác định kích thước mẫu tối thiểu n sao cho với độ tin cậy bằng $(1 - \alpha)$ cho trước, độ dài khoảng tin cậy không vượt quá giá trị I_o cho trước được giải quyết bằng *phương pháp mẫu kép* sau đây:

Trước hết điều tra một mẫu sơ bộ kích thước $m \geq 2$: $W_1 = (X_1, ..., X_m)$ và từ đó tìm được phương sai mẫu của mẫu sơ bộ đó:

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

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

Sau đó lập mẫu thứ hai kích thước n – m

$$W_2 = (X_{m+1}, \ldots, X_n)$$

Có thể chứng minh được rằng thống kê:

$$T = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)\sqrt{n}}{S}$$

phân phối theo quy luật Student với (m – 1) bậc tự do. Vì vậy, có thể tìm giá trị $t_{\alpha/2}^{(m-1)}$ sao cho:

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{S}{\sqrt{n}} t_{\alpha/2}^{(m-1)} < \mu < \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{S}{\sqrt{n}} t_{\alpha/2}^{(m-1)}\right] = 1 - \alpha$$

do đó $I_0 = \frac{2S}{\sqrt{n}} t_{\alpha/2}^{(m-1)}$

Từ đây suy ra kích thước mẫu cần tìm:

$$n \geq \left[\frac{S^2}{\varepsilon_0^2} t_{\alpha/2}^{2(m-1)}\right] \qquad (7.28)$$

Như vậy, dựa vào mẫu sơ bộ đã có ta tìm được kích thước mẫu chính thức đáp ứng yêu cầu về chất lượng của ước lượng. Trên thực tế chỉ cần điều tra tiếp mẫu thứ hai kích thước n – m là đủ.

Với độ tin cậy (1 – α) khá lớn, để có khoảng tin cậy cụ thể của μ, người ta lập mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$, từ đó tính

được các giá trị $\overline{x}$ và s, thay chúng vào các công thức khoảng tin cậy vừa tìm được ở trên ta có các khoảng tin cậy cụ thể bằng số phải tìm.

Thí dụ 3. Để xác định trọng lượng trung bình của các bao bột trong kho, người ta đem cân ngẫu nhiên 15 bao của kho đó và tìm được $\overline{x} = 39{,}8$ kg; $s^2 = 0{,}144$. Hãy tìm khoảng tin cậy đối xứng của trọng lượng trung bình của các bao bột trong kho với yêu cầu độ tin cậy của việc ước lượng là 99%. Giả thiết trọng lượng đóng bao của các bao bột là biến ngẫu nhiên phân phối chuẩn.

Giải. Gọi X là trọng lượng bột đóng bao, theo giả thiết X phân phối chuẩn. Vậy trọng lượng đóng bao trung bình chính là giá trị μ. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng giá trị của tham số μ của phân phối $N(\mu, \sigma^2)$ khi chưa biết σ^2 của X. Vậy theo biểu thức (7.27) ta có khoảng tin cậy là:

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}; \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)})$$

Cân ngẫu nhiên 15 bao bột, gọi X_i ($i = \overline{1,15}$) là trọng lượng của bao thứ i ta có mẫu ngẫu nhiên $W = (X_1, ..., X_{15})$.

Với độ tin cậy $1 - \alpha = 0{,}99$ thì $\alpha/2 = 0{,}005$ tra bảng phân phối Student có $t_{0{,}005}^{(14)} = 2{,}977$. Vậy với độ tin cậy 0,99 khoảng tin cậy đối xứng của μ là

$$(\overline{X} - \frac{S}{\sqrt{15}} 2{,}977; \overline{X} + \frac{S}{\sqrt{15}} 2{,}977)$$

Với mẫu cụ thể ta tính được $\overline{x} = 39{,}8$, $s^2 = 0{,}144 \rightarrow s = 0{,}379$. Vậy với độ tin cậy 0,99 qua mẫu cụ thể này, khoảng tin cậy

đối xứng của μ là:

$$(39{,}8 - \frac{0{,}379}{\sqrt{15}}2{,}977;\ 39{,}8 + \frac{0{,}379}{\sqrt{15}}2{,}977)$$

hay $(39{,}5023 < \mu < 40{,}0977)$.

Thí dụ 4. Phỏng vấn 5 gia đình có 3 người về chi phí hàng tháng cho nhu yếu phẩm thu được các số liệu sau: 150 ngàn đồng, 180 ngàn, 200 ngàn, 250 ngàn, 300 ngàn. Vậy phải phỏng vấn bao nhiêu gia đình cùng loại để với độ tin cậy 95% sai số của việc ước lượng chi phí trung bình hàng tháng cho nhu yếu phẩm không vượt quá 30 ngàn đồng. Giả thiết, chi phí hàng tháng cho nhu yếu phẩm là biến ngẫu nhiên phân phối chuẩn.

Giải. Gọi X là chi phí hàng tháng cho nhu yếu phẩm, theo giả thiết X phân phối chuẩn. Vậy chi phí trung bình chính là giá trị μ. Đây là bài toán xác định kích thước mẫu tối thiểu cho việc ước lượng tham số μ của phân phối $N(\mu, \sigma^2)$ khi chưa biết phương sai σ^2.

Theo phương pháp mẫu kép, từ mẫu sơ bộ kích thước $n = 5$ ta tìm được:

$$\bar{x} = \frac{150 + 180 + 200 + 250 + 300}{5} = 216 \text{ ngàn}$$

$$s^2 = [(150 - 216)^2 + (180 - 216)^2 + (200 - 216)^2 + (250 - 216)^2 + (300 - 216)^2]:4 = 3530$$

$$\varepsilon_0 = 30 \text{ ngàn}$$

$$t^4_{0,025} = 2{,}776$$

Vậy theo công thức (7.28) có:

$$n \geq \left[\frac{3530}{30^2}(2{,}776)^2\right] = 31 \text{ gia đình}$$

Như vậy phải phỏng vấn thêm 31 – 5 = 26 gia đình nữa.

Chú ý rằng khi kích thước mẫu n > 30 thì phân phối Student đã xấp xỉ phân phối chuẩn hóa do đó khi sử dụng các công thức (7.22 – 7.28) có thể dùng các giá trị phân phối chuẩn hóa để thay thế cho các giá trị của phân phối Student tương ứng.

Thí dụ 5. Để xác định kích thước trung bình của chi tiết do một máy sản xuất người ta lấy ngẫu nhiên 200 chi tiết để đo kích thước và thu được bảng số liệu sau (bảng 7.1). Với độ tin cậy 95% hãy ước lượng bằng khoảng tin cậy đối xứng kích thước trung bình của chi tiết do máy đó sản xuất. Giả thiết kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn.

Bảng 7.1

Kích thước chi tiết (cm)	Số chi tiết tương ứng
54,795 - 54,805	6
54,805 - 54,815	14
54,815 - 54,825	33
54,825 - 54,835	47
54,835 - 54,845	45
54,845 - 54,855	33
54,855 - 54,865	15
54,865 - 54,875	7
	n = 200

Giải. Gọi X là kích thước chi tiết do máy đó sản xuất. Theo giả thiết X phân phối chuẩn. Vậy kích thước trung bình của chi tiết chính là tham số μ. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng giá trị của tham số μ của phân phối $N(\mu,\sigma^2)$ khi chưa biết phương sai σ^2. Vậy khoảng tin cậy của μ là:

$$(\overline{X} - \frac{S}{n} t_{\alpha/2}^{(n-1)}; \overline{X} + \frac{S}{n} t_{\alpha/2}^{(n-1)})$$

Do n = 200 > 30 nên với 1 – α = 0,95 thì $t_{\alpha/2}^{(n-1)} = t_{0,025}^{(199)} \approx u_{0,025} = 1,96$. Với mẫu cụ thể để tìm $\bar{x}$ và s lập bảng sau:

x_i	n_i	$n_i x_i$	$n_i\ x_i^2$
54,80	6	328,80	18018,240
54,81	14	767,34	42057,905
54,82	33	1809,06	99172,669
54,83	47	2577,01	14197,450
54,84	45	1467,80	135334,150
54,85	33	1810,05	99281,242
54,86	15	822,90	45144,294
54,87	7	384,09	21075,018
	n = 200	10967,05	601380,950

Từ đó $\bar{x} = \frac{10967,05}{200} = 54,83525$

$$ms = \frac{60138095}{200} - (54,83525)^2 = 0,0002559$$

$$s = \sqrt{\frac{200}{199}0{,}0002559} = 0{,}0164$$

Vậy với độ tin cậy 0,95 qua mẫu cụ thể này khoảng tin cậy đối xứng của μ là:

$$(54{,}83525 - \frac{0{,}0164}{\sqrt{200}}1{,}96; 54{,}83525 + \frac{0{,}0164}{\sqrt{200}}1{,}96)$$

hay $(54{,}83298 < \mu < 54{,}83752)$

Thí dụ. Với các số liệu trong thí dụ A và bằng phần mềm Stata chúng ta ước lượng với độ tin cậy 0,95 thu nhập trung bình hàng năm của dân cư ba vùng và thu được kết quả sau

.centile x1 x2 x3.normal

- Normal, based on observed centiles -

Variable	Obs	Percentile	Centile	[95%Conf	Interval]
x1	100	50	1643	1631.245	1654.755
x2	100	50	1657.5	1645.848	1669.152
x3	100	50	1624	1612.068	1635.932

Nếu mẫu được chọn từ tổng thể hữu hạn theo phương thức không hoàn lại và $n > 0{,}1N$ thì biểu thức (7.20) của thống kê G trở thành:

$$G = T = \frac{\bar{X} - \mu}{\sqrt{\frac{N-n}{N-1}\cdot\frac{S^2}{n}}}$$

và việc xây dựng các công thức ước lượng cũng được tiến hành tương tự như mục trước.

2.3. Ước lượng hiệu hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu, trong đó các biến ngẫu nhiên X_1 và X_2 cùng phân phối chuẩn với các tham số đặc trưng tương ứng là μ_1, μ_2 và σ_1^2, σ_2^2, với μ_1 và μ_2 chưa biết. Để ước lượng sự chênh lệch $\mu_1 - \mu_2$ giữa hai kỳ vọng toán từ hai tổng thể lập hai mẫu ngẫu nhiên độc lập kích thước tương ứng là n_1 và n_2

$$W_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$$

từ đó tìm được các thống kê đặc trưng mẫu tương ứng là $\overline{X}_1, \overline{X}_2$ và S_1^2, S_2^2.

Để chọn thống kê G thích hợp ta xét các trường hợp sau:

1. Nếu đã biết các phương sai σ_1^2 và σ_2^2 ***của các tổng thể***

Lúc đó chọn thống kê

$$G = U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Từ mục 6 chương VI ta đã biết U phân phối chuẩn hóa N(0,1). Do đó với độ tin cậy $1 - \alpha$ cho trước tìm được cặp giá trị α_1 và α_2 ($\alpha_1 + \alpha_2 = \alpha$) và hai giá trị tới hạn chuẩn tương ứng là $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn:

$$P(U < u_{1\ \alpha_1}) = \alpha_1$$

$$P(U > u_{\alpha_2}) = \alpha_2$$

Từ đó:

$$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1 - \alpha$$

Bằng các phép biến đổi như đã làm ở các mục trước ta thu được khoảng tin cậy mức $(1 - \alpha)$ của $\mu_1 - \mu_2$ như sau:

$$P\left[(\overline{X}_1 - \overline{X}_2) - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\, u_{\alpha_2} < \mu_1 - \mu_2 \right.$$

$$\left. < (\overline{X}_1 - \overline{X}_2) + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\, u_{\alpha_1} \right] = 1 - \alpha \qquad (7.29)$$

Từ công thức tổng quát (7.29) có thể xây dựng được các công thức cụ thể như đã làm ở các mục trước:

+ Khoảng tin cậy đối xứng

$$P\left[(\overline{X}_1 - \overline{X}_2) - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\, u_{\alpha/2} < \mu_1 - \mu_2 \right.$$

$$\left. < (\overline{X}_1 - \overline{X}_2) + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\, u_{\alpha/2} \right] = 1 - \alpha \qquad (7.30)$$

+ Khoảng tin cậy bên phải

$$P\left[(\mu_1 - \mu_2) > (\overline{X}_1 - \overline{X}_2) - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\, u_{\alpha} \right] = 1 - \alpha \qquad (7.31)$$

+ Khoảng tin cậy bên trái

$$P\left[(\mu_1 - \mu_2) < (\overline{X}_1 - \overline{X}_2) + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\, u_{\alpha} \right] = 1 - \alpha \qquad (7.32)$$

2. Nếu chưa biết các phương sai σ_1^2 và σ_2^2 của các tổng thể, song giả thiết rằng $\sigma_1^2 = \sigma_2^2$

Lúc đó chọn thống kê

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (7.33)$$

trong đó $$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Ta đã biết thống kê T phân phối $T(n_1 + n_2 - 2)$ do đó tiến hành tương tự như ở các mục trước ta thu được khoảng tin cậy $(1 - \alpha)$ của hiệu $\mu_1 - \mu_2$ như sau:

$$P\left[(\overline{X}_1 - \overline{X}_2) - S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}t_{\alpha_2}^{(n_1+n_2-2)} < \mu_1 - \mu_2 < \right.$$

$$\left. < (\overline{X}_1 - \overline{X}_2) + S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}t_{\alpha_1}^{(n_1+n_2-2)}\right] = 1 - \alpha \qquad (7.34)$$

Từ công thức (7.34) có thể xây dựng các khoảng tin cậy đối xứng, khoảng tin cậy bên phải và bên trái như đã làm ở mục trước.

3. Nếu chưa biết các phương sai σ_1^2 và σ_2^2 ***của các tổng thể và không có căn cứ để cho rằng chúng bằng nhau***

Lúc đó chọn lập thống kê

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (7.35)$$

Ta đã biết thống kê T phân phối Student với số bậc tự do là:

$$k = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)C^2 + (1 - C)^2(n_1 - 1)}$$

với
$$C = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

do đó khoảng tin cậy mức $(1 - \alpha)$ của hiệu $\mu_1 - \mu_2$ như sau:

$$P\left[(\overline{X}_1 - \overline{X}_2) - \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\, t_{\alpha_2}^{(k)} < \mu_1 - \mu_2 < \right.$$

$$\left. < (\overline{X}_1 - \overline{X}_2) + \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\, t_{\alpha_1}^{(k)}\right] = 1 - \alpha \qquad (7.36)$$

Từ đó có thể xây dựng các khoảng tin cậy đối xứng, bên phải, bên trái tương ứng.

Việc xác định kích thước mẫu để ước lượng hiệu $\mu_1 - \mu_2$ của hai kỳ vọng toán được tiến hành như sau: Giả sử ta chủ trương lấy ra hai mẫu độc lập có kích thước như nhau và bằng n và giả sử các phương sai tổng thể đã biết và bằng σ_1^2 và σ_2^2. Lúc đó từ (7.30) suy ra kích thước n của hai mẫu đảm bảo với độ tin cậy $1 - \alpha$ cho trước độ dài của khoảng tin cậy không vượt quá giá trị I_0 cho trước là:

$$n \geq \left[\frac{(\sigma_1^2 + \sigma_2^2)}{\varepsilon_0^2} u_{\alpha/2}^2\right] \qquad (3.37)$$

nếu chưa biết σ_1^2 và σ_2^2 thì thay bằng ước lượng của chúng là S_1^2 và S_2^2 với S_1^2 và S_2^2 là phương sai của hai mẫu sơ bộ kích thước m_1 và m_2 được rút ra từ hai tổng thể nghiên cứu.

Thí dụ 6. Từ một chuồng nuôi lợn người ta lấy ra ngẫu nhiên bốn con đem cân và thu được trọng lượng tương ứng của chúng là 64, 66, 89 và 77 kg. Từ một chuồng khác lấy ra ba con đem cân thu được trọng lượng là 56, 71 và 53 kg. Với độ tin cậy 95% hãy ước lượng sự khác biệt về trọng lượng trung bình của hai chuồng lợn đó. Giả thiết trọng lượng của lợn phân phối chuẩn.

Giải. Gọi X_1 và X_2 tương ứng là trọng lượng của lợn ở hai chuồng nói trên, theo giả thiết X_1 và X_2 phân phối chuẩn. Vậy trọng lượng trung bình là μ_1 và μ_2. Đây là bài toán ước lượng hiệu số $\mu_1 - \mu_2$ khi chưa biết các phương sai của tổng thể. Nếu có thể cho rằng phương sai của chúng bằng nhau (chẳng hạn cả hai chuồng cùng nuôi một giống lợn và được chăm sóc như nhau) thì có thể từ công thức ước lượng (7.34) suy ra khoảng tin cậy đối xứng sau:

$$P\left[(\overline{X}_1 - \overline{X}_2) - S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha/2}^{(n_1+n_2-2)} < \mu_1 - \mu_2 < \right.$$

$$\left. < (\overline{X}_1 - \overline{X}_2) + S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha/2}^{(n_1+n_2-2)}\right] = 1-\alpha \qquad (7.38)$$

Từ hai mẫu cụ thể ta tìm được:

$$n_1 = 4 \qquad n_2 = 3$$

$$\overline{x}_1 = 74 \qquad \overline{x}_2 = 60$$

$$s_1^2 = 132{,}67 \qquad s_2^2 = 93$$

Từ đó:

$$S_p^2 = \frac{3.132{,}67 + 2.93}{4+3-2} = 116{,}8$$

và $$t_{\alpha/2}^{(n_1+n_2-2)} = t_{0,025}^{(5)} = 2{,}57$$

$$14 - \sqrt{116{,}8} \cdot \sqrt{\frac{1}{4}+\frac{1}{3}}.2{,}57 < \mu_1 - \mu_2 < 14 + \sqrt{116{,}8} \cdot \sqrt{\frac{1}{4}+\frac{1}{3}}.2{,}57$$

$$[-7{,}21 < \mu_1 - \mu_2 < 35{,}21]$$

4. Ước lượng hiệu hai kỳ vọng toán khi mẫu gồm các số liệu theo cặp

Ở các phần trước ta luôn giả thiết rằng có thể lấy ra hai mẫu độc lập từ hai tổng thể nghiên cứu. Trong nhiều trường hợp thực tế phải so sánh hai kỳ vọng toán trong điều kiện hai mẫu được rút ra từ cùng một tổng thể và nói chung là chúng phụ thuộc nhau. Lúc đó nếu hai mẫu bao gồm các cặp giá trị (X_1, Y_1) (X_2, Y_2) ... (X_n, Y_n) thì lúc đó với mỗi cặp giá trị của hai mẫu có thể xác định sự sai lệch trên từng cặp

$$D_i = X_i - Y_i \qquad (i = \overline{1,n}) \tag{7.39}$$

Lúc đó có thể coi D là biến ngẫu nhiên phân phối chuẩn với các tham số là μ_D và σ_D^2 và việc xử lý D cũng giống như đã làm với một mẫu ngẫu nhiên. Từ mẫu bao gồm n sai lệch D_i ta tìm được:

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{7.40}$$

và $$Se(\overline{D}) = \frac{S_D}{\sqrt{n}} \quad \text{với } S_D = \frac{\sqrt{\sum_{i=1}^{n}(D_i - \overline{D})^2}}{\sqrt{n-1}} \qquad (7.41)$$

Lúc đó thống kê

$$G = T = \frac{(\overline{D} - \mu_D)\sqrt{n}}{S_D} \sim T(n-1)$$

nên khoảng tin cậy mức $(1-\alpha)$ của μ_D là:

$$P\left[\overline{D} - \frac{S_D}{\sqrt{n}} t_{\alpha_2}^{(n-1)} < \mu_D < \overline{D} + \frac{S_D}{\sqrt{n}} t_{\alpha_1}^{(n-1)}\right] = 1 - \alpha \qquad (7.42)$$

Từ đó có thể xây dựng các khoảng tin cậy đối xứng, bên phải, bên trái tương ứng.

Lúc đó kích thước mẫu tối thiểu n theo độ tin cậy $(1-\alpha)$ cho trước và độ dài của khoảng tin cậy tối đa là I_0 cho trước được xác định bằng công thức

$$n \geq \left[\frac{S_D^2}{\varepsilon_0^2}(t_{\alpha/2}^{m-1})^2\right] \qquad (7.43)$$

trong đó S_D^2 là phương sai của mẫu sơ bộ kích thước m.

Thí dụ 7. Giả sử để theo dõi tốc độ tăng trọng trung bình của một đàn lợn người ta lấy ngẫu nhiên 4 con và cân trọng lượng của chúng vào đầu tháng và cuối tháng. Kết quả thu được như sau:

Thứ tự	TL đầu tháng X_2	TL cuối tháng X_1	d_i	$(d_i - \bar{d})$	$(d_i - \bar{d})^2$
1	57	64	7	-4	16
2	57	66	9	-2	4
3	73	89	16	5	25
4	65	77	12	1	1
			$\bar{d} = \frac{44}{4} = 11$		$S_d^2 = \frac{46}{3} = 15,3$

Với độ tin cậy 95% hãy ước lượng mức tăng trọng trung bình của đàn lợn trong tháng đó.

Giải. Gọi X_1 và X_2 là trọng lượng của lợn vào cuối tháng và đầu tháng. Giả sử X_1 và X_2 phân phối chuẩn. Lúc đó μ_1 và μ_2 là trọng lượng trung bình tương ứng và mức tăng trọng trung bình được xác định bằng giá trị $\mu_D = \mu_1 - \mu_2$.

Từ biểu thức (7.42) ta có khoảng tin cậy đối xứng của μ_D như sau:

$$P\left[\bar{D} - \frac{S_D}{\sqrt{n}} t_{\alpha/2}^{(n-1)} < \mu_D < \bar{D} + \frac{S_D}{\sqrt{n}} t_{\alpha/2}^{(n-1)}\right] = 1 - \alpha$$

Từ mẫu cụ thể ta tính được:

$$\bar{d} = 11;\ S_d^2 = 15,3;\ t_{\alpha/2}^{(n-1)} = t_{0,025}^{(3)} = 3,18$$

$$\text{Vậy} \quad 11 - \frac{\sqrt{15,3}}{\sqrt{4}} 3,18 < \mu_D < 11 + \frac{\sqrt{15,3}}{\sqrt{4}} 3,18$$

$$\Rightarrow \quad (4,78 < \mu_D < 17,22)$$

2.4. Ước lượng xác suất p của biến ngẫu nhiên phân phối theo quy luật không - một

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối theo một quy luật nào đó khác với quy luật chuẩn. Trong trường hợp này, để ước lượng giá trị của kỳ vọng toán m chưa biết ta có thể chọn thống kê

$$G = U = \frac{(\overline{X} - m)\sqrt{n}}{\sigma}$$

nếu đã biết phương sai σ^2 của X, hoặc thống kê

$$G = U = \frac{(\overline{X} - m)\sqrt{n}}{S}$$

nếu chưa biết phương sai σ^2 của X.

Từ mục §6 chương VI ta đã biết nếu kích thước mẫu n đủ lớn, cả hai thống kê trên đều phân phối xấp xỉ chuẩn hóa N(0,1). Do đó vẫn có thể tiến hành thủ tục ước lượng bằng khoảng tin cậy như đã xét ở phần trên. Sau đây ta sẽ xét một trường hợp cụ thể khá thông dụng trong thực tế là bài toán ước lượng kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật không - một: A(p).

Giả sử trong tổng thể kích thước N có M phần tử mang dấu hiệu nghiên cứu. Nếu lấy ngẫu nhiên ra 1 phần tử và gọi X là số phần tử mang dấu hiệu nghiên cứu được lấy ra thì X là biến ngẫu nhiên phân phối theo quy luật không - một:

X	0	1
P	1-p	p

trong đó p là xác suất để lấy ngẫu nhiên một phần tử thì được phần tử mang dấu hiệu nghiên cứu

$$p = \frac{M}{N}$$

Ta đã biết trong quy luật không - một thì $E(X) = p$ và $V(X) = \frac{p(1-p)}{n}$, như vậy ước lượng kỳ vọng toán của quy luật này cũng chính là ước lượng xác suất p, mà p lại là tần suất của tổng thể, phản ánh cơ cấu của tổng thể theo dấu hiệu nghiên cứu đó. Như vậy đây chính là bài toán ước lượng cơ cấu của tổng thể. Từ mục §6 chương VI ta thấy nếu thỏa mãn điều kiện:

$$n > 5 \text{ và } \frac{\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right|}{\sqrt{n}} < 0,3$$

thì thống kê

$$G = U = \frac{(f-p)}{Se(f)} = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \qquad (7.44)$$

phân phối xấp xỉ N(0,1). Vì vậy, với độ tin cậy $(1 - \alpha)$ cho trước có thể tìm được hai giá trị tới hạn chuẩn $u_{1-\alpha/2}$ và $u_{\alpha/2}$ thỏa mãn điều kiện:

$$P(U < u_{1-\alpha/2}) = \frac{\alpha}{2}$$

$$P(U > u_{\alpha/2}) = \frac{\alpha}{2}$$

Từ đó $P(u_{1-\alpha/2} < U < u_{\alpha/2}) = 1 - (\alpha/2 + \alpha/2) = 1 - \alpha$

Do $u_{\alpha/2} = -u_{1-\alpha/2}$ nên

$$P(-u_{\alpha/2} < U < u_{\alpha/2}) = 1 - \alpha \qquad (7.45)$$

Phép biến đổi tương đương đối với biểu thức trong ngoặc của (7.45) được tiến hành như sau: Thay U từ (7.44) vào ta có

$$-u_{\alpha/2} < \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} < u_{\alpha/2}$$

tương đương với $\left|\frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}}\right| < u_{\alpha/2}$

Bình phương hai vế ta thu được $n(f-p)^2 < p(1-p)\,u^2_{\alpha/2}$. Khai triển và chuyển vế ta thu được bất phương trình:

$$(n + u^2_{\alpha/2})p^2 - (2nf + u^2_{\alpha/2})p + nf^2 < 0$$

Giải ra ta được:

$$p_1, p_2 = \frac{nf + \frac{1}{2}u^2_{\alpha/2} \pm u_{\alpha/2}\sqrt{nf(1-f) + \frac{1}{4}u^2_{\alpha/2}}}{n + u^2_{\alpha/2}} \qquad (7.46)$$

Như vậy với độ tin cậy $(1 - \alpha)$ khoảng tin cậy đối xứng của p là:

$$(p_1 < p < p_2) \qquad (7.47)$$

với $p_1$và p_2 được xác định từ (7.46)

Việc áp dụng các công thức (7.46) và (7.47) khá phức tạp và chỉ cho phép tìm khoảng tin cậy đối xứng của p. Do đó nếu có thể điều tra một mẫu có kích thước n khá lớn ($n \geq 100$) thì ta có thể chọn thống kê

$$G = U = \frac{(f-p)\sqrt{n}}{\sqrt{f(1-f)}} \qquad (7.48)$$

Nó cũng phân phối xấp xỉ N(0,1), do đó với độ tin cậy $(1 - \alpha)$ cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$, từ đó tìm được hai giá trị tới hạn chuẩn tương ứng là $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn điều kiện:

$$P(U < u_{1-\alpha_1}) = \alpha_1$$

và $$P(U > u_{\alpha_2}) = \alpha_2$$

từ đó : $$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha$$

Thay giá trị của U từ (7.48) vào và sử dụng tính chất $u_{\alpha_1} = -u_{1-\alpha_1}$ sau phép biến đổi tương đương ta có

$$p\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha_2} < p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha_1}\right) = 1 - \alpha \qquad (7.49)$$

Như vậy, với độ tin cậy $(1 - \alpha)$ khoảng tin cậy của p có dạng:

$$\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha_2}; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha_1}\right) \qquad (7.50)$$

Từ biểu thức tổng quát (7.50) có thể xây dựng được các khoảng tin cậy cụ thể:

- Khoảng tin cậy đối xứng khi $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

$$\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2}; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2}\right) \qquad (7.51)$$

- Khoảng tin cậy bên phải khi $\alpha_1 = 0; \alpha_2 = \alpha$

$$\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha}; +\infty\right) \qquad (7.52)$$

- Khoảng tin cậy bên trái khi $\alpha_2 = 0; \alpha_1 = \alpha$

$$\left(-\infty;\ f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_\alpha\right) \tag{7.53}$$

- Độ dài khoảng tin cậy ngắn nhất trong trường hợp khoảng tin cậy đối xứng:

$$I = 2\varepsilon = \frac{2\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2} \tag{7.54}$$

- Kích thước mẫu tối thiểu n đảm bảo độ tin cậy $(1 - \alpha)$ cho trước và độ dài khoảng tin cậy không vượt quá giá trị I_0 cho trước

$$n \geq \left[\frac{f(1-f)}{\varepsilon_0^2} u_{\alpha/2}^2\right] \tag{7.55}$$

trong đó f là tần suất của mẫu sơ bộ kích thước $m \geq 2$.

Việc chuyển từ các khoảng tin cậy ngẫu nhiên sang các khoảng tin cậy bằng số qua một mẫu cụ thể cũng được tiến hành như đã làm ở các phần trên.

Thí dụ 8. Kiểm tra ngẫu nhiên 400 sản phẩm do một máy sản xuất thấy có 20 phế phẩm. Với độ tin cậy 0,95 hãy ước lượng tỷ lệ phế phẩm tối đa của máy đó.

Giải. Gọi p là tỷ lệ phế phẩm của máy đó. Như vậy p là cơ cấu của tập hợp sản phẩm do máy đó sản xuất theo dấu hiệu "phế phẩm". Đây là bài toán ước lượng tham số p của quy luật phân phối A(p) bằng khoảng tin cậy bên trái. Vậy khoảng tin cậy của p có dạng (7.53):

$$\left(-\infty;\ f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_\alpha\right)$$

Qua mẫu cụ thể ta có $f = \frac{20}{400} = 0,05$. Với $1 - \alpha = 0,95$ $\rightarrow u_{0.05} = 1,645$. Vậy với độ tin cậy 0,95 qua mẫu cụ thể này khoảng tin cậy của p là:

$$(-\alpha; 0,05 + \frac{\sqrt{0,05.0,95}}{\sqrt{400}} 1,645)$$

hay $p < 0,0679$, hay tỷ lệ phế phẩm tối đa của máy đó là 6,79%.

Thí dụ 9. Một vùng có 2000 hộ gia đình. Để điều tra nhu cầu tiêu dùng một loại hàng hóa tại vùng đó người ta nghiên cứu ngẫu nhiên 100 gia đình và thấy có 60 gia đình có nhu cầu về loại hàng hóa trên. Với độ tin cậy 0,95 hãy ước lượng bằng khoảng tin cậy đối xứng số gia đình trong vùng có nhu cầu về loại hàng hóa đó.

Giải: Gọi M là số gia đình trong vùng có nhu cầu về loại hàng hóa đó thì tỷ lệ gia đình có nhu cầu này là:

$$p = \frac{m}{2000}$$

Như vậy bài toán được đưa về việc ước lượng bằng khoảng tin cậy đối xứng cơ cấu p của tổng thể. Theo công thức (7.51) khoảng tin cậy của p là:

$$\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2}; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2} \right)$$

Qua mẫu cụ thể ta có $f = \frac{60}{100} = 0,6$; $u_{\alpha/2} = u_{0,025} = 1,96$

Vậy với độ tin cậy 0,95 khoảng tin cậy đối xứng của p qua mẫu cụ thể này là:

$$\left(0,6-\frac{\sqrt{0,6.0,4}}{\sqrt{100}}1,96;\ 0,6+\frac{\sqrt{0,6.0,4}}{\sqrt{100}}1,96\right)$$

hay $(0,504 < p < 0,696)$.

Do $M = pN = p.2000$ nên ta có khoảng tin cậy đối xứng của M với độ tin cậy 0,95 qua mẫu cụ thể này là $(1008 \le M \le 1392)$.

Nếu sử dụng các công thức (7.46) và (7.47) thì kết quả là:

$$p_1 = \frac{0,6.100+\frac{1}{2}(1,96)^2-1,96\sqrt{\frac{1}{4}(1,96)^2+0,6.0,4.100}}{100+(1,96)^2} = 0,502$$

$$p_2 = \frac{0,6.100+\frac{1}{2}(1,96)^2+1,96\sqrt{\frac{1}{4}(1,96)^2+0,6.0,4.100}}{100+(1,96)^2} = 0,691$$

vậy $(0,502 < p < 0,691)$

do đó $(1004 \le M \le 1382)$

2.5. Ước lượng hiệu hai tham số p của hai biến ngẫu nhiên phân phối không - một

Giả sử có hai tổng thể nghiên cứu, trong đó các biến ngẫu nhiên X_1 và X_2 phân phối không - một với tham số tương ứng là p_1 và p_2. Nếu từ hai tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước n_1 và n_2 và tìm được các tần suất tương ứng là f_1 và f_2 thì có thể chọn lập thống kê

$$G = U = \frac{(f_1-f_2)-(p_1-p_2)}{S_f} \qquad (7.56)$$

$$\text{với } S_f = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}$$

Từ mục 6 chương VI ta đã biết nếu $n_1 > 30$ và $n_2 > 30$ thì biến ngẫu nhiên U sẽ phân phối xấp xỉ N(0,1). Do đó với cách tiến hành giống như ở mục trước ta thu được khoảng tin cậy mức $(1-\alpha)$ của hiệu $p_1 - p_2$ như sau:

$$P\left[(f_1 - f_2) - S_f u_{\alpha_2} < p_1 - p_2 < (f_1 - f_2) + S_f u_{\alpha_1}\right] = 1 - \alpha \qquad (7.57)$$

Xuất phát từ công thức (7.57) có thể xây dựng các khoảng tin cậy đối xứng, khoảng tin cậy bên phải và bên trái cho hiệu $p_1 - p_2$.

Công thức tìm kích thước mẫu tối thiểu cần điều tra với cả 2 mẫu sao cho với độ tin cậy $1 - \alpha$ cho trước, độ dài khoảng tin cậy không vượt quá $I_0 = 2\varepsilon_0$ cho trước có dạng:

$$n \geq \frac{\left[f_1(1-f_1) + f_2(1-f_2)\right]}{\varepsilon_0^2} u_{\alpha/2}^2 \qquad (7.58)$$

với f_1 và f_2 là tần suất của các mẫu sơ bộ kích thước m.

Thí dụ 10. Doanh nghiệp dự định đưa sản phẩm của mình vào hai thị trường khác nhau. Bán thử sản phẩm cho 100 khách hàng tiềm năng của thị trường thứ nhất thì có 50 người mua. Còn với thị trường thứ hai, khi bán thử sản phẩm cho 50 khách hàng tiềm năng thì có 20 người mua. Với độ tin cậy 95% hãy ước lượng mức độ chênh lệch về thị phần mà doanh nghiệp có thể đạt được tại hai thị trường đó.

Giải. Gọi p_1 và p_2 tương ứng là thị phần mà doanh nghiệp có thể đạt được ở hai thị trường. Vậy mức độ chênh lệch về thị phần là hiệu $p_1 - p_2$. Với n_1 và n_2 đủ lớn, để ước

lượng khoảng tin cậy đối xứng của $p_1 - p_2$, trong công thức (7.57) ta lấy $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ và thu được công thức ước lượng

$$P[(f_1 - f_2) - S_f u_{\alpha/2} < p_1 - p_2 < (f_1 - f_2) + S_f u_{\alpha/2}] = 1 - \alpha$$

Với hai mẫu cụ thể

$$n_1 = 100 \qquad n_2 = 50$$

$$f_1 = 0,5 \qquad f_2 = 0,4$$

ta tìm được

$$S_f = \sqrt{\frac{0,5.0,5}{100} + \frac{0,4.0,6}{50}} = 0,0854$$

với $1 - \alpha = 0,95 \Rightarrow u_{\alpha/2} = u_{0,025} = 1,96$

Vậy

$$[(0,5 - 0,4) - 0,0854.1,96 < p_1 - p_2 < (0,5 - 0,4) + 0,0854.1,96]$$

$$[0,0674 < p_1 - p_2 < 0,2674]$$

Vậy với độ tin cậy 95%, thị phần mà doanh nghiệp có thể đạt được ở thị trường thứ nhất cao hơn ở thị trường thứ hai từ 6,74% đến 26,74%.

2.6. Ước lượng phương sai của biến ngẫu nhiên phân phối theo quy luật chuẩn

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối theo quy luật chuẩn $N(\mu,\sigma^2)$ nhưng chưa biết phương sai σ^2 của nó. Để ước lượng σ^2 từ tổng thể lập mẫu ngẫu nhiên kích thước n

$$W = (X_1, X_2, \ldots X_n)$$

Để chọn thống kê G thích hợp ta xét hai trường hợp sau:

1. Đã biết kỳ vọng toán μ của biến ngẫu nhiên gốc X trong tổng thể. Lúc đó ta chọn thống kê

$$G = \chi^2 = \frac{nS^{*2}}{\sigma^2} \tag{7.59}$$

Từ mục §6 chương VI ta đã biết thống kê χ^2 nói trên phân phối theo quy luật "khi bình phương" với n bậc tự do: $\chi^2(n)$. Do đó với độ tin cậy $(1 - \alpha)$ cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ từ đó tìm được hai giá trị tới hạn "khi bình phương" tương ứng là $\chi^{2(n)}_{1-\alpha_1}$ và $\chi^{2(n)}_{\alpha_2}$ thỏa mãn điều kiện:

$$P(\chi^2 < \chi^{2(n)}_{1-\alpha_1}) = \alpha_1 \text{ và}$$

$$P(\chi^2 > \chi^{2(n)}_{\alpha_2}) = \alpha_2 \text{ do đó}$$

$$P(\chi^{2(n)}_{1-\alpha_1} < \chi^2 < \chi^{2(n)}_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha \tag{7.60}$$

Thay giá trị của χ^2 từ (7.59) vào (7.60) và giải ra theo σ^2, ta thu được biểu thức tương đương:

$$P\left(\frac{nS^{*2}}{\chi^{2(n)}_{\alpha_2}} < \sigma^2 < \frac{nS^{*2}}{\chi^{2(n)}_{1-\alpha_1}}\right) \tag{7.61}$$

Như vậy, với độ tin cậy $(1 - \alpha)$ khoảng tin cậy của σ^2 có dạng:

$$\left(\frac{nS^{*2}}{\chi^{2(n)}_{\alpha_2}} < \sigma^2 < \frac{nS^{*2}}{\chi^{2(n)}_{1-\alpha_1}}\right) \tag{7.62}$$

Từ khoảng tin cậy tổng quát (7.62) có thể xây dựng được khoảng tin cậy trong một số trường hợp cụ thể sau:

- Nếu $\alpha_1 = \alpha_2 = \alpha/2$, khoảng tin cậy có dạng:

$$\left(\frac{nS^{*2}}{\chi^{2(n)}_{\alpha/2}}; \frac{nS^{*2}}{\chi^{2(n)}_{1-\alpha/2}}\right) \tag{7.63}$$

Ta chú ý rằng khoảng tin cậy này không đối xứng.

- Nếu $\alpha_1 = 0; \alpha_2 = \alpha$, ta có khoảng tin cậy bên phải của σ^2:

$$\left(\frac{nS^{*2}}{\chi^{2(n)}_{\alpha}}; +\infty\right) \tag{7.64}$$

- Nếu $\alpha_2 = 0; \alpha_1 = \alpha$, ta có khoảng tin cậy bên trái của σ^2:

$$\left(0; \frac{nS^{*2}}{\chi^{2(n)}_{1-\alpha}}\right) \tag{7.65}$$

Với một mẫu cụ thể $w = x_1, x_2, \ldots x_n$ có thể xác định được một khoảng tin cậy cụ thể bằng số của σ^2 giống như đã làm ở các phần trên.

Thí dụ 11. Mức hao phí nguyên liệu cho một đơn vị sản phẩm là biến ngẫu nhiên phân phối chuẩn với trung bình là 20 gam. Để ước lượng mức độ phân tán của mức hao phí này người ta cân thử 25 sản phẩm và thu được kết quả sau:

Hao phí nguyên liệu (gam)	19,5	20,0	20,5
Số sản phẩm tương ứng	5	18	2

Với độ tin cậy $1 - \alpha = 0,90$, hãy ước lượng σ^2 nếu:

$$\alpha_1 = \alpha_2 = \frac{\alpha}{2} = 0,05$$

Giải. Gọi X là mức hao phí nguyên liệu cho 1 đơn vị sản

phẩm, X phân phối chuẩn với kỳ vọng toán đã biết $\mu = 20$. Đây là bài toán ước lượng phương sai của phân phối $N(\mu,\sigma^2)$ khi đã biết μ. Vậy theo công thức (7.63) khoảng tin cậy của σ^2 là:

$$\left(\frac{nS^{*2}}{\chi^{2(n)}_{\alpha/2}} < \sigma^2 < \frac{nS^{*2}}{\chi^{2(n)}_{1-\alpha/2}}\right)$$

Tra bảng giá trị χ^2 ta có:

$$\chi^{2(25)}_{\alpha/2} = \chi^{2(25)}_{0,05} = 37,65$$

$$\chi^{2(25)}_{1-\alpha/2} = \chi^{2(25)}_{0,95} = 14,61$$

Để tìm s^{*2} ta lập bảng

X_i	n_i	$(X_i - \mu)$	$(X_i - \mu)^2$	$n_i (X_i - \mu)^2$
19,5	5	–0,5	0,25	1,25
20,0	18	0,0	0,00	0,00
20,5	2	0,5	0,25	0,50
	n = 25			Σ = 1,75

$$s^{*2} = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \mu)^2 = \frac{1,75}{25} = 0,07$$

Vậy với độ tin cậy 0,90, qua mẫu cụ thể này, khoảng tin cậy của σ^2 là:

$$\left(\frac{25.0,07}{37,65}; \frac{25.0,07}{14,61}\right) \text{ hay } (0,0464 < \sigma^2 < 0,1198)$$

Để tìm khoảng tin cậy của σ ta chỉ cần lấy căn bậc hai của cả ba vế trong các công thức ước lượng trên.

2. Chưa biết kỳ vọng toán μ của biến ngẫu nhiên gốc X trong tổng thể. Lúc đó ta chọn thống kê

$$G = \chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad (7.66)$$

Từ mục §6 chương VI ta đã biết thống kê χ^2 nói trên phân phối theo quy luật "khi bình phương" với (n – 1) bậc tự do. Vì vậy, với độ tin cậy (1 – α) cho trước, có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$, từ đó tìm được hai giá trị tới hạn "khi bình phương" tương ứng là $\chi^{2(n-1)}_{1-\alpha_1}$ và $\chi^{2(n-1)}_{\alpha_2}$ thỏa mãn điều kiện:

$$P(\chi^2 < \chi^{2(n-1)}_{1-\alpha_1}) = \alpha_1 \text{ và}$$

$$P(\chi^2 > \chi^{2(n-1)}_{\alpha_2}) = \alpha_2 \text{ do đó}$$

$$P(\chi^{2(n-1)}_{1-\alpha_1} < \chi^2 < \chi^{2(n-1)}_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha \qquad (7.67)$$

Thay giá trị của χ^2 từ (7.66) vào (7.67) và giải ra σ^2 ta có:

$$P\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha_2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha_1}}\right) = 1 - \alpha$$

Vậy với độ tin cậy (1 – α), khoảng tin cậy của σ^2 là:

$$\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha_2}}; \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha_1}}\right) \qquad (7.68)$$

Từ công thức (7.68) ta cũng xây dựng được các khoảng tin cậy trong những trường hợp cụ thể sau:

- Nếu $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, khoảng tin cậy của σ^2 là:

$$\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha/2}}; \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha/2}}\right) \tag{7.69}$$

- Nếu $\alpha_1 = 0$; $\alpha_2 = \alpha$, ta có khoảng tin cậy bên phải của σ^2:

$$\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha}}; +\infty\right) \tag{7.70}$$

- Nếu $\alpha_2 = 0$; $\alpha_1 = \alpha$, ta có khoảng tin cậy bên trái của σ^2:

$$\left(0; \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha}}\right) \tag{7.71}$$

Việc xác định các khoảng tin cậy bằng số qua một mẫu cụ thể cũng tiến hành giống như đã làm ở các phần trên.

Thí dụ 12. Với độ tin cậy 0,95 hãy ước lượng phương sai của kích thước các chi tiết trên cơ sở các số liệu mẫu cho trong bảng (7.1), biết $\alpha_1 = \alpha_2 = \alpha/2 = 0{,}025$.

Giải. Đây là bài toán ước lượng phương sai của phân phối $N(\mu, \sigma^2)$ khi chưa biết μ. Vậy khoảng tin cậy của σ^2 có dạng:

- Nếu $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, khoảng tin cậy của σ^2 là:

$$\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha/2}}; \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha/2}}\right)$$

Qua mẫu cụ thể ta đã tìm được:

$$s^2 = 0{,}0002689; \quad n = 200$$

Tra bảng χ^2 : $\chi^{2(199)}_{0,975} \approx 198{,}98$; $\chi^{2(199)}_{0,025} \approx 284{,}8$

Vậy với độ tin cậy 0,95 qua mẫu cụ thể này, khoảng tin cậy của σ^2 là:

$$\left(\frac{199.0,0002689}{284,8}; \frac{199.0,0002689}{198,98}\right)$$

hay $(0,000188 < \sigma^2 < 0,000269)$.

Tương tự như ở mục trước, để tìm khoảng tin cậy của σ ta chỉ cần lấy căn bậc hai của các công thức ước lượng tương ứng.

2.7. Ước lượng tỷ số của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu trong đó các biến X_1 và X_2 cùng phân phối chuẩn với các phương sai α_1^2 và α_2^2 chưa biết. Từ các tổng thể trên lập hai mẫu ngẫu nhiên độc lập kích thước n_1 và n_2, từ đó tìm được các phương sai mẫu S_1^2 và S_2^2 và tạo lập thống kê

$$G = F = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \quad (S_1^2 > S_2^2) \tag{7.72}$$

Từ mục §6 chương VI ta đã biết thống kê F phân phối $F(n_1 - 1, n_2 - 1)$. Do đó với độ tin cậy $1 - \alpha$ cho trước tìm được cặp giá trị α_1 và α_2 $(\alpha_1 + \alpha_2 = \alpha)$ từ đó tìm được hai giá trị tới hạn $f_{1-\alpha_1}^{(n_1-1,n_2-1)}$ và $f_{\alpha_2}^{(n_1-1,n_2-1)}$ thỏa mãn điều kiện

$$P\left(F < f_{1-\alpha_1}^{(n_1-1,n_2-1)}\right) = \alpha_1$$

$$P\left(F > f_{\alpha_2}^{(n_1-1,n_2-1)}\right) = \alpha_2$$

do đó

$$P\left(f_{1-\alpha_1}^{(n_1-1,n_2-1)} < F < f_{\alpha_2}^{(n_1-1,n_2-1)}\right) = 1-\alpha \qquad (7.73)$$

Thay giá trị của F từ (7.72) vào (7.73) và sau khi chuyển vế ta thu được biểu thức

$$P\left[\frac{S_2^2}{S_1^2} f_{1-\alpha_1}^{(n_1-1,n_2-1)} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2} f_{\alpha_2}^{(n_1-1,n_2-1)}\right] = 1-\alpha$$

Từ đó

$$P\left[\frac{S_1^2}{S_2^2} f_{1-\alpha_2}^{(n_2-1,n_1-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha_1}^{(n_2-1,n_1-1)}\right] = 1-\alpha \quad (7.74)$$

Từ (7:74) có thể xây dựng được các khoảng tin cậy tương ứng với mỗi cặp giá trị α_1 và α_2.

Việc tìm khoảng tin cậy của $\frac{\sigma_1}{\sigma_2}$ được thực hiện bằng cách lấy căn bậc hai của cả ba vế của biểu thức (7.74).

Thí dụ 13. Giá cổ phiếu của hai công ty A và B là các biến ngẫu nhiên phân phối chuẩn. Theo dõi giá cổ phiếu của hai công ty đó trong 10 ngày tìm được phương sai mẫu tương ứng là 0,51 và 0,2. Với độ tin cậy 0,9 hãy ước lượng tỷ số của hai phương sai của giá cổ phiếu của hai công ty đó.

Giải. Gọi X_1 và X_2 là giá cổ phiếu của hai công ty A và B. Theo giả thiết X_1 và X_2 phân phối chuẩn. Vậy các phương sai là σ_1^2 và σ_2^2. Từ (7.74) ta lấy $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ và thu được khoảng tin cậy hai phía sau đây:

$$P\left[\frac{S_1^2}{S_2^2} f_{1-\alpha/2}^{(n_2-1,n_1-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}^{(n_2-1,n_1-1)}\right] = 1-\alpha$$

Từ hai mẫu cụ thể ta có:

$$n_1 = 10 \qquad n_2 = 10$$

$$s_1^2 = 0,51 \qquad s_2^2 = 0,2$$

và với $1 - \alpha = 0,9$ thì

$$f_{0,05}^{(9,9)} = 3,18 \text{ và } f_{0,95}^{(9,9)} = \frac{1}{f_{0,05}^{(9,9)}} = \frac{1}{3,18} = 0,31$$

Từ đó $$\frac{0,51}{0,2}.0,31 < \frac{\sigma_1^2}{\sigma_2^2} < \frac{0,51}{0,2}.3,18$$

$$\Rightarrow 0,79 < \frac{\sigma_1^2}{\sigma_2^2} < 8,11$$

2.8. Ước lượng trung vị của tổng thể nghiên cứu

Cho đến nay, đối với các tham số đặc trưng xu hướng trung tâm của tổng thể nghiên cứu, ta mới chỉ quan tâm đến việc ước lượng trung bình mà chưa xét đến trung vị hay mốt. Đó cũng là điều hợp lý vì tham số của tổng thể mà ta quan tâm chủ yếu là tổng số mà tổng số lại gắn liền với trung bình. Chẳng hạn để đánh giá tổng doanh số bán của các đại lý chỉ cần tìm doanh số trung bình của mỗi đại lý sau đó đem nhân với tổng số đại lý là được.

Không thể ước lượng tổng doanh số dựa trên trung vị vì nó chỉ sử dụng giá trị ở chính giữa và bỏ qua mọi giá trị khác, nhất là những giá trị đột xuất, khác biệt so với các giá trị khác. Trong kinh tế người ta cũng thường quan tâm đến giá trị tổng số như tổng sản phẩm quốc nội, tổng sản phẩm công nghiệp, nông nghiệp... Trong những trường hợp này ta thường dùng trung bình vì nó chứa đựng thông tin về mọi phần tử của tổng thể nghiên cứu.

Tuy nhiên trong một số trường hợp trung vị lại tỏ ra hữu dụng. Chẳng hạn khi số liệu điều tra về tổng thể không đầy đủ hay không hoàn toàn tin cậy thì nên dùng trung vị thay cho trung bình. Hơn nữa trung vị có ưu thế hơn trung bình là nó không đòi hỏi biến ngẫu nhiên trong tổng thể phải phân phối chuẩn, do đó sẽ thuận lợi nhiều trong quy nạp thống kê.

Sau đây ta sẽ xem xét phương pháp ước lượng trung vị của tổng thể.

Để ước lượng trung vị m_d của tổng thể, người ta cũng xuất phát từ x_d của mẫu như là một ước lượng điểm của m_d. Tuy nhiên, để nâng cao độ chính xác của ước lượng ta sẽ đưa ra phương pháp để xây dựng khoảng tin cậy cho m_d.

Ta sẽ minh họa việc tìm khoảng tin cậy cho m_d qua thí dụ sau: Giả sử lấy ngẫu nhiên 9 gia đình tại một thành phố và điều tra thu nhập hàng năm X của họ thu được các số liệu được sắp xếp từ thấp đến cao như sau:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0		6 triệu		x_d = 9 triệu		13 triệu		20 triệu

từ đó ta có $x_d = x_5 = 9$ triệu. Để xây dựng khoảng tin cậy cho m_d ta có thể dịch chuyển sang hai phía, chẳng hạn hai giá trị, như vậy

$$6 \text{ triệu} \le m_d \le 13 \text{ triệu}. \qquad (7.75)$$

Song vấn đề là độ tin cậy của khoảng tin cậy (7.75) là bao nhiêu. Để giải đáp vấn đề này ta lập luận như sau: Nếu trung vị của thu nhập là m_d thì một nửa số gia đình của thành phố đó phải có thu nhập hàng năm lớn hơn giá trị đó. Song điều đó không có nghĩa là một nửa thu nhập của mẫu

cũng sẽ phải lớn hơn m_d mà chỉ có nghĩa là mỗi giá trị của mẫu ngẫu nhiên rút ra từ tổng thể có xác suất lớn hơn m_d là 0,5 giống như xác suất để được mặt sấp khi tung đồng xu. Vậy là ta có hai biến cố tương đương: "Giá trị của mẫu ngẫu nhiên lớn hơn m_d "tương đương với" được mặt sấp khi tung đồng xu".

Với quan niệm như vậy thì số gia đình có thu nhập cao hơn m_d sẽ là biến ngẫu nhiên phân phối theo quy luật nhị thức với các tham số là n (kích thước mẫu) và p = 0,5. Như vậy theo công thức của quy luật nhị thức, áp dụng vào thí dụ đang xét, có thể tìm được xác suất để có 7 gia đình trong 9 gia đình được điều tra sẽ có thu nhập lớn hơn hoặc bằng trung vị m_d. Nó được tính theo công thức

$$P(X \geq 7) = P_9(7) + P_9(8) + P_9(9) = 0,09$$

Tương tự như vậy có thể tìm được xác suất để có 7 gia đình có thu nhập nhỏ hơn hoặc bằng m_d. Nó cũng bằng 0,09. Vậy xác suất để khoảng tin cậy (7.75) chứa đựng giá trị trung vị m_d là $1 - 2.0,09 = 0,82$.

Hiển nhiên là nếu n đủ lớn thì có thể dùng phân phối chuẩn để thay thế cho quy luật nhị thức. Chẳng hạn nếu điều tra thu nhập của 25 gia đình lấy ngẫu nhiên từ một thành phố và thu được các số liệu đã sắp xếp theo trình tự tăng dần như sau:

4, 5, 5, 5, 5, 6, 7, 9, 10, 12, 13, 14, 14, 15, 17, 18, 18, 19, 23, 25, 27, 30, 32, 89, 40, 52

Lúc đó trung vị của mẫu là giá trị thứ 13: x_{13}= 14. Đó là ước lượng điểm của m_d.

Nếu ta chọn khoảng tin cậy của m_d là

$$7 \le md \le 27$$

thì độ tin cậy của ước lượng là

$$1 - 2.P(X \ge 19)$$

Ta tìm xác suất $P(X \ge 19)$. Với n đủ lớn X phân phối xấp xỉ chuẩn với

$$\mu = np = 25.0,5 = 12,5$$

và $$\sigma = \sqrt{np(1-p)} = 2,5$$

Do đó

$$P(X \ge 19) = P\left(\frac{X - 12,5}{2,5} \ge \frac{19 - 12,5}{2,5}\right) = P(U \ge 2,6)$$

$$= 0,005 = \alpha \rightarrow \alpha/2 = 0,0025$$

Từ đó độ tin cậy của ước lượng là

$$1 - 2 \,.\, 0,0025 = 0,9975$$

Ngược lại nếu muốn độ tin cậy của ước lượng là 0,95 thì $\alpha = 0,05$

$$\Rightarrow u_\alpha = 1,645 \Rightarrow X_c = 1,645 \,.\, 2,5 + 12,5 = 16,6 \approx 17$$

Vậy $12 \le m_d \le 19$

Các ký hiệu và công thức cơ bản

* Ước lượng không chệch

$$E(\theta) = \theta$$

* Ước lượng vững

$$\lim_{n\to\infty} P\left(|\hat{\theta} - \theta| < \varepsilon\right) = 1$$

* Ước lượng hiệu quả nhất: Là ước lượng không chệch $\hat{\theta}$ có $V(\hat{\theta})$ đạt min.

* Bất đẳng thức Cramer - Rao:

$$V(\hat{\theta}) \geq \frac{1}{nE\left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right]^2}$$

* Hàm hợp lý

$$L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

* Khoảng tin cậy của μ:

+ Trường hợp đã biết σ^2

$$P\left[\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha_2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha_1}\right] = 1 - \alpha$$

Độ dài khoảng tin cậy $I = \frac{2\sigma}{\sqrt{n}} u_{\alpha/2}$

Kích thước mẫu n sao cho $I \leq I_0; n \geq \frac{\sigma^2}{\varepsilon_0^2}(u_{\alpha/2})^2$

+ Trường hợp chưa biết σ^2

$$P\left[\overline{X} - \frac{S}{\sqrt{n}} t_{\alpha_2}^{(n-1)} < \mu < \overline{X} + \frac{S}{\sqrt{n}} t_{\alpha_1}^{(n-1)}\right] = 1 - \alpha$$

Độ dài khoảng tin cậy: $I = \frac{2S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}$

Kích thước mẫu n sao cho $I \le I_0 : n \ge \frac{S^2}{\varepsilon_0^2} (t_{\alpha/2}^{(m-1)})^2$

* Khoảng tin cậy của σ^2

+ Trường hợp đã biết μ:

$$P\left[\frac{nS^{*2}}{\chi_{\alpha_2}^{2(n)}} < \sigma^2 < \frac{nS^{*2}}{\chi_{1-\alpha_1}^{2(n)}}\right] = 1 - \alpha$$

+ Trường hợp chưa biết μ:

$$P\left[\frac{(n-1)S^2}{\chi_{\alpha_2}^{2(n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha_1}^{2(n-1)}}\right] = 1 - \alpha$$

* Khoảng tin cậy của p

+ Trường hợp $n < 100$; $P(p_1 < p < p_2)$

Trong đó

$$p_1, p_2 = \frac{2nf + u_{\alpha/2}^2 \pm u_{\alpha/2}\sqrt{4nf(1-f) + u_{\alpha/2}^2}}{2(n + u_{\alpha/2}^2)}$$

+ Trường hợp $n \ge 100$

$$P\left[f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha_2} < p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha_1}\right] = 1 - \alpha$$

Độ dài khoảng tin cậy

$$I = \frac{2\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2}$$

Kích thước mẫu n sao cho $I \le I_0 : n \ge \frac{f(1-f)}{\varepsilon_0^2}(u_{\alpha/2})^2$

* Khoảng tin cậy cho hiệu $\mu_1 - \mu_2$

+ Trường hợp đã biết σ_1^2 và σ_2^2

$$P\left[(\overline{X}_1 - \overline{X}_2) - u_{\alpha_2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + u_{\alpha_1}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

+ Trường hợp chưa biết σ_1^2 và σ_2^2 song giả thiết $\sigma_1^2 = \sigma_2^2$:

$$P\left[(\overline{X}_1 - \overline{X}_2) + t_{\alpha_2}^{(n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{\alpha_1}^{(n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha$$

trong đó $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}$

+ Trường hợp chưa biết σ_1^2 và σ_2^2 song giả thiết $\sigma_1^2 \neq \sigma_2^2$

$$P\left[(\overline{X}_1 - \overline{X}_2) + t_{\alpha_2}^{(k)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{\alpha_1}^{(k)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right] = 1 - \alpha$$

+ Trường hợp hai mẫu phụ thuộc theo cặp

$$P\left[\overline{D}-\frac{S_D}{\sqrt{n}}t_{\alpha_2}^{(n-1)}<\mu_D<\overline{D}+\frac{S_D}{\sqrt{n}}t_{\alpha_1}^{(n-1)}\right]=1-\alpha$$

* Khoảng tin cậy cho hiệu $p_1 - p_2$

$$P\left[(f_1-f_2)-S_f.u_{\alpha_2}<p_1-p_2<(f_1-f_2)+S_f.u_{\alpha_1}\right]=1-\alpha$$

trong đó $S_f=\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}}$

* Khoảng tin cậy cho tỷ số $\frac{\sigma_1^2}{\sigma_2^2}$

$$P\left[\frac{S_1^2}{S_2^2}f_{1-\alpha_2}^{(n_2-1,n_1-1)}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{S_1^2}{S_2^2}f_{\alpha_1}^{(n_2-1,n_1-1)}\right]=1-\alpha$$

Câu hỏi ôn tập

1. Để ước lượng tham số θ của tổng thể, tại sao người ta dùng thống kê $\hat{\theta}$ của mẫu thỏa mãn các điều kiện

a. Không chệch b. Vững

c. Hiệu quả nhất d. Đủ

2. Tại sao trung vị lại là một ước lượng kém hiệu quả hơn trung bình?

3. Có sự khác biệt nào trong kết luận của phương pháp hàm ước lượng và phương pháp ước lượng hợp lý tối đa?

4. Một nhân viên chọn ngẫu nhiên một mẫu n = 12 hóa đơn trong số các hóa đơn bán hàng của công ty và thu được các giá trị sau (đơn vị: Ngàn đồng): 875, 1231, 453, 522, 2130, 1550, 390, 760, 498, 999, 1320, 1021. Hãy tìm một ước lượng của giá trị trung bình của các hóa đơn bán hàng và ước lượng của phương sai của các giá trị của hóa đơn bán hàng.

5. Ở câu hỏi 4, giả sử nhân viên muốn ước lượng tỷ lệ các hóa đơn có giá trị trên 1000. Hãy tìm ước lượng của tham số đó.

6. Một nhân viên nghiên cứu thị trường phỏng vấn ngẫu nhiên 18 người xem họ cố (C) hay không (K) tiêu dùng một loại sản phẩm và thu được dãy câu trả lời như sau: CKKCCCKCKCCCKCKCCK. Hãy ước lượng tỷ lệ khách hàng tiêu dùng sản phẩm đó.

7. Cho biến ngẫu nhiên X phân phối chuẩn với $\sigma^2 = 1$. Hãy tìm mối liên hệ giữa độ tin cậy, độ chính xác của ước lượng và kích thước mẫu trong trường hợp ước lượng μ bằng khoảng tin cậy đối xứng.

8. Với độ tin cậy 0,95 người ta ước lượng được tham số θ của tổng thể nằm trong khoảng từ 62 đến 69. Từ đó có thể kết luận được rằng xác suất để θ nằm trong khoảng (62; 69) bằng 0,95 được không? Tại sao?

9. Một chủ cửa hàng muốn ước lượng lợi nhuận trung bình hàng ngày của cửa hàng đó. Người đó lấy mẫu n = 60 ngày để ước lượng song không chắc lợi nhuận hàng ngày có phân phối chuẩn hay không do đó cho rằng kết quả ước lượng là không chính xác. Hãy bình luận về điều đó.

10. Giám đốc xí nghiệp giao cho nhân viên A ước lượng bằng khoảng tin cậy 95% tỷ lệ phế phẩm của xí nghiệp với một mẫu sản phẩm kích thước n = 100. Sau đó đề nghị nhân viên B ước lượng lại với một mẫu n = 100 sản phẩm khiác và thu được kết quả khác. Giám đốc cho rằng kết quả ước lượng của một trong hai nhân viên là sai. Vậy hai nhân viên A và B phải làm gì để thuyết phục giám đốc?

11. Một chủ cửa hàng muốn ước lượng bằng khoảng tin cậy 95% số khách hàng trung bình vào cửa hàng của ông ta mỗi ngày. Song khoảng tin cậy thu được quá rộng nên mất ý nghĩa. Nếu chủ cửa hàng không muốn thay đổi độ tin cậy của ước lượng thì phải làm gì?

Chương VIII

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

§1. KHÁI NIỆM CHUNG

Ở chương VII ta đã nghiên cứu các tham số đặc trưng của tổng thể trên cơ sở thông tin của mẫu bằng phương pháp ước lượng. Chương này tiếp tục nghiên cứu dấu hiệu của tổng thể bằng một phương pháp khác là kiểm định giả thuyết thống kê. Với những thông tin bổ sung phương pháp này cho phép giải quyết nhiều bài toán đa dạng hơn liên quan đến dấu hiệu nghiên cứu trong tổng-thể.

1.1. Giả thuyết thống kê

Giả sử dấu hiệu nghiên cứu trong tổng thể có thể xem như biến ngẫu nhiên X. Nếu chưa biết dạng phân phối xác suất của nó, song có cơ sở để giả thiết rằng X phân phối theo một quy luật A nào đó, người ta đưa ra giả thuyết: Biến ngẫu nhiên X phân phối theo quy luật A.

Cũng có trường hợp dạng phân phối xác suất của X đã biết song tham số đặc trưng của nó lại chưa biết, nếu có cơ sở

để giả thiết rằng giá trị của tham số bằng θ_0 (θ_0 là hằng số đã biết), người ta đưa ra giả thuyết: $\theta = \theta_0$.

Khi nghiên cứu hai hay nhiều biến ngẫu nhiên thuộc các tổng thể khác nhau hay thuộc cùng một tổng thể thường phải xét xem chúng độc lập hay phụ thuộc nhau, các tham số đặc trưng của chúng có bằng nhau hay không. Nếu chưa biết một cách chắc chắn song có cơ sở để nhận định về các vấn đề đó cũng có thể đưa ra các giả thuyết tương ứng.

Từ đó có thể đưa ra định nghĩa sau:

Định nghĩa. *Giả thuyết thống kê là giả thuyết về dạng phân phối xác suất của biến ngẫu nhiên, về các tham số đặc trưng của biến ngẫu nhiên hoặc về tính độc lập của các biến ngẫu nhiên.*

Giả thuyết thống kê đưa ra được ký hiệu là H_0 và được gọi là *giả thuyết gốc*.

Khi đưa ra một giả thuyết thống kê, người ta còn nghiên cứu kèm theo nó mệnh đề mâu thuẫn với nó, gọi là *giả thuyết đối* và ký hiệu là H_1 để khi giả thuyết H_0 bị bác bỏ thì thừa nhận giả thuyết H_1. H_0 và H_1 tạo nên cặp giả thuyết thống kê.

Chẳng hạn ta nghiên cứu nhu cầu thị trường về một loại hàng hoá nào đó. Ta có thể đưa ra các cặp giả thuyết thống kê sau:

* H_0: Nhu cầu X của thị trường phân phối theo quy luật chuẩn; H_1: Nhu cầu X của thị trường không phân phối theo quy luật chuẩn.

* H_0: Nhu cầu trung bình về loại hàng hóa này là $\mu = 1000$ đơn vị/tháng, lúc đó các giả thuyết đối tương ứng với nó có thể là H_1: $\mu > 1000$; H_1: $\mu < 1000$ hoặc H_1: $\mu \neq 1000$.

* H_0: Nhu cầu X của thị trường và thu nhập Y của khách hàng độc lập nhau; H_1: X và Y phụ thuộc nhau...

Trong thực tế người ta còn phân biệt các giả thuyết chứa đựng một mệnh đề hoặc nhiều mệnh đề.

Giả thuyết *đơn* là giả thuyết chỉ chứa đựng một mệnh đề. Chẳng hạn nếu μ là tham số của quy luật chuẩn và ta đưa ra giả thuyết H_0: $\mu = 5$ thì đó là giả thuyết đơn.

Giả thuyết *kép* là giả thuyết chứa đựng một số hữu hạn hoặc vô hạn các giả thuyết đơn. Chẳng hạn giả thuyết H_0: $\mu > 5$ bao gồm một số vô hạn các giả thuyết đơn dạng H_0: $\mu = b_i$ trong đó b_i là mọi số lớn hơn 5.

Việc kiểm định một giả thuyết kép thường khá phức tạp, do đó ở đây ta chỉ hạn chế ở việc nghiên cứu trường hợp giả thuyết gốc là giả thuyết đơn.

Vì các giả thuyết thống kê có thể đúng hoặc sai nên cần kiểm định, tức là tìm ra kết luận về tính thừa nhận được hay không thừa nhận được của giả thuyết đó. Việc kiểm định này gọi là *kiểm định thống kê* vì nó dựa vào thông tin thực nghiệm của mẫu để kết luận.

Phương pháp chung để kiểm định một giả thuyết thống kê như sau: Trước hết giả sử H_0 đúng và từ đó dựa vào thông tin của mẫu rút ra từ tổng thể tìm được một biến cố A nào đó sao cho xác suất xảy ra biến cố A bằng α bé đến mức có thể sử dụng nguyên lý xác suất nhỏ tức là có thể coi A không xảy ra trong một phép thử về biến cố này. Lúc đó trên một mẫu cụ thể thực hiện một phép thử đối với biến cố A, nếu A xảy ra thì điều đó chứng tỏ H_0 sai và bác bỏ nó, còn nếu A không xảy ra thì ta chưa có cơ sở để bác bỏ H_0.

Để cụ thể hóa phương pháp trên ta nghiên cứu một số khái niệm sau:

1.2. Tiêu chuẩn kiểm định giả thuyết thống kê

Từ biến ngẫu nhiên gốc X trong tổng thể lập mẫu ngẫu nhiên kích thước n

$$W = (X_1, X_2, ..., X_n)$$

và chọn lập thống kê

$$G = f(X_1, X_2, ..., X_n, \theta_0)$$

trong đó θ_0 là tham số liên quan đến giả thuyết cần kiểm định. Điều kiện đặt ra đối với thống kê G là nếu H_0 đúng thì quy luật phân phối xác suất của G hoàn toàn xác định. Thống kê G được gọi là tiêu chuẩn kiểm định.

1.3. Miền bác bỏ giả thuyết

Sau khi đã chọn được tiêu chuẩn kiểm định G, do quy luật phân phối xác suất của G đã biết nên với một xác suất khá bé bằng α cho trước (thường α được lấy bằng 0,05 hoặc 0,01) có thể tìm được miền W_α tương ứng sao cho với điều kiện giả thuyết H_0 đúng xác suất để G nhận giá trị thuộc miền W_α bằng α. Điều kiện này được viết như sau:

$$P(G \in W_\alpha / H_0) = \alpha \qquad (8.1)$$

Biến cố $(G \in W_\alpha)$ đóng vai trò như biến cố A nói trên và vì α khá bé nên theo nguyên lý xác suất nhỏ có thể coi như nó không xảy ra trong một phép thử.

Giá trị α được gọi là *mức ý nghĩa* của kiểm định và miền W_α được gọi là *miền bác bỏ* giả thuyết H_0 với mức ý nghĩa α. Hiển nhiên với một mức ý nghĩa α cho trước có thể tìm được vô số miền bác bỏ tương ứng.

Lúc đó miền giá trị còn lại của G, ký hiệu là $\overline{W}_\alpha$ được gọi là miền không bác bỏ giả thuyết (đôi khi để cho đơn giản người ta gọi là miền thừa nhận giả thuyết). Điểm giới hạn phân chia miền bác bỏ và miền không bác bỏ được gọi là giá trị tới hạn.

1.4. Giá trị quan sát của tiêu chuẩn kiểm định

Thực hiện một phép thử đối với mẫu ngẫu nhiên $W = (X_1, X_2,..., X_n)$ thu được một mẫu cụ thể $w = (x_1, x_2,..., x_n)$ và qua đó tính được một giá trị cụ thể của tiêu chuẩn kiểm định G:

$$G_{qs} = f(x_1, x_2,..., x_n, \theta_\alpha)$$

Giá trị này được gọi là giá trị quan sát của tiêu chuẩn kiểm định.

1.5. Quy tắc kiểm định giả thuyết thống kê

Sau khi đã tính được giá trị quan sát G_{qs} của tiêu chuẩn kiểm định, ta so sánh giá trị này với miền bác bỏ W_α và kết luận theo quy tắc sau:

1. Nếu giá trị quan sát của tiêu chuẩn kiểm định thuộc miền bác bỏ $G_{qs} \in W_\alpha$ thì điều đó có thể giải thích rằng H_0 sai và do đó ta bác bỏ H_0, thừa nhận H_1.

2. Nếu giá trị quan sát của tiêu chuẩn kiểm định không thuộc miền bác bỏ $G_{qs} \notin W_\alpha$ thì điều đó chưa khẳng định rằng H_0 đúng mà chỉ có nghĩa là qua mẫu cụ thể này chưa khẳng định được là H_0 sai. Do đó ta chỉ có thể nói: Qua mẫu cụ thể này chưa có cơ sở để bác bỏ H_0 (trên thực tế là vẫn thừa nhận H_0).

1.6. Sai lầm loại một và sai lầm loại hai

Với quy tắc kiểm định như trên có thể mắc hai loại sai lầm:

1. Sai lầm loại 1: Bác bỏ giả thuyết H_0 trong khi H_0

đúng. Ta thấy xác suất mắc phải loại sai lầm này đúng bằng mức ý nghĩa α. Thật vậy, mặc dù H_0 đúng thì xác suất để ($G \in W_\alpha$) vẫn bằng α. Nhưng nếu $G \in W_\alpha$ thì ta lập tức bác bỏ H_0. Như vậy, ta có thể mắc sai lầm loại 1 với xác suất bằng α. Sai lầm này có thể sinh ra do kích thước mẫu quá nhỏ, do phương pháp lấy mẫu v.v...

2. Sai lầm loại 2: Thừa nhận giả thuyết H_0 trong khi H_0 sai, hay giá trị quan sát G_{qs} không thuộc miền bác bỏ W_α trong khi H_1 đúng.

Giả sử xác suất mắc sai lầm loại 2 là β:

$$P(G \notin W_\alpha/H_1) = \beta \qquad (8.2)$$

Lúc đó biến cố không mắc sai lầm loại 2 là biến cố để G thuộc miền bác bỏ và do đó ta bác bỏ H_0 trong khi H_1 đúng:

$$G \in W_\alpha/H_1$$

Biến cố này đối lập với biến cố ($G \notin W_\alpha/H_1$) nên xác suất của nó là:

$$P(G \in W_\alpha/H_1) = 1 - \beta \qquad (8.3)$$

xác suất 1 – β được gọi là *lực kiểm định.*

Quan hệ giữa việc kiểm định giả thuyết và các loại sai lầm có thể mô tả trong bảng sau: (Bảng 8.1)

Bảng 8.1

Tình huống / Quyết định	H_0 đúng	H_0 sai
Bác bỏ H_0	Sai lầm loại 1 xác suất = α	Quyết định đúng xác suất = 1 – β
Không bác bỏ H_0	Quyết định đúng xác suất = 1 – α	Sai lầm loại 2 xác suất = β

Ta thấy rằng sai lầm loại 1 và sai lầm loại 2 mâu thuẫn nhau, tức là với một mẫu kích thước n xác định thì không thể cùng một lúc giảm xác suất mắc hai loại sai lầm nói trên được. Khi ta giảm α đi thì đồng thời sẽ làm tăng β và ngược lại. Chẳng hạn nếu lấy $\alpha = 0$ thì sẽ không bác bỏ bất kỳ giả thuyết nào, kể cả giả thuyết sai, như vậy β sẽ đạt cực đại.

Trong thực tế người ta tiến hành như sau: Sau khi đã ấn định một mức ý nghĩa α và với mẫu kích thước n xác định thì trong vô số các miền bác bỏ W_α tương ứng có thể tìm được, ta chọn ra miền bác bỏ W_α sao cho xác suất mắc sai lầm loại 2 là nhỏ nhất hay lực kiểm định là lớn nhất.

Như vậy, cần tìm miền bác bỏ W_α thỏa mãn các điều kiện sau:

$$P(G \in W_\alpha / H_0) = \alpha \text{ cho trước}$$

và $$P(G \in W_\alpha / H_1) = 1 - \beta \rightarrow \max.$$

Dựa vào định lý Neyman - Pearson được trình bày trong các tài liệu đầy đủ hơn có thể tìm được những miền bác bỏ "tốt nhất" như vậy.

Những miền bác bỏ được xây dựng trong những phần dưới đây đều thỏa mãn điều kiện nêu trên, tức là đều là những miền bác bỏ "tốt nhất" với mức ý nghĩa và kích thước mẫu xác định trước.

Tuy nhiên, vẫn còn lại vấn đề là lựa chọn mức ý nghĩa bằng bao nhiêu?

Điều này tùy thuộc vào từng trường hợp cụ thể và căn cứ vào "hậu quả" mà sai lầm loại 1 và sai lầm loại 2 mang lại.

1.7. Thủ tục kiểm định giả thuyết thống kê

Qua nội dung được trình bày ở phần trên có thể xây dựng

một thủ tục để kiểm định giả thuyết thống kê. Trên thực tế việc kiểm định giả thuyết thống kê có thể tiến hành theo hai thủ tục khác nhau.

1. Kiểm định với giá trị cho trước của α

Khi chỉ kiểm soát khả năng mắc sai lầm loại 1 thì thủ tục kiểm định được tiến hành như sau:

a) Xây dựng giả thuyết gốc H_0 cần kiểm định.

b) Từ tổng thể nghiên cứu lập mẫu ngẫu nhiên kích thước n.

c) Chọn tiêu chuẩn kiểm định G wà xác định quy luật phân phối xác suất của nó với điều kiện giả thuyết H_0 là đúng.

d) Với mức ý nghĩa α cho trước xác định miền bác bỏ tốt nhất tùy thuộc vào giả thuyết đối H_1.

e) Lập mẫu cụ thể và tìm giá trị của tiêu chuẩn kiểm định trên mẫu.

g) So sánh giá trị quan sát của tiêu chuẩn kiểm định với miền bác bỏ và kết luận.

h) Đánh giá xác suất mắc sai lầm loại 2 theo các giá trị khác nhau của H_1.

2. Kiểm định với giá trị cho trước của α và β

Khi kiểm soát cả khả năng mắc sai lầm loại 1 và loại 2 thì thủ tục kiểm định được tiến hành như sau:

a) Xây dựng giả thuyết H_0 cần kiểm định.

b) Chọn tiêu chuẩn kiểm định G wà xác định quy luật phân phối xác suất của nó với điều kiện giả thuyết H_0 là đúng.

c) Với α và β cho trước xác định kích thước mẫu cần điều tra để việc kiểm định phạm hai sai lầm trên với xác suất không vượt quá mức cho trước.

d) Theo kết quả ở phần c, điều tra một mẫu cụ thể và tiến hành tiếp như ở trường hợp trước.

Sau đây ta sẽ vận dụng các thủ tục kiểm định trên vào một số giả thuyết thống kê thông dụng hơn cả trong nghiên cứu kinh tế - xã hội.

§2. KIỂM ĐỊNH THAM SỐ

2.1. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật chuẩn khi đã biết phương sai

Giả sử biến ngẫu nhiên gốc X trong tổng thể phân phối theo quy luật chuẩn $N(\mu, \sigma^2)$ với phương sai đã biết nhưng chưa biết kỳ vọng toán μ. Nếu có cơ sở để giả thiết rằng giá trị của nó bằng μ_0 ta đưa ra giả thuyết thống kê H_0: $\mu = \mu_0$. Để kiểm định giả thuyết trên từ tổng thể lập mẫu kích thước n:

$$W = (X_1, X_2, \ldots, X_n)$$

Vì đã biết phương sai σ^2 của biến ngẫu nhiên gốc X trong tổng thể nên tiêu chuẩn kiểm định được chọn là thống kê.

$$G = U = \frac{\overline{X} - \mu_0}{SE(\overline{X})} = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma} \qquad (8.4)$$

Nếu giả thuyết H_0 đúng thì ta có:

$$U = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

và từ mục §6 Chương VI ta có U phân phối N(0,1).

1. Kiểm định với giá trị cho trước của α. Nếu cho trước mức ý nghĩa α thì tùy thuộc vào dạng của giả thuyết đối H_1 miền bác bỏ "tốt nhất" được xây dựng theo các trường hợp sau:

a) H_0: $\mu = \mu_0$; H_1: $\mu > \mu_0$. Lúc đó với α cho trước có thể tìm giá trị tới hạn chuẩn u_α sao cho

$$P(G \in W_\alpha / H_0) = P(U > u_\alpha) = \alpha$$

Ta thu được miền bác bỏ bên phải W_α được xác định bằng biểu thức:

$$W_\alpha = \left\{ U = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma}; U > u_\alpha \right\} \qquad (8.5)$$

b) H_0: $\mu = \mu_0$; H_1: $\mu < \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn chuẩn $u_{1-\alpha}$ sao cho

$$P(G \in W_\alpha / H_0) = P(U < u_{1-\alpha}) = P(U < -u_\alpha) = \alpha$$

Ta thu được miền bác bỏ bên trái W_α được xác định bằng biểu thức:

$$W_\alpha = \left\{ U = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma}; U < -u_\alpha \right\} \qquad (8.6)$$

c) H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được hai giá trị tới hạn chuẩn là $u_{1-\alpha}/2$ và $u_\alpha/2$ sao cho

$$P(G \in W_\alpha / H_0) = P(U < u_{1-\alpha/2}) + P(U > u_{\alpha/2})$$
$$= P(U < -u_{\alpha/2}) + P(U > u_{\alpha/2})$$
$$= P(|U| > u_{\alpha/2}) = \alpha$$

Ta thu được miền bác bỏ hai phía được xác định bằng biểu thức:

$$W_\alpha = \left\{ U = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma}; |U| > u_{\alpha/2} \right\} \quad (8.7)$$

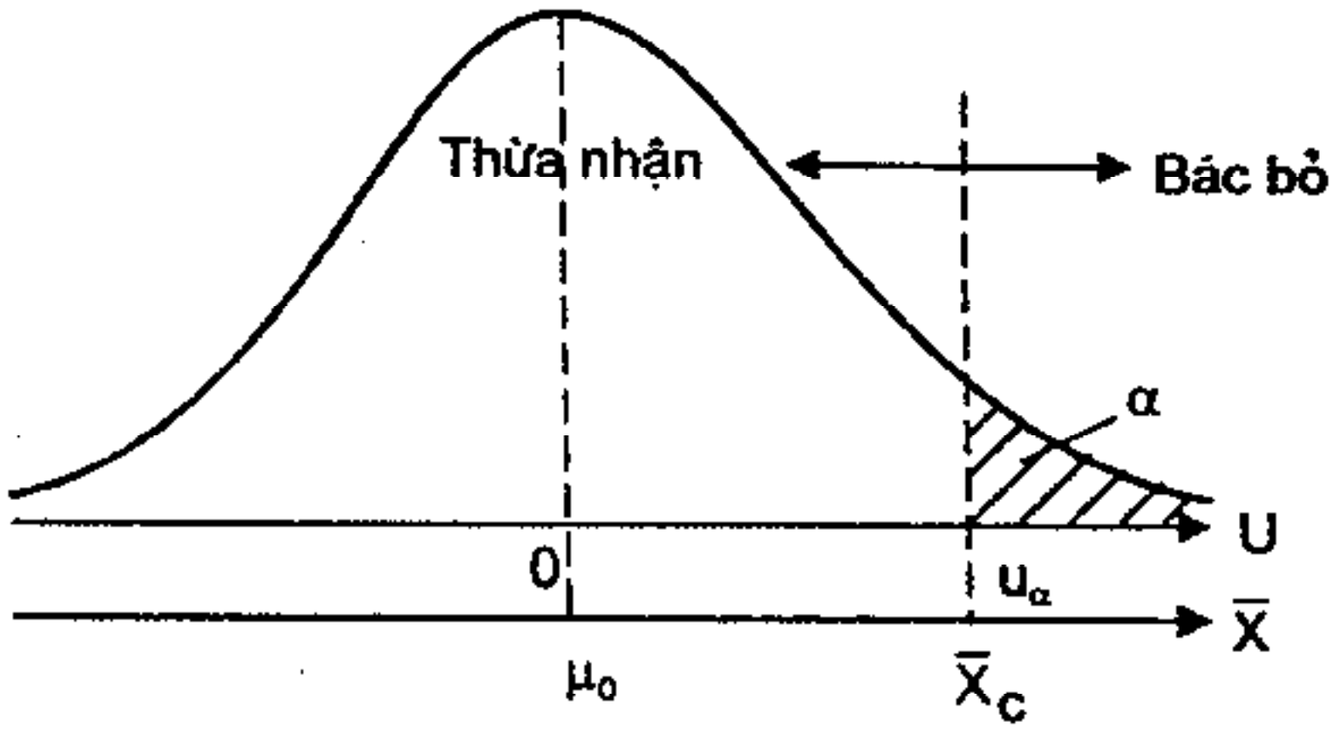
Lập mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ và tính giá trị quan sát tiêu chuẩn kiểm định

$$U_{qs} = \frac{(\overline{x} - \mu_0)\sqrt{n}}{\sigma}$$

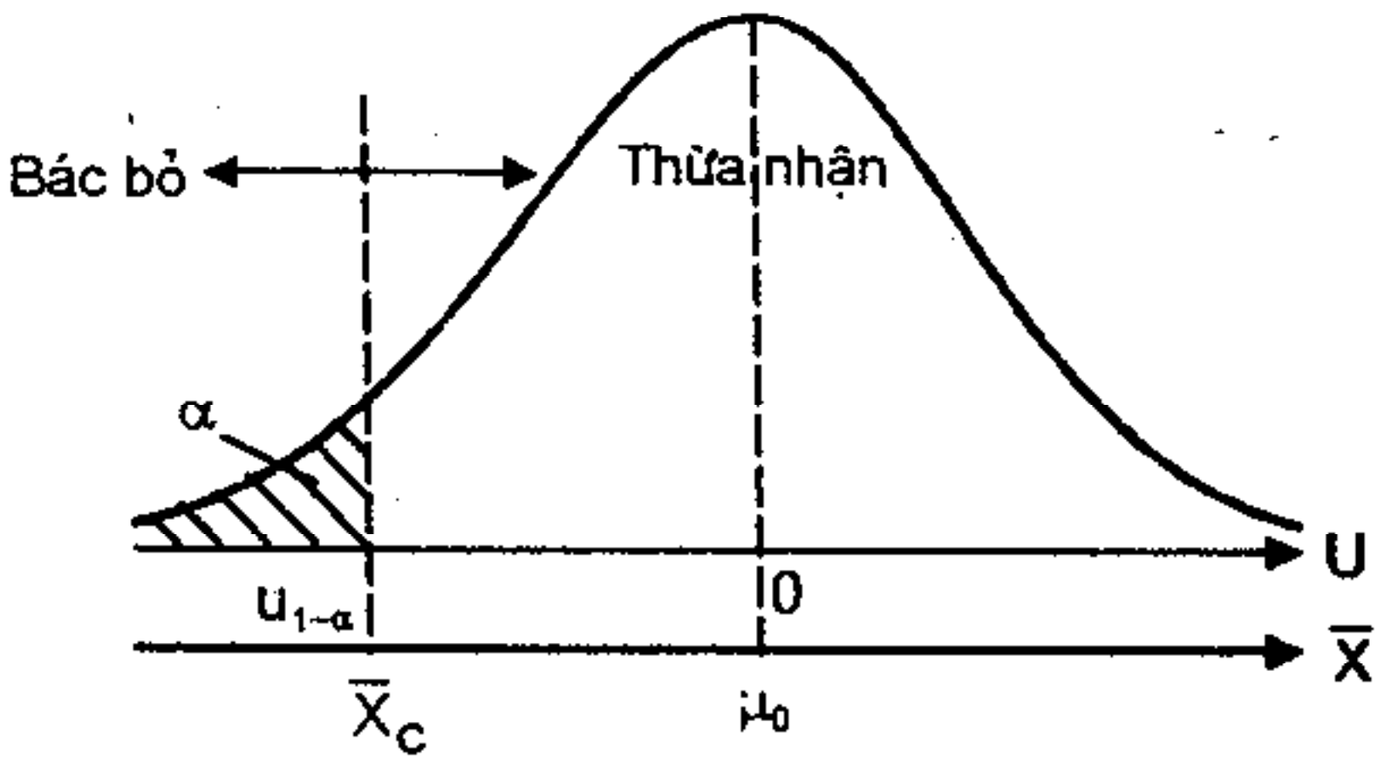
và so sánh với W_α để kết luận:

- Nếu $U_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1:
- Nếu $U_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

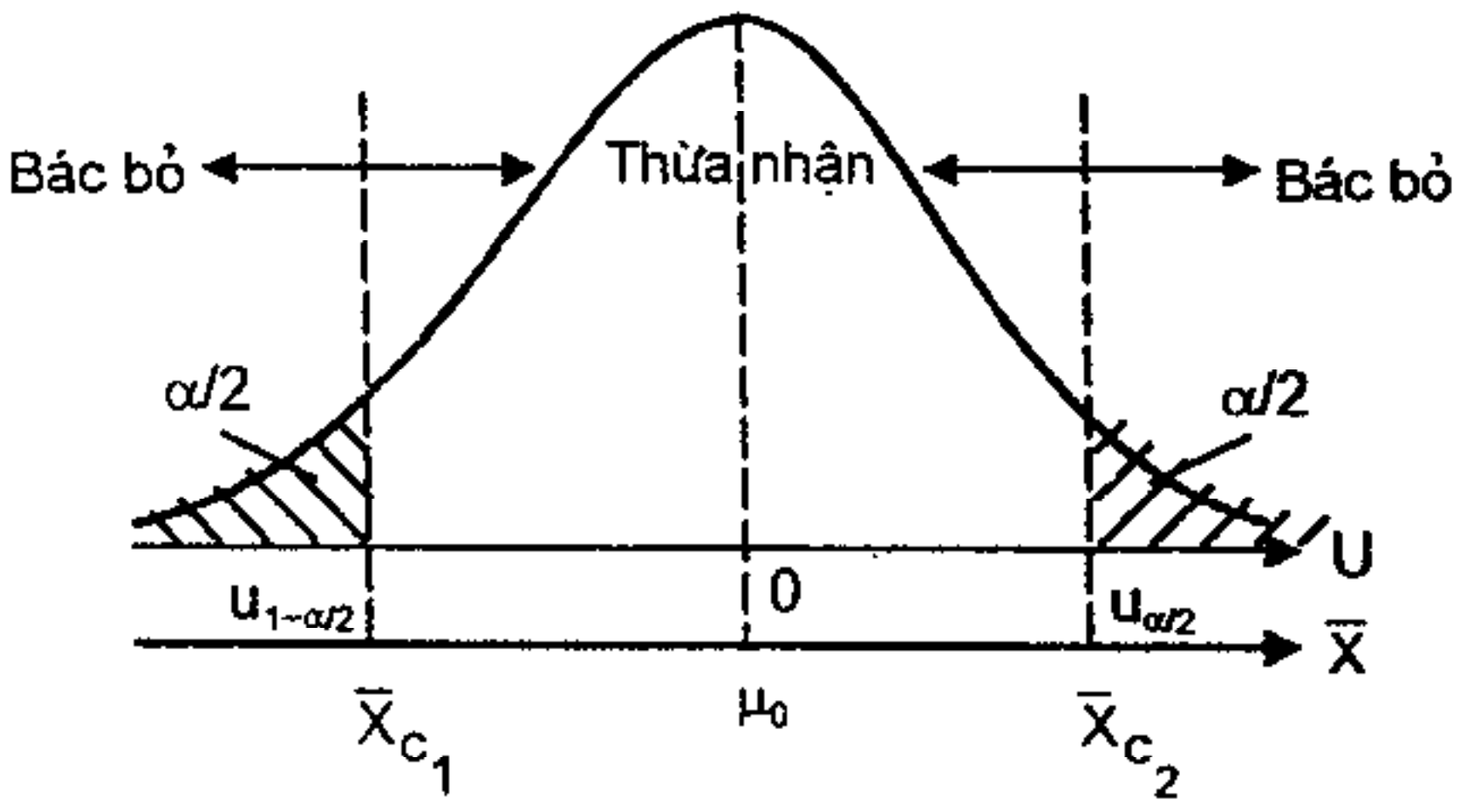
Các miền bác bỏ được xây dựng theo các công thức (8.5) (8.6) và (8.7) có thể mô tả trên đồ thị như sau (H.8.1) trong đó $\overline{x}_C$ là giá trị tới hạn tương ứng của $\overline{x}$ khi $U = u_\alpha$.



Hình 8.1a. **Miền bác bỏ bên phải**



Hình 8.1.b. **Miền bác bỏ bên trái**



Hình 8.1.c. **Miền bác bỏ hai phía**

Việc kiểm định bằng cách sử dụng tiêu chuẩn (8.4) thường được gọi là kiểm định U (U-test).

Thí dụ 1. Trong năm trước trọng lượng trung bình trước khi xuất chuồng của bò ở một trại chăn nuôi là 380 kg. Năm nay người ta áp dụng thử một chế độ chăn nuôi mới với hy vọng là bò sẽ tăng trọng nhanh hơn. Sau thời gian áp dụng thử người ta lấy ngẫu nhiên 50 con bò trước khi xuất chuồng đem cân và tính được trọng lượng trung bình của chúng là 390 kg. Vậy với mức ý nghĩa $\alpha = 0,01$ có thể cho rằng trọng lượng trung bình của bò trước khi xuất chuồng đã tăng lên hay không? Giả thiết trọng lượng của bò là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn là 35,2 kg.

Giải. Gọi X là trọng lượng của bò trước khi xuất chuồng. Theo giả thiết X phân phối chuẩn với $\sigma = 35,2$. Vậy trọng lượng xuất chuồng trung bình là μ. Đây là bài toán kiểm định giá trị của tham số μ của biến ngẫu nhiên phân phối chuẩn khi đã biết phương sai của tổng thể. Vậy từ (8.5) ta có:

Cặp giả thuyết thống kê có dạng:

$$H_0: \mu = 380; \; H_1: \mu > 380$$

Tiêu chuẩn kiểm định

$$U = \frac{(\overline{X} - 380)\sqrt{50}}{35,2}$$

trong đó $\overline{X}$ là trung bình mẫu ngẫu nhiên kích thước $n = 50$.

Cũng từ (8.5), với $\alpha = 0,01$ ta có:

$$u_\alpha = u_{0,01} = 2,33$$

Vậy miền bác bỏ bên phải là $(2,33; +\infty)$.

Từ mẫu cụ thể ta có $\bar{x} = 390$. Vậy giá trị quan sát của tiêu chuẩn kiểm định.

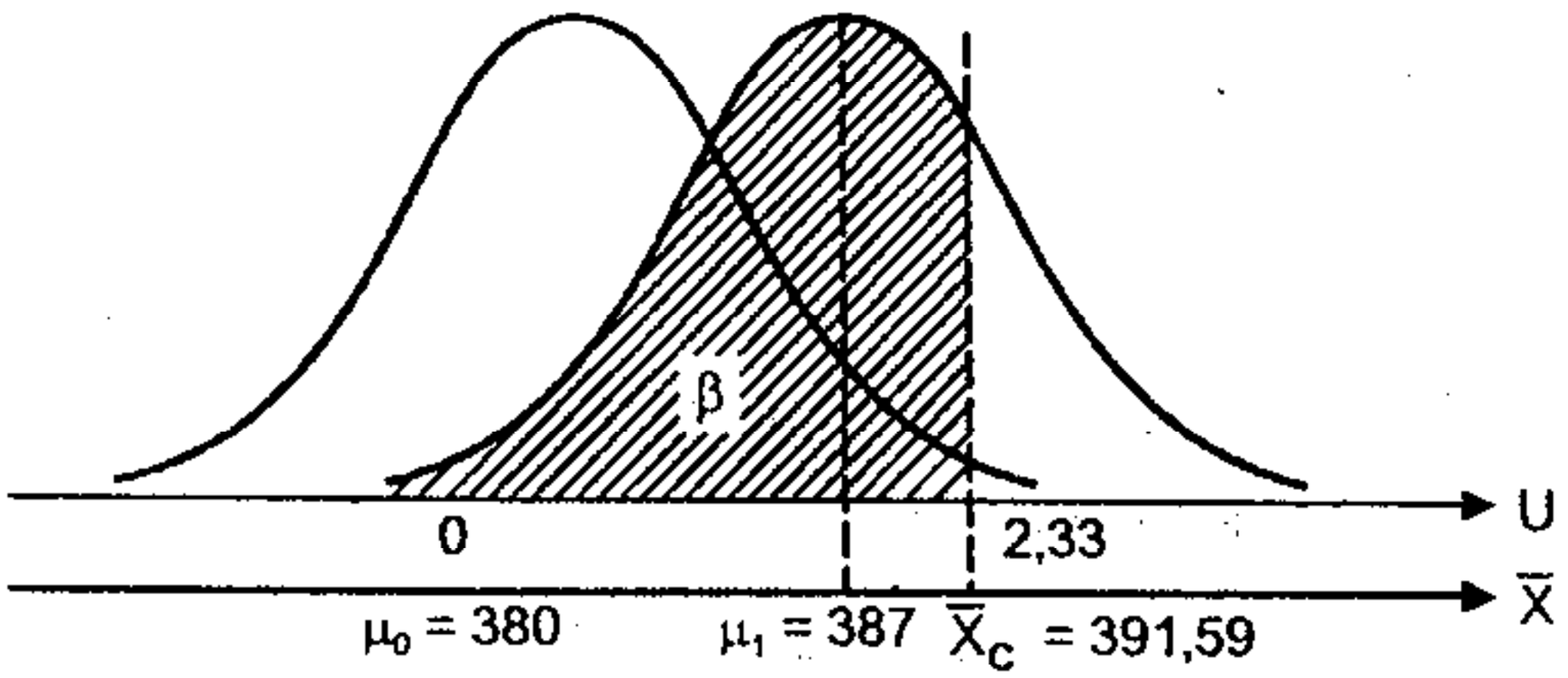
$$U_{qs} = \frac{(390 - 380)\sqrt{50}}{35,2} = 2,01$$

Như vậy $U_{qs} \notin W_{\alpha}$.

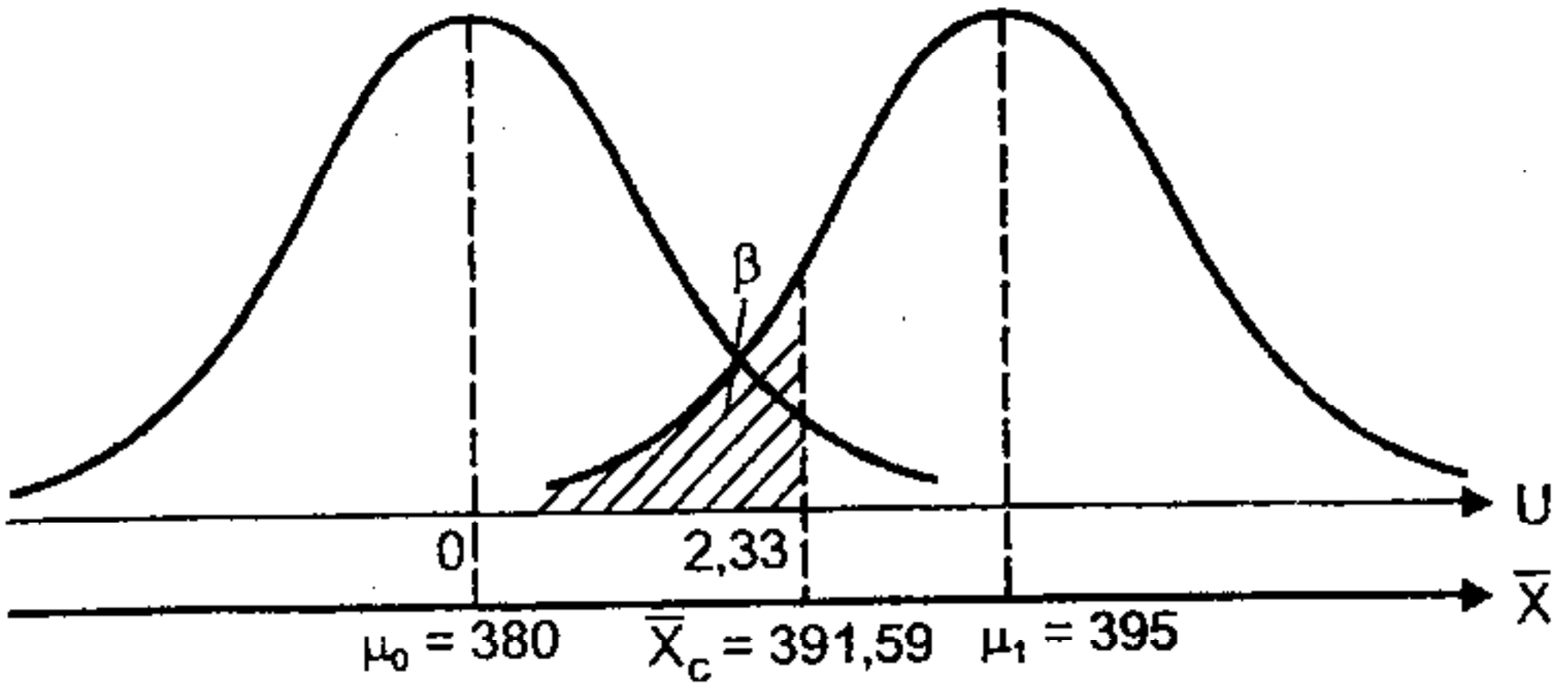
Kết luận: Với mức ý nghĩa $\alpha = 0,01$, qua mẫu cụ thể đã cho ta chưa có cơ sở để bác bỏ H_0 (trên thực tế là vẫn chấp nhận H_0: $\mu = 380$). Kết quả trên cũng cho thấy trung bình mẫu thu được qua mẫu cụ thể đã cho không khác biệt một cách có ý nghĩa so với trung bình tổng thể.

2. Tìm β

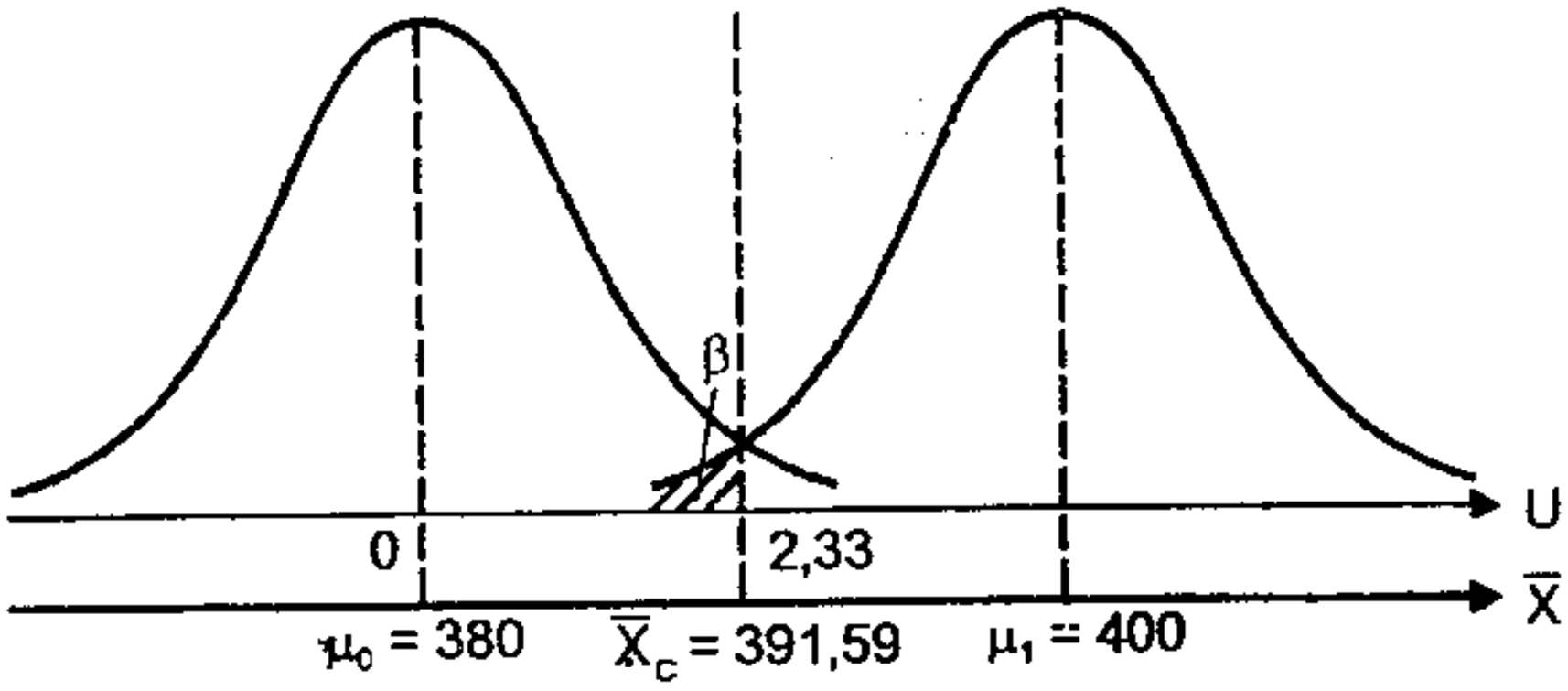
Ta cũng có thể tính tiếp giá trị β là xác suất để mắc sai lầm loại 2. Nếu giả thuyết H_0: $\mu = 380$ thì xác suất để thừa nhận giả thuyết sai H_0 sẽ phụ thuộc vào việc giá trị thực của μ sai lệch nhiều hay ít so với 380. Chẳng hạn nếu giá trị thực của μ là 400 thì β sẽ nhỏ hơn so với trường hợp giá trị thực của μ là 387. Vậy tùy thuộc vào giá trị thực của μ mà ta có các giá trị β khác nhau. Điều đó được minh họa trên hình (8.2) với ba giá trị thực khác nhau của μ là 387, 395 và 400.



Hình 8.2a. **Khi H_1: $\mu = 387$**



Hình 8.2b. **β khi H_1: μ = 395**



Hình 8.2c. **β khi H_1: μ = 400**

Giả sử ta vẫn kiểm định cặp giả thuyết H_0: $\mu = 380$; H_1: $\mu > 380$ và với mức ý nghĩa $\alpha = 0,01$ ta xác định β. Phần gạch chéo trên hình (8.2) biểu diễn giá trị của β vì đó chính là xác suất để $\overline{X}$ rơi vào miền thừa nhận khi giả thuyết H_0 sai và giá trị thực của μ là 387, 395 và 400. Lúc đó lực kiểm định $1 - \beta$ chính là miền giá trị còn lại nằm trong miền bác bỏ H_0.

Nếu ký hiệu như trước đây μ_0 là giá trị giả thuyết của μ và μ_1 là giá trị thực của μ thì với miền bác bỏ bên phải giá trị β được xác định như sau:

Phải kiểm định cặp giả thuyết $H_0: \mu = \mu_0$; $H_1: \mu > \mu_0$.

Do H_0 sai nên biến ngẫu nhiên $\frac{\overline{X} - \mu_0}{Se(\overline{X})}$ không có phân phối N(0,1). Thay vào đó, biến ngẫu nhiên $\frac{\overline{X} - \mu_1}{Se(\overline{X})}$ sẽ phân phối N(0,1).

Vì thế:

$$\beta = P\left(\frac{\overline{X} - \mu_0}{Se(\overline{X})} < u_\alpha\right) = P\left(\frac{\overline{X} - \mu_0}{Se(\overline{X})} - \frac{\mu_1}{Se(\overline{X})} < u_\alpha - \frac{\mu_1}{Se(\overline{X})}\right)$$

$$= P\left(\frac{\overline{X} - \mu_1}{Se(\overline{X})} < u_\alpha - \frac{\mu_1 - \mu_0}{Se(\overline{X})}\right) = P\left(U < u_\alpha - \frac{\mu_1 - \mu_0}{Se(\overline{X})}\right)$$

$$\Rightarrow \beta = P\left(U < u_\alpha - \frac{\mu_1 - \mu_0}{Se(\overline{X})}\right) \quad (8.8)$$

Bằng cách chứng minh tương tự ta thu được biểu thức của β khi miền bác bỏ là bên trái

$$\beta = P\left[U < u_\alpha - \frac{(\mu_0 - \mu_1)}{Se(\overline{X})}\right] \quad (8.9)$$

Từ đó ta có công thức chung để tìm xác suất mắc sai lầm loại hai β khi miền bác bỏ là một phía (bên phải hoặc bên trái) như sau:

$$\beta = P\left[U < u_\alpha - \frac{|\mu_0 - \mu_1|}{Se(\overline{X})}\right] \quad (8.10)$$

Nếu miền bác bỏ là hai phía thì β được xác định bằng công thức:

$$\beta \approx P\left[U < u_{\alpha/2} - \frac{|\mu_0 - \mu_1|}{Se(\overline{X})}\right] \qquad (8.11)$$

Từ đó suy ra giá trị của lực kiểm định $1 - \beta$.

Thí dụ 2. Tiếp tục thí dụ 1, tìm xác suất mắc sai lầm loại 2 và lực kiểm định nếu trọng lượng xuất chuồng trung bình của đàn bò thực sự là 395 kg.

Giải. Vì giả thuyết đối là H_1: $\mu > 380$ nên với $\alpha = 0,01$; $\mu_0 = 380$; $\mu_1 = 395$, theo công thức (8.10) ta có:

$$\beta = P\left[U < u_{0,01} - \frac{|380 - 395|}{\frac{35,2}{\sqrt{50}}}\right] = P[U < 2,33 - 3,01]$$

$$= P[U < -0,68] = P[U > 0,68] = 0,2483$$

Vậy $1 - \beta = 1 - 0,2483 = 0,7517$.

3. Kiểm định với α và β cho trước

Từ công thức (8.10) và (8.11) ta thấy rằng nếu đã cố định trước xác suất mắc sai lầm loại 1 và với giá trị xác định của μ_1 thì xác suất mắc sai lầm loại 2 sẽ phụ thuộc vào kích thước mẫu n. Kích thước mẫu càng lớn thì càng có nhiều thông tin hơn về μ và giá trị của β sẽ càng nhỏ hơn. Ngược lại, nếu giá trị μ_1 đã xác định trước thì có thể xác định kích thước mẫu n sao cho đảm bảo xác suất mắc sai lầm loại 1 và loại 2 tương ứng là α và β cho trước.

Giả sử ta phải kiểm định cặp giả thuyết H_0: $\mu = \mu_0$ và H_1: $\mu > \mu_0$. Hơn nữa giả sử ta muốn xác suất mắc sai lầm loại 1 là α và xác suất mắc sai lầm loại 2 không vượt quá giá trị β với giá trị thực của μ sai lệch so với μ_0 không vượt quá Δ cho trước.

Từ hình (8.2) ta có:

$$u_\alpha + u_\beta = u_\alpha - u_{1-\beta} = \frac{(\overline{X}_c - \mu_0)}{Se(\overline{X})} - \frac{(\overline{X}_c - \mu_1)}{Se(\overline{X})}$$

$$= \frac{(\mu_1 - \mu_0)}{Se(\overline{X})} = \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}$$

suy ra: $n = \dfrac{\sigma^2(u_\alpha + u_\beta)^2}{(\mu_1 - \mu_0)^2}$

Từ đó nếu đòi hỏi $\mu_1 - \mu_0 \le \Delta$ thì ta có:

$$n \ge \left[\frac{\sigma^2(u_\alpha + u_\beta)^2}{\Delta^2}\right] \qquad (8.12)$$

Nếu miền bác bỏ là hai phía thì kích thước mẫu tối thiểu được xác định bằng công thức:

$$n \ge \left[\frac{\sigma^2(u_{\alpha/2} + u_\beta)^2}{\Delta^2}\right] \qquad (8.13)$$

Thí dụ 3. Trở lại thí dụ 1 và với xác suất mắc sai lầm loại 1 cho trước là 0,01, nếu muốn xác suất mắc sai lầm loại 2 không vượt quá 0,05 thì phải đem cân thử tối thiểu là bao nhiêu con bò nếu quả thực trọng lượng trung bình của bò trước khi xuất chuồng là không quá 390 kg.

Giải. Theo công thức (8.12) ta có:

$\alpha = 0,01 \rightarrow u_\alpha = u_{0.01} = 2,33$

$\beta = 0,05 \rightarrow u_\beta = u_{0.05} = 1,645$

$\mu_1 - \mu_0 = 390 - 380 = 10$

Từ đó: $n \geq \left[\frac{(35,2)^2(2,33+1,645)^2}{10^2}\right] = 195,78$

Vậy n = 196 con.

Vậy để tiếp tục kiểm định, trước hết cần đem cân ngẫu nhiên 196 con để thu được trung bình mẫu $\bar{x}$ cho quá trình kiểm định tiếp theo.

4. Giá trị xác suất (P-Value) của kiểm định

Thủ tục kiểm định được trình bày ở trên có tính chất truyền thống và thường được gọi là kiểm định theo cách tiếp cận cổ điển, theo đó ta xác định các bộ phận của một giả thuyết thống kê theo các sai lầm loại 1 và loại 2 tương ứng với xác suất α và β. Trong những năm gần đây nhiều nhà nghiên cứu thường sử dụng một cách tiếp cận khác. Thay vì kiểm định giả thuyết với một giá trị α định trước thì họ cho rằng ta nên định rõ các giả thuyết cơ sở H_0 và giả thuyết đối H_1, sau đó thu thập số liệu mẫu và xác định mức độ khẳng định việc bác bỏ giả thuyết H_0. Mức độ khẳng định này thường được gọi là giá trị P (P-value) của kiểm định.

Ta sẽ minh họa việc tính giá trị P qua thí dụ sau.

Thí dụ 4: Trở lại thí dụ 1.

a) Thay vì định trước giá trị α, hãy xác định giá trị P của kiểm định.

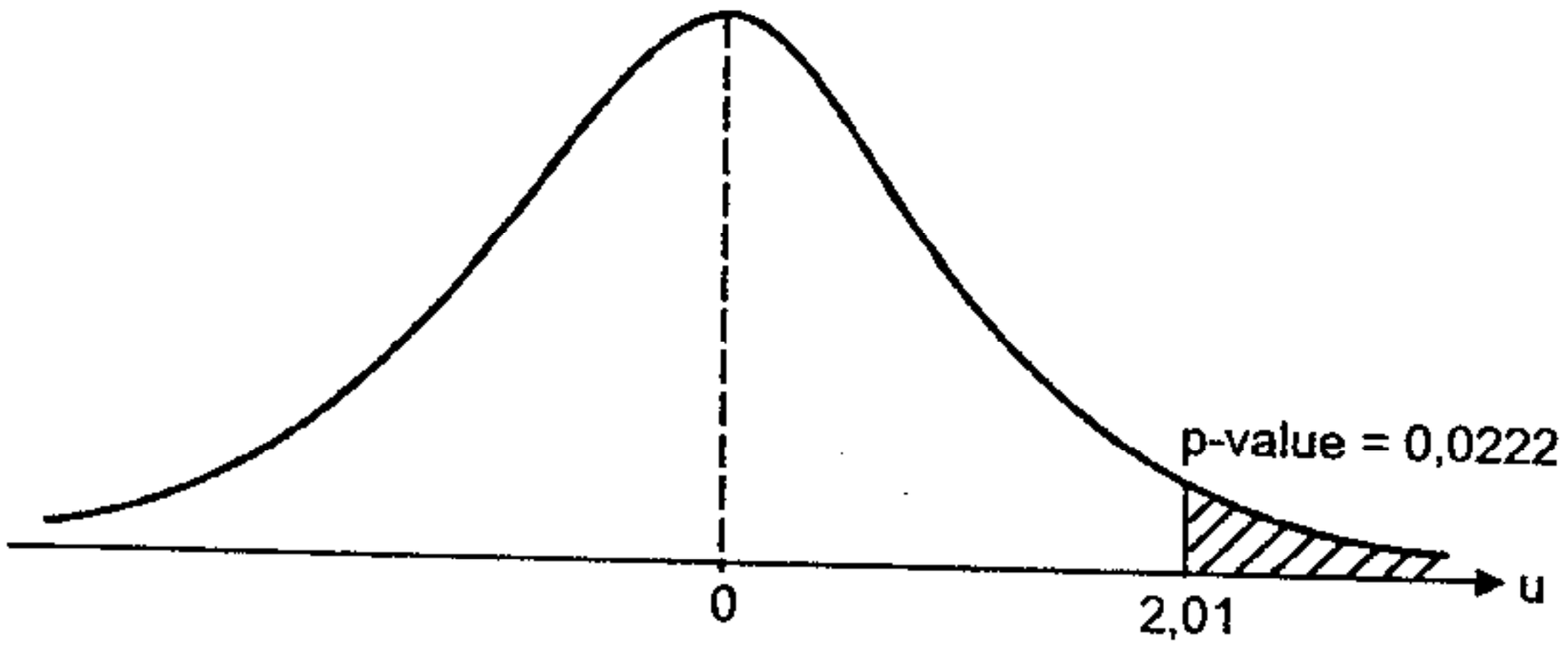
b) Giá trị P sẽ thay đổi như thế nào nếu trung bình mẫu tìm được không phải bằng 390 mà là 397.

Giải: a) Cặp giả thuyết thống kê có dạng H_0: μ = 380; H_1: μ > 380.

Từ mẫu điều tra trên 50 con bò ta tìm được giá trị quan sát của tiêu chuẩn kiểm định là:

$$U_{qs} = \frac{(\overline{x} - 380)\sqrt{50}}{35,2} = \frac{(390 - 380)\sqrt{50}}{35,2} = 2,01$$

Giá trị P của kiểm định (tức là mức độ khẳng định việc bác bỏ H_1) là xác suất để giá trị quan sát của $\overline{X}$ lớn hơn 390 nếu giả thuyết H_0 là đúng. Giá trị này có thể tính bằng cách sử dụng giá trị quan sát 2,01 của tiêu chuẩn kiểm định và tìm xác suất để U lớn hơn 2,01. Nó bằng 0,0222. Giá trị này được minh họa trên đồ thị như sau (Hình 8.3).



Hình 8.3. **P-value**

b) Với $\overline{x}$ ta tìm được:

$$U_{qs} = \frac{(397 - 380)\sqrt{50}}{35,2} = 3,415$$

Vì giá trị gần nhất của bảng giá trị tới hạn U là 3,5 nên giá trị P có thể lấy gần bằng 0,0002. P-value ≈ 0,0002.

Như đã thấy qua thí dụ 4, giá trị P là xác suất nhỏ nhất để kết quả quan sát được qua mẫu hiện tại mâu thuẫn nhiều

hơn với H_0 so với kết quả quan sát mẫu trước đó. Nói cách khác nó khẳng định độ tin cậy được của H_0. Giá trị xác suất này càng nhỏ thì mức độ khẳng định của mẫu về việc bác bỏ H_0 càng rõ rệt hơn hay H_0 càng kém tin cậy hơn. Chẳng hạn kiểm định với giá trị P bằng 0,01 cho thấy mức độ khẳng định để bác bỏ H_0 rõ ràng hơn là kiểm định với giá trị P bằng 0,2.

Trên đây là giá trị P trong kiểm định bên phải. Nếu tiến hành kiểm định bên trái với cặp giả thuyết H_0: $\mu = 380$; H_1: $\mu < 380$ ta cũng thu được kết quả tương tự.

Từ đó ta có công thức tính giá trị P (P-value) cho kiểm định giả thuyết thống kê như sau:

Nếu H_1: $\mu > \mu_0$ thì

$$\text{P-value} = P(U > U_{qs}) \qquad (8.14)$$

Nếu H_1: $\mu < \mu_0$ thì

$$\text{P-value} = (U < U_{qs}) \qquad (8.15)$$

Nếu H_1: $\mu \neq \mu_0$ thì

$$\text{P-value} = P(U > |U_{qs}|) \qquad (8.16)$$

Trong thực tế việc kiểm định theo giá trị P (P-value) thường được tiến hành theo nguyên tắc sau:

- Nếu P-value > 0,1 thì thường người ta thừa nhận H_0.
- Nếu 0,05 < P-value < 0,1 thì cần cân nhắc cẩn thận trước khi bác bỏ H_0.
- Nếu 0,01 < P-value < 0,05 thì nghiêng về hướng bác bỏ H_0 nhiều hơn.
- Nếu 0,001 < P-value < 0,01 thì có thể ít băn khoăn khi bác bỏ H_0.

- Nếu P-value < 0,001 thì có thể hoàn toàn yên tâm khi bác bỏ H_0.

Mặt khác nếu quy định trước mức ý nghĩa α thì có thể dùng P-value để kết luận theo α. Lúc đó nguyên tắc kiểm định như sau:

- Nếu P-value < α thì bác bỏ H_0, thừa nhận H_1.

- Nếu P-value > α thì chưa có cơ sở bác bỏ H_0.

Theo cách kiểm định này thì việc sử dụng P-value lại chính là kiểm định theo cách tiếp cận truyền thống.

5. Mối liên hệ giữa miền bác bỏ và khoảng tin cậy

Ta có thể chứng tỏ được rằng việc tìm miền bác bỏ với mức ý nghĩa α cũng chính là tìm khoảng tin cậy tương ứng với độ tin cậy $1 - \alpha$. Như ở phần trên khi kiểm định cặp giả thuyết H_0: $\mu = \mu_0$ và H_1: $\mu \neq \mu_0$ ta đã đòi hỏi là xác suất để tiêu chuẩn kiểm định

$$U = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

thuộc vào miền bác bỏ hai phía bằng α, do đó xác suất để U thuộc miền thừa nhận giả thuyết $(-u_{\alpha/2}, u_{\alpha/2})$ bằng $1 - \alpha$, tức là thỏa mãn điều kiện

$$P\left[-u_{\alpha/2} < \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} < u_{\alpha/2}\right] = 1 - \alpha$$

hay
$$P\left[\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\alpha/2}\right] = 1 - \alpha$$

Ta lại thu được công thức ước lượng khoảng tin cậy đối xứng giá trị μ khi đã biết σ.

Chú ý rằng hai thủ tục trên, mặc dù đi đến cùng một kết quả song cách giải thích những kết quả đó lại khác nhau. Miền bác bỏ xác định các giới hạn (giá trị tới hạn) trong đó chứa đựng $(1 - \alpha)\%$ số các giả thuyết H_0 được chấp nhận khi lặp lại nhiều lần phép thử. Còn khoảng tin cậy thì xác định các giới hạn trong đó $(1 - \alpha)\%$ phép thử sẽ chứa giá trị thực của tham số cần ước lượng.

2.2. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên phân phối chuẩn khi chưa biết phương sai

Lúc đó tiêu chuẩn kiểm định là thống kê

$$G = T = \frac{(\overline{X} - \mu_0)}{Se(\overline{X})} = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \qquad (8.17)$$

Nếu giả thuyết H_0 đúng thì ta có:

$$T = \frac{(\overline{X} - \mu_0)}{S} = \frac{(\overline{X} - \mu)\sqrt{n}}{S}$$

Và từ mục §6 Chương VI ta đã biết T phân phối T (n – 1). Từ đó, tùy thuộc vào dạng của giả thuyết đối H_1, miền bác bỏ "tốt nhất" được xây dựng theo các trường hợp sau:

a) H_0: $\mu = \mu_0$; H_1: $\mu > \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn Student $t_\alpha^{(n-1)}$ sao cho

$$P(G \in W_\alpha / H_0) = P(T > t_\alpha^{(n-1)}) = \alpha$$

Ta thu được miền bác bỏ bên phải W_α sau:

$$W_\alpha = \left\{ T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S};\ T > t_\alpha^{(n-1)} \right\} \qquad (8.18)$$

b) H_0: $\mu = \mu_0$; H_1: $\mu < \mu_0$

Lúc đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn Student $t_{1-\alpha}^{(n-1)}$ sao cho

$$P(G \in W_\alpha / H_0) = P(T < t_{1-\alpha}^{(n-1)}) = P(T < -t_\alpha^{(n-1)}) = \alpha$$

Ta thu được miền bác bỏ bên trái W_α sau:

$$W_\alpha = \left\{ T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S}; T < -t_\alpha^{(n-1)} \right\} \qquad (8.19)$$

c) H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$

Lúc đó với mức ý nghĩa cho trước có thể tìm được hai giá trị tới hạn Student là $t_{\alpha/2}^{(n-1)}$ và $t_{1-\alpha/2}^{(n-1)}$ sao cho

$$P(G \in W_\alpha / H_0) = P(T < t_{1-\alpha/2}^{(n-1)}) + P(T > t_{\alpha/2}^{(n-1)})$$

$$= P(T < t_{\alpha/2}^{(n-1)}) + P(T > t_{\alpha/2}^{(n-1)})$$

$$= P(|T| > t_{\alpha/2}^{(n-1)}) = \alpha$$

Ta thu được miền bác bỏ hai phía W_α sau:

$$W_\alpha = \left\{ T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S}; |T| > t_{\alpha/2}^{(n-1)} \right\} \qquad (8.20)$$

Với một mẫu cụ thể tính được $\overline{x}$, s; từ đó tính được giá trị quan sát T của tiêu chuẩn kiểm định

$$T_{qs} = \frac{(\overline{x} - \mu_0)\sqrt{n}}{s}$$

So sánh T_{qs} với miền bác bỏ W_α và kết luận:

- Nếu $T_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1;

- Nếu $T_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Thí dụ 5: Trọng lượng đóng bao của các bao gạo trong kho là biến ngẫu nhiên phân phối chuẩn với trọng lượng trung bình theo quy định là 50kg.

Nghi ngờ bị đóng thiếu, người ta đem cân ngẫu nhiên 25 bao và thu được các số liệu sau (bảng 8.2).

Bảng 8.2

Trọng lượng bao (kg)	Số bao tương ứng
48,0 – 48,5	2
48,5 – 49,0	5
49,0 – 49,5	10
49,5 – 50,0	6
50,0 – 50,5	2
	n = 25

Với ý nghĩa $\alpha = 0,01$ hãy kết luận về điều nghi ngờ nói trên.

Giải. Gọi X là trọng lượng đóng bao. Theo giả thiết X phân phối chuẩn. Vậy trọng lượng đóng bao trung bình chính là tham số μ. Đây là bài toán kiểm định giả thuyết về tham số μ của phân phối chuẩn $N(\mu, \sigma^2)$ khi chưa biết σ^2.

Cặp giả thuyết thống kê: $H_0: \mu = 50$; $H_1: \mu < 50$

Vậy theo công thức (8.19) ta có tiêu chuẩn kiểm định là:

$$T = \frac{(\overline{X} - 50)\sqrt{25}}{S}$$

trong đó $\overline{X}$ và S là trung bình và độ lệch chuẩn của mẫu ngẫu nhiên kích thước n = 25.

Cũng từ (8.19)

$$-t_{\alpha}^{(n-1)} = -t_{0,01}^{(24)} = -2,402$$

Vậy miền bác bỏ $(-\infty; -2,402)$

Từ mẫu cụ thể ta lập bảng tính $\overline{x}$ và s.

x_i	n_i	$x_i n_i$	$n_i x_i^2$
48,25	2	96,5	4656,125
48,75	5	243,75	11882,8125
49,25	10	492,5	24255,625
49,75	6	298,5	14850,375
50,25	2	100,5	5050,125
	n = 25	Σ = 1231,75	Σ = 60695,062

Từ đó: $\overline{x} = \frac{1231,75}{25} = 49,27$

$$MS = \frac{60695,062}{25} - (49,27)^2 = 0,27$$

$$S = \sqrt{\frac{25}{24}.0,27} = 0,53$$

Giá trị quan sát của tiêu chuẩn kiểm định

$$T_{qs} = \frac{(49,27-50)\sqrt{25}}{0,53} = -6,887$$

Vậy $T_{qs} \in W_\alpha$: Bác bỏ H_0, thừa nhận H_1, tức là qua mẫu cụ thể này thừa nhận gạo bị đóng thiếu với mức ý nghĩa 0,01.

Việc kiểm định bằng cách sử dụng tiêu chuẩn (8.17) thường được gọi là kiểm định T (T - test).

Trong trường hợp này xác suất mắc sai lầm loại 2 được tìm bằng các công thức sau:

- Nếu là kiểm định một phía (bên phải hoặc bên trái) thì

$$\beta = P\left[T < t_\alpha^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(x)}\right] \qquad (8.21)$$

- Nếu là kiểm định hai phía thì

$$\beta = P\left[T < t_{\alpha/2}^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(x)}\right] \qquad (8.22)$$

trong đó $Se(\bar{x}) = \frac{s}{\sqrt{n}}$.

Và kích thước mẫu tối thiểu n cần điều tra để xác suất mắc sai lầm loại 1 và loại 2 không quá α và β tương ứng và giá trị thực μ_1 không sai lệch so với μ_0 quá giá trị Δ được tính bằng các công thức sau:

- Nếu là kiểm định một phía

$$n \geq \left[\frac{s^2}{\Delta^2}(t_\alpha^{(m-1)} + t_\beta^{(m-1)})^2\right] \qquad (8.23)$$

- Nếu là kiểm định hai phía

$$n \geq \left[\frac{s^2}{\Delta^2}(t_{\alpha/2}^{(m-1)} + t_\beta^{(m-1)})^2\right] \qquad (8.24)$$

Trong đó s^2 là phương sai của mẫu sơ bộ kích thước $m \geq 2$ và $t_\alpha^{(m-1)}$, $t_{\alpha/2}^{(m-1)}$ và $t_\beta^{(m-1)}$ là các giá trị tới hạn tương ứng mức α, $\frac{\alpha}{2}$ và β và số bậc tự do là $m - 1$.

Việc tìm các giá trị P để kiểm định theo cách tiếp cận P-value cũng tiến hành tương tự như đã làm ở mục trước.

Thí dụ 6: Tiếp tục thí dụ 5.

a) Tìm xác suất mắc sai lầm loại 2 nếu trọng lượng đóng bao trung bình thực là 49,5 kg.

b) Tìm giá trị P.

Giải: a) Vì đây là kiểm định bên trái nên theo công thức (8.21) ta có:

$$\beta = P\left[T < t_\alpha^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})}\right]$$

$$Se(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0,53}{\sqrt{25}} = 0,106$$

với $\alpha = 0,01 \Rightarrow t_\alpha^{(n-1)} = t_{0,01}^{(24)} = 2,492$.

$$\text{Vậy } \beta = P\left[T < 2,492 - \frac{|50 - 49,5|}{0,106}\right] = P[T < -2,675]$$

$$= P[T > 2,675] \approx 0,005$$

b) Từ thí dụ 5 ta đã tìm được.

$$T_{qs} = -6,887$$

Vì đây là kiểm định bên trái nên

$$\text{P-value} = P(T < T_{qs}) = P(T < -6,887) = P(T > 6,887)$$

Với số bậc tự do là 24, giá trị tối đa của T là 3,467 do đó P-value < 0,001.

Ví dụ A: Bằng Stata việc kiểm định giả thuyết H_0: $\mu_0 = 1650$ cho kết quả sau:

• ttest x 1 = 1650

One-sample t test Number of obs = 100

Variable	Mean	Std. Err.	t	P > |t|	[95% Conf.	Interval]
x1	1649.73	4.73749	348.229	0.000	1640.33	1659.13

Degrees of freedom: 99

Ho: mean (x1) = 1650

Ha: mean < 1650 Ha: mean ~= 1650 Ha: mean > 1650

t = –0.0570 t = –0.0570 t = –0.0570

P < t = 0.4773 P > |t| = 0.9547 P > t = 0.5227

2.3. Kiểm định giả thuyết về hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu trong đó các biến ngẫu nhiên X_1 và X_2 cùng phân phối chuẩn với các kỳ vọng toán là μ_1, μ_2 và các phương sai là σ_1^2 và σ_2^2. Nếu μ_1 và μ_2 chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau người ta đưa ra giả thuyết thống kê.

$$H_0: \mu_1 = \mu_2$$

Để kiểm định giả thuyết trên ta xét một số trường hợp sau:

1. Nếu đã biết các phương sai σ_1^2 và σ_2^2 của các biến ngẫu nhiên gốc trong tổng thể và từ hai tổng thể trên có thể rút ra hai mẫu độc lập kích thước n_1 và n_2:

$$W_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$$

Lúc đó tiêu chuẩn kiểm định được chọn là thống kê

$$G = U = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (8.25)$$

Từ mục §6 Chương VI ta đã biết thống kê U phân phối N(0,1). Nếu giả thuyết H_0 đúng thì thống kê U có dạng

$$U = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (8.26)$$

và cũng phân phối N(0,1). Vì vậy, với mức ý nghĩa bằng α cho trước và tùy thuộc vào dạng của giả thuyết đối H_1, với phương pháp xây dựng giống như đã làm ở các phần trên ta thu được các miền bác bỏ W_α tương ứng sau:

a) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 > \mu_2$

Miền bác bỏ bên phải là

$$W_\alpha = \left\{ U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; U > u_\alpha \right\} \qquad (8.27)$$

b) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$

Miền bác bỏ bên trái là

$$W_\alpha = \left\{ U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; U < -u_\alpha \right\} \quad (8.28)$$

c) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

Miền bác bỏ 2 phía là

$$W_\alpha = \left\{ U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; |U| > u_{\alpha/2} \right\} \quad (8.29)$$

Lập hai mẫu cụ thể từ X_1 và X_2 và tính được các trung bình mẫu cụ thể

$$\overline{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}; \quad \overline{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$$

và tính được giá trị quan sát của tiêu chuẩn kiểm định

$$U_{qs} = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Xem xét U_{qs} có thuộc W_α không để kết luận.

Thí dụ 7: Tại một xí nghiệp người ta xây dựng hai phương án gia công cùng một loại chi tiết. Để đánh giá xem chi phí trung bình về nguyên liệu theo hai phương án ấy có khác nhau hay không người ta tiến hành sản xuất thử và thu được các kết quả sau:

Phương án 1:	2,5	3,2	3,5	3,8	3,5	
Phương án 2:	2,0	2,7	2,5	2,9	2,3	2,6

Với mức ý nghĩa $\alpha = 0,05$, hãy kết luận về vấn đề trên biết rằng chi phí nguyên liệu theo cả hai phương án gia công đều là các biến ngẫu nhiên phân phối chuẩn với $\sigma_1^2 = \sigma_2^2 = 0,16$.

Giải. Gọi X_1 và X_2 tương ứng là chi phí nguyên liệu theo hai phương án gia công trên. Theo giả thiết X_1 và X_2 phân phối chuẩn. Vậy chi phí nguyên liệu trung bình theo các phương án đó là μ_1 và μ_2. Vậy đây là bài toán kiểm định H_0: $\mu_1 = \mu_2$ với H_1: $\mu_1 \neq \mu_2$ khi đã biết σ_1^2 và σ_2^2.

Theo (8.29) tiêu chuẩn kiểm định là

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{0,16}{5} + \frac{0,16}{6}}}$$

Do $\alpha = 0,05 \Rightarrow u_{\alpha/2} = u_{0,025} = 1,96$. Vậy miền bác bỏ là $(-\infty; -1,96)$ và $(1,96; +\infty)$.

Từ mẫu cụ thể ta tính được

$$\overline{x}_1 = \frac{2,5 + 3,2 + 3,5 + 3,8 + 3,5}{5} = 3,3$$

$$\overline{x}_2 = \frac{2,0 + 2,7 + 2,5 + 2,9 + 2,3 + 2,6}{6} = 2,5$$

và giá trị quan sát của tiêu chuẩn kiểm định

$$U_{qs} = \frac{3,3 - 2,5}{\sqrt{\frac{0,16}{5} + \frac{0,16}{6}}} = 3,33$$

Vậy $U_{qs} \in W_\alpha$ bác bỏ H_0 thừa nhận H_1 tức là chi phí nguyên liệu theo hai phương án gia công trên thực sự khác nhau.

Việc tìm xác suất mắc sai lầm loại 2, kích thước mẫu n và giá trị P cũng được tiến hành như đã làm ở các mục trước. Chẳng hạn với α và β định trước và nếu muốn $|\mu_1 - \mu_2| \leq \Delta$ thì kích thước mẫu tối thiểu trong trường hợp kiểm định một phía là

$$n \geq \left[\frac{2\hat{\sigma}^2 (u_\alpha + u_\beta)^2}{\Delta^2} \right] \tag{8.30}$$

Còn trong trường hợp kiểm định hai phía là

$$n \geq \left[\frac{2\hat{\sigma}^2 (u_{\alpha/2} + u_\beta)^2}{\Delta^2} \right] \tag{8.31}$$

trong đó $n_1 = n_2 = n$ và $\hat{\sigma}^2 = \frac{\sigma_1^2 + \sigma_2^2}{2}$.

Chú ý rằng phương pháp kiểm định vừa trình bày ở trên còn có thể áp dụng khi μ_1 và μ_2 khác nhau. Chẳng hạn cần kiểm định giả thuyết H_0: $\mu_1 - \mu_2 = D_0$ với D_0 là một giá trị định trước thì tiêu chuẩn kiểm định sẽ là:

$$G = U = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{8.32}$$

và thủ tục kiểm định được tiến hành hoàn toàn giống như đã trình bày ở trên.

2. Nếu chưa biết các phương sai σ_1^2 và σ_2^2 của các biến ngẫu nhiên gốc trong tổng thể song giả định rằng chúng bằng nhau ($\sigma_1^2 = \sigma_2^2$). Việc kiểm định $\sigma_1^2 = \sigma_2^2$ sẽ được xét ở mục 2.7. Mặt khác, ta vẫn giả định rằng có thể điều tra được 2 mẫu độc lập kích thước n_1 và n_2.

Lúc đó tiêu chuẩn kiểm định được chọn là

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (8.33)$$

Với $S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$

Ta biết rằng T phân phối Student với $n_1 + n_2 - 2$ bậc tự do.

Với điều kiện giả thuyết H_0 là đúng thì tiêu chuẩn kiểm định trở thành

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (8.34)$$

và vẫn phân phối $T(n_1 + n_2 - 2)$. Do đó tùy thuộc vào giả thuyết đối H_1 ta có các miền bác bỏ mức α như sau:

a) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 > \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}; T > t_\alpha^{(n_1+n_2-2)} \right\} \qquad (8.35)$$

b) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}; T < -t_\alpha^{(n_1+n_2-2)} \right\} \qquad (8.36)$$

c) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}; |T| > t_{\alpha/2}^{(n_1+n_2-2)} \right\} \quad (8.37)$$

Qua hai mẫu cụ thể tính được $\overline{x}_1, \overline{x}_2, s_1^2, s_2^2$ và giá trị quan sát T_{qs} của tiêu chuẩn kiểm định, từ đó so sánh với W_α và kết luận.

Việc suy ra một cách tương ứng cho việc kiểm định giả thuyết $H_0 = \mu_1 - \mu_2 = D_0$ cũng tiến hành tương tự như ở mục trước.

Việc xác định β, P-value và kích thước mẫu n cũng tiến hành tương tự. Sự thay đổi duy nhất so với công thức (8.30) và (8.31) là tính giá trị của $\hat{\sigma}^2$ theo công thức

$$\hat{\sigma}^2 = S_p^2$$

Thí dụ 8: Một nghiên cứu được thực hiện đối với 20 người ở một phường và 19 người ở một phường khác trong thành phố để xem thu nhập trung bình hàng năm (tính bằng triệu đồng) của dân cư hai phường đó có thực sự khác nhau hay không. Các số liệu mẫu thu được như sau:

$n_1 = 20$ $\qquad$ $n_2 = 19$

$\overline{x}_1 = 18{,}27$ $\qquad$ $\overline{x}_2 = 16{,}78$

$s_1^2 = 8{,}74$ $\qquad$ $s_2^2 = 6{,}58$

Vậy với mức ý nghĩa 0,05 có thể cho rằng thu nhập trung bình của dân cư ở hai phường đó khác nhau hay không? Giả thiết thu nhập hàng năm của dân cư hai phường cùng phân phối chuẩn với phương sai như nhau.

Giải: Gọi X_1 và X_2 tương ứng là thu nhập hàng năm của dân cư hai phường đó. Theo giả thiết X_1 và X_2 phân phối chuẩn với các phương sai $\sigma_1^2 = \sigma_2^2$. Do đó để kiểm định cặp giả thuyết

$$H_0: \mu_1 = \mu_2;\ H_1: \mu_1 \neq \mu_2$$

ta sử dụng công thức kiểm định (8.37)

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}};\ |T| > t_{\alpha/2}^{(n_1+n_2-2)} \right\}$$

với $\alpha = 0,05 \Rightarrow t_{\alpha/2}^{(n_1+n_2-2)} = t_{0,025}^{(37)} \approx 2,021$

Vậy miền bác bỏ là $(-\infty; -2,021)$ và $(2,021; +\infty)$.

Từ mẫu cụ thể ta tính được

$$S_p = \sqrt{\frac{19.8,74 + 18.6,58}{20 + 19 - 2}} = 2,773$$

do đó $T_{qs} = \dfrac{18,27 - 16,78}{2,773\sqrt{\dfrac{1}{20} + \dfrac{1}{19}}} = 1,677$

$T_{qs} \neq W_\alpha$ do đó với mức ý nghĩa 0,05 qua hai mẫu cụ thể đã cho, chưa có cơ sở để bác bỏ H_0, tức là có thể xem thu nhập trung bình hàng năm của dân cư hai phường đó là như nhau.

3. Nếu chưa biết các phương sai σ_1^2 và σ_2^2 của các tổng thể và không thể cho rằng chúng bằng nhau ($\sigma_1^2 \neq \sigma_2^2$). Lúc đó nếu có thể điều tra được từ tổng thể hai mẫu độc lập kích thước n_1 và n_2 thì chọn lập thống kê

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (8.38)$$

Từ mục 6 chương VI ta đã biết T phân phối Student với số bậc tự do là

$$k = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)C^2 + (1 - C)^2(n_1 - 1)}$$

với $$C = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^{2\cdot}}{n_2}}$$

Nếu giả thuyết H_0 là đúng thì tiêu chuẩn kiểm định trở thành

$$G = T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (8.39)$$

và vẫn phân phối T(k). Vì vậy, miền bác bỏ mức α được xác định bằng các công thức sau:

a) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 > \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}};\ T > t_\alpha^{(k)} \right\} \qquad (8.40)$$

b) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}; T < -t_\alpha^{(k)} \right\} \quad (8.41)$$

c) H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}; |T| > t_{\alpha/2}^{(k)} \right\} \quad (8.42)$$

Việc xác định xác suất mắc sai lầm loại hai, P-value và kích thước mẫu điều tra cũng tiến hành giống như ở mục trước. Thay đổi duy nhất là giá trị $\hat{\sigma}^2$ trong công thức xác định n được tính bằng công thức xấp xỉ

$$\hat{\sigma}^2 = \frac{m_1 S_1^2 + m_2 S_2^2}{m_1 + m_2}$$

với S_1^2 và S_2^2 là phương sai của các mẫu sơ bộ kích thước m_1 và m_2.

Thí dụ 9. Để kiểm nghiệm hiệu quả của một loại thuốc tẩy giun cho lợn, người ta bắt ngẫu nhiên 14 con lợn từ một trại chăn nuôi và chia thành hai nhóm:

Nhóm I: Cho uống thuốc tẩy giun

Nhóm II: Không cho uống thuốc tẩy giun

Sau thời gian dùng thuốc, khi giết thịt, hai nhóm lợn trên cho kết quả sau về số giun có trong những con lợn thuộc hai nhóm trên.

Nhóm I:	18	43	28	50	16	32	13
Nhóm II:	40	54	26	63	21	37	39

a) Với mức ý nghĩa 0,05 hãy kết luận xem loại thuốc tẩy giun nói trên có thực sự hiệu quả hay không.

b) Tìm β nếu $|\mu_1 - \mu_2| = 10$

c) Tìm P-value

Giả thiết số lượng giun phân phối chuẩn.

Giải: a) Gọi X_1 và X_2 là số giun trong mỗi con lợn thuộc hai nhóm trên. Theo giả thuyết X_1 và X_2 phân phối chuẩn với σ_1^2 và σ_2^2 chưa biết và không thể cho rằng chúng bằng nhau. Vậy số giun trung bình là μ_1 và μ_2. Để kiểm định cặp giả thuyết H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$ ta dùng công thức (8.41)

$$W_\alpha = \left\{ T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} ; T < -t_\alpha^{(k)} \right\}$$

từ 2 mẫu cụ thể tính được

$n_1 = 7$ $\qquad$ $n_2 = 7$

$\overline{x}_1 = 28,57$ $\qquad$ $\overline{x}_2 = 40$

$s_1^2 = 198,62$ $\qquad$ $s_2^2 = 215,33$

Ta có:

$$C = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 0,4798$$

$$\Rightarrow \quad k = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)C^2 + (1 - C)^2(n_1 - 1)}$$

$$= \frac{6.6}{6.0{,}4798^2 + 0{,}5202^2.6} \approx 12$$

với $\alpha = 0{,}05 \Rightarrow t_{0,05}^{(12)} = 1{,}782$

Vậy miền bác bỏ là $(-\infty; -1{,}782)$

$$T_{qs} = \frac{28{,}57 - 40}{\sqrt{\frac{198{,}62}{7} + \frac{215{,}33}{7}}} = -1{,}49$$

$T_{qs} \notin W_\alpha$ vậy với mức ý nghĩa 0,05 chưa có cơ sở bác bỏ H_0 hay chưa thể nói loại thuốc tẩy giun được thử nghiệm là có hiệu quả.

b) Ta tìm β theo công thức

$$\beta = P\left[T < t_\alpha^{(k)} - \frac{|\mu_1 - \mu_2|}{S(\overline{X}_1 - \overline{X}_2)}\right]$$

$$S^2(\overline{X}_1 - \overline{X}_2) \approx \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} = \frac{198{,}62}{7} + \frac{215{,}33}{7} = 59{,}14$$

$$\Rightarrow \quad S(\overline{X}_1 - \overline{X}_2) \approx 7{,}69$$

$$\beta = P\left[T < 1{,}782 - \frac{10}{7{,}69}\right] = P[T < +0{,}48]$$

$$= 1 - P\,[T > 0{,}48]$$

Với k = 12 giá trị nhỏ nhất của T là 1,356 nên $\beta < 0{,}9$

c) P-value = $P[T < T_{qs}] = P[T < -1,49] = P[T > 1,49]$

Với k = 12 giá trị 1,49 nằm trong khoảng hai giá trị 1,356 và 1,782. Vậy

$$0,05 < \text{P-value} < 0,1$$

Trong các trường hợp kiểm định trên ta luôn giả thiết biến ngẫu nhiên trọng các tổng thể phân phối chuẩn. Trong thực tế nếu X_1 và X_2 không phân phối chuẩn (việc kiểm định phân phối chuẩn sẽ được đề cập ở mục 3) song nếu điều tra được 2 mẫu độc lập với kích thước $n_1 > 30$ và $n_2 > 30$ thì các thống kê

$$U = \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

hoặc thống kê

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

vẫn phân phối xấp xỉ N(0,1) theo định lý giới hạn trung tâm. Do đó vẫn có thể áp dụng các thủ tục kiểm định nói trên với các biến ngẫu nhiên X_1 và X_2 phân phối theo các quy luật bất kỳ, chỉ cần thỏa mãn hai điều kiện:

- Các mẫu điều tra là độc lập
- Kích thước của cả hai mẫu đều > 30.

Thí dụ: Bằng phần mềm Stata, với các số liệu của bảng A, kiểm định H_0: $\mu_1 = \mu_2$ với các phương sai tổng thể chưa biết song giả thiết bằng nhau cho ta kết quả sau:

ttest x1=x2, unpaired

Two-sample t test with equal variances

x1 = Number of obs = 100

x2 = Number of obs = 100

Variable	Mean	Std.Err.	t	P > |t|	[95% Conf.	Interval]
x1	1649.73	4.73749	348.2	0.	1640.33	1659.13
x2	1655.99	4.740828	349.304	0.000	1646.58	1665.397
diff	-6.26	6.702183	- 934024	0.3514	-19.47682	6.956823

Degrees of freedom: 198

Ho: mean (x1) – mean(x2) = diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = –0.9340	t = –0.9340	t = –0.9340
P < t = 0.1757	P > |t| = 0.3514	P > t = 0.8243

Và nếu kiểm định $\mu_1 = \mu_2$ khi các phương sai tổng thể khác nhau cho kết quả sau:

ttest x1 = x2, unpaired unequal

Two-sample t test with unequal variances

x1 = Number of obs = 100

x2 = Number of obs = 100

Variable	Mean	Std.Err.	t	P > |t|	[95% Conf.	Interval]
x1	1649.73	4.73749	348.229	0.0000	1640.33	1659.13
x2	1655.99	4.740828	349.304	0.0000	1646.583	1665.397
diff	–6.26	6.702183	-.934024	0.3514	-19.47682	6.956823

Satterthwaite's degrees of freedom: 197.9999

Ho: mean (x1) – mean(x2) = diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = –0.9340	t = –0.9340	t = –0.9340
P < t = 0.1757	P > |t| = 0.3514	P > t = 0.8243

Thí dụ 10: Người ta cân trẻ sơ sinh ở hai khu vực thành thị và nông thôn, thu được kết quả sau (Bảng 8.3)

Bảng 8.3

Khu vực	Số trẻ được cân	Trọng lượng trung bình	Phương sai
Nông thôn	$n_1 = 2500$	$\bar{x}_1 = 3,0$	$x_1^2 = 200$
Thành thị	$n_2 = 500$	$\bar{x}_2 = 3,1$	$x_2^2 = 5$

Với mức ý nghĩa 0,01 có thể coi trọng lượng trung bình của trẻ sơ sinh ở hai khu vực bằng nhau được không?

Giải: Gọi trọng lượng trẻ sơ sinh ở nông thôn và thành thị tương ứng là X_1 và X_2. Vậy trọng lượng trẻ sơ sinh trung bình chính là m_1 và m_2. Đây là bài toán kiểm định cặp giả thuyết H_0: $m_1 = m_2$; H_1: $m_1 \neq m_2$ khi chưa biết các phương sai σ_1^2 và σ_2^2. Do $n_1 > 30$ và $n_2 > 30$ nên theo công thức (8.42) tiêu chuẩn kiểm định có dạng:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{2500} + \frac{S_2^2}{500}}} = -0,33$$

Do $\alpha = 0,01 \rightarrow \frac{\alpha}{2} = 0,005 \Rightarrow u_{0,005} = 2,576$ nên miền bác bỏ có dạng $(-\infty; -2,576)$ và $(2,576; +\infty)$

Qua mẫu cụ thể tính được

$$U_{qs} = \frac{3,0 - 3,1}{\sqrt{\frac{200}{2500} + \frac{5}{500}}} = -0,33$$

Vậy $U_{qs} \notin W_\alpha$ chưa có cơ sở để bác bỏ H_0, như vậy có thể coi trọng lượng trẻ sơ sinh ở nông thôn và thành thị là như nhau...

Sau đây là kết quả giải bài toán bằng phần mềm Stata:

ttesti 2500 3.0 14.14 500 3.1 2.24, **unequal**

x: Number of obs = 2500

y: Number of obs = 500

Variable	Mean	Std.Err.	t	P > |t|	[95% Conf.	Interval]
X	3	.2828	10.6082	0.000	2.445454	3.554546
Y	3.1	.10017	30.9456	0.000	2.903182	3.296818
diff	-.1	.3000184	-.33331	0.738	-.6882679	.4882679

Statterthwaite's degrees of freedom: 2934.1293

Ho: mean(x) – mean(y) = diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = –0.3333	t = –0.3333	t = –0.333
P < t = 0.3695	P > |t| = 0.7389	P > t = 0.6305

4. Trường hợp hai mẫu điều tra phụ thuộc theo từng cặp

Ở các phần trên khi kiểm định về hai kỳ vọng toán ta luôn giả thiết rằng có thể điều tra được hai mẫu độc lập từ các tổng thể nghiên cứu. Trong thực tế có nhiều trường hợp

hai mẫu điều tra được rút ra từ cùng một tổng thể thì khả năng chúng phụ thuộc nhau càng cao. (Vấn đề này sẽ được kiểm định ở mục 3.7). Lúc đó các phương pháp kiểm định ở trên không thể áp dụng được vì các kết quả kiểm định không còn đáng tin cậy nữa. Ở phần này ta sẽ xét việc kiểm định khi hai mẫu điều tra có cùng kích thước n, trong đó các giá trị của mẫu phụ thuộc tương ứng với nhau theo từng cặp.

Giả sử có hai tổng thể nghiên cứu trong đó có các biến ngẫu nhiên X_1 và X_2 cùng phân phối chuẩn với các phương sai chưa biết. Với mức ý nghĩa α phải kiểm định giả thuyết thống kê

$$H_0: \mu_1 = \mu_2 \text{ hay } H_0: \mu_1 - \mu_2 = 0$$

Từ hai tổng thể rút ra hai mẫu ngẫu nhiên kích thước n

$$W_1 = (X_{11}, X_{12},..., X_{1n})$$

$$W_2 = (X_{21}, X_{22},..., X_{2n})$$

Do các giá trị của 2 mẫu tương ứng với nhau theo từng cặp do đó ta thiết lập biến ngẫu nhiên D với các giá trị có thể có $D_1, D_2,..., D_n$ là hiệu số của các cặp giá trị tương ứng của 2 mẫu, như vậy

$$D_i = X_{1i} - X_{2i} \ (i = \overline{1,n}) \qquad (8.43)$$

Lúc đó kỳ vọng toán của D là

$$\mu_D = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

và việc kiểm định giả thuyết $H_0: \mu_1 - \mu_1 = 0$ có thể viết dưới dạng $H_0: \mu_D = 0$. Bài toán đưa về dạng kiểm định giá trị của tham số μ như đã xét ở mục 2.1.

Từ các giá trị D_i tìm được ở (8.43) ta xác định trung bình mẫu và phương sai mẫu

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{8.44}$$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 \tag{8.45}$$

Từ đó chọn lập tiêu chuẩn kiểm định

$$G = T = \frac{\overline{D}\sqrt{n}}{S_D} \tag{8.46}$$

Ta biết rằng T phân phối Student với n – 1 bậc tự do, vì vậy với mức ý nghĩa α ta có các miền bác bỏ sau:

a) H_0: $\mu_D = 0$; H_1: $\mu_D > 0$

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D};\ T > t_\alpha^{(n-1)} \right\} \tag{8.47}$$

b) H_0: $\mu_D = 0$; H_1: $\mu_D < 0$

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D};\ T < -t_\alpha^{(n-1)} \right\} \tag{8.48}$$

c) H_0: $\mu_D = 0$; H_1: $\mu_D \neq 0$

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D};\ |T| > t_{\alpha/2}^{(n-1)} \right\} \tag{8.49}$$

Việc tìm β, P-value và kích thước mẫu n cũng tiến hành giống như ở mục 2.1. Chẳng hạn kích thước mẫu n khi kiểm định một phía là:

$$n \geq \left[\frac{S_D^2}{\Delta^2} (t_\alpha^{(n-1)} + t_\beta^{(n-1)})^2 \right] \tag{8.50}$$

và khi kiểm định 2 phía là:

$$n \geq \left[\frac{S_D^2}{\Delta^2} (t_{\alpha/2}^{(m-1)} + t_{\beta}^{(m-1)})^2 \right] \qquad (8.51)$$

trong đó S_D^2 là phương sai của mẫu sơ bộ kích thước $m \geq 2$, còn $t_{\alpha}^{(m-1)}$, $t_{\alpha/2}^{(m-1)}$ và $t_{\beta}^{(m-1)}$ là các giá trị tới hạn Student tương ứng.

Thí dụ 11: Theo dõi doanh số bán của một công ty (tính bằng triệu đồng) trong 15 ngày đầu tháng 3 và 15 ngày đầu tháng 5 thu được kết quả sau:

Ngày	Tháng 3	Tháng 5	di
1	7,6	7,3	0,3
2	10,2	9,1	1,1
3	9,5	8,4	1,1
4	1,3	1,5	–0,2
5	3,0	2,7	0,3
6	6,3	5,0	0,5
7	5,3	4,9	0,4
8	6,2	5,3	0,9
9	2,2	2,0	0,2
10	4,8	4,2	0,6
11	11,3	11,0	0,3
12	12,1	11,0	1,1
13	6,9	6,1	0,8
14	7,6	6,7	0,0
15	8,4	7,5	0,9

Nếu giả thiết doanh số hàng ngày phân phối chuẩn thì với mức ý nghĩa $\alpha = 0,05$ có thể cho rằng doanh số bán trung bình hàng ngày trong tháng 5 có giảm sút so với tháng 3 hay không?

Giải. Gọi X_1 và X_2 tương ứng là doanh số bán hàng ngày trong tháng 3 và tháng 5. Theo giả thiết X_1 và X_2 phân phối chuẩn. Vậy các doanh số trung bình là μ_1 và μ_2. Ta phải kiểm định cặp giả thuyết

$$H_0: \mu_1 - \mu_2 = 0;\ H_0: \mu_1 - \mu_2 > 0;$$

hay $$H_0: \mu_D = 0;\ H_1: \mu_D > 0$$

Theo công thức (8.47) công thức kiểm định là

$$W_\alpha = \left\{ T = \frac{\overline{D}\sqrt{n}}{S_D};\ T > t_\alpha^{(n-1)} \right\}$$

với $\alpha = 0,05 \Rightarrow t_\alpha^{(n-1)} = t_{0,05}^{(14)} = 1,761$

Vậy miền bác bỏ là $(1,761; +\infty)$

Từ các số liệu mẫu tìm được

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = 0,61$$

$$s_D^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^2 - \frac{(\sum d_i)^2}{n}\right] = 0,156$$

$$s_D = 0,394$$

Từ đó $$T_{qs} = \frac{0,61.\sqrt{5}}{0,394} = 6$$

Vì $T_{qs} \in W_\alpha$ nên bác bỏ H_0, thừa nhận H_1 tức là với mức ý

nghĩa 0,05 qua hai mẫu cụ thể trên, doanh số trung bình hàng ngày của tháng 5 thực sự giảm sút so với tháng 3.

Trên đây là một số kiểm định tham số đối với hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn. Việc mở rộng kiểm định cho k kỳ vọng toán (k > 2) của k biến ngẫu nhiên phân phối chuẩn trong k tổng thể nghiên cứu sẽ được giải quyết bằng phương pháp phân tích phương sai (xem chương IX). Còn nếu các dấu hiệu không phân phối chuẩn thì việc kiểm định có thể tiến hành bằng các phương pháp kiểm định phi tham số (xem mục 3 của chương VIII).

2.4. Kiểm định giả thuyết về tham số p của biến ngẫu nhiên phân phối không - một

Giả sử trong tổng thể nghiên cứu biến ngẫu nhiên gốc X phân phối không - một với tham số là p. Như đã thấy ở mục 2 chương VII, p chính là cơ cấu của tổng thể theo dấu hiệu nghiên cứu. Nếu chưa biết p song có cơ sở giả thiết rằng giá trị của nó bằng p_0, ta đưa ra giả thuyết thống kê

$$H_0: p = p_0$$

Từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

Để kiểm định H_0 ta xét các trường hợp sau:

1. Nếu n và p thỏa mãn điều kiện

$$n > 5 \text{ và } \frac{\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right|}{\sqrt{n}} < 0,3$$

thì lập thống kê

$$G = U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}, \qquad (8.52)$$

Từ mục 6 chương VI ta đã biết U phân phối xấp xỉ N (0,1). Do đó với mức ý nghĩa α và tùy thuộc vào giả thuyết đối H_1, các miền bác bỏ được xác định như sau:

a) H_0: $p = p_0$; H_1: $p > p_0$

$$W_\alpha = \left\{ U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}; U > u_\alpha \right\} \qquad (8.53)$$

b) H_0: $p = p_0$; H_1: $p < p_0$

$$W_\alpha = \left\{ U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}; U < -u_\alpha \right\} \qquad (8.54)$$

c) H_0: $p = p_0$; H_1: $p \neq p_0$

$$W_\alpha = \left\{ U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}; |U| > u_{\alpha/2} \right\} \qquad (8.55)$$

Với mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ tìm được giá trị quan sát U_{qs} của tiêu chuẩn kiểm định, so sánh với W_α và kết luận.

Thí dụ 12. Tỷ lệ khách hàng tiêu dùng một loại sản phẩm ở địa phương A là 60%. Sau một chiến dịch quảng cáo người ta muốn đánh giá xem chiến dịch quảng cáo này có thực sự mang lại hiệu quả hay không. Để làm điều đó người ta đã phỏng vấn ngẫu nhiên 400 khách hàng thì thấy có 250 người tiêu dùng loại sản phẩm nói trên. Với mức ý nghĩa 0,05 hãy kết luận về hiệu quả của chiến dịch quảng cáo đó.

Giải. Gọi p là tỷ lệ khách hàng tiêu dùng loại sản phẩm đó ở địa phương A. Đây là bài toán kiểm định tham số p của phân phối A(p). Cặp giả thuyết thống kê có dạng:

$$H_0: p = 0,6;\ H_1: p > 0,6$$

vì $n > 5$ và $\dfrac{\left|\sqrt{\dfrac{0,6}{0,4}} - \sqrt{\dfrac{0,4}{0,6}}\right|}{\sqrt{400}} = 0,02 < 0,3$

nên ta dùng công thức kiểm định (8.53)

Với $\alpha = 0,05 \Rightarrow u_\alpha = u_{0.05} = 1,645$

Vậy miền bác bỏ là $(1,645; +\infty)$

Với $f = \dfrac{250}{400} = 0,625$ ta có:

$$U_{qs} = \frac{(0,625 - 0,6)\sqrt{400}}{\sqrt{0,6.0,4}} = 1,02$$

$U_{qs} \notin W_\alpha$ nên chưa có cơ sở bác bỏ H_0, tức là chưa thể nói chiến dịch quảng cáo có hiệu quả.

Việc tìm xác suất mắc sai lầm loại 2 trong các trường hợp kiểm định một phía hoặc hai phía, tìm kích thước mẫu n theo α; β và $p_1 - p_0 < \Delta$ cho trước cũng như việc xác định giá trị p cũng được tiến hành giống như đã làm ở các mục trước.

Thí dụ 13. Tiếp tục thí dụ 12

a) Tìm xác suất mắc sai lầm loại 2 nếu $p_1 = 0,65$.

b) Tìm kích thước mẫu cần điều tra để với các điều kiện như ở câu a, xác suất mắc sai lầm loại 2 chỉ là 0,1.

Giải: a) Vì là kiểm định bên phải nên ta có công thức tính β như sau:

$$\beta = P\left[U < u_\alpha - \frac{|p_0 - p_1|}{Se(f)}\right]$$

ta có $Se(f) = \sqrt{\frac{f(1-f)}{n}} = \sqrt{\frac{0,625.0,375}{400}} = 0,0242$

và với $\alpha = 0,05 \Rightarrow u_\alpha = u_{0,05} = 1,645$

do đó

$$\beta = P\left[U < 1,645 - \frac{|0,6 - 0,65|}{0,0242}\right]$$

$$= P[U < -0,421] = P[U > 0,421] = 0,3372$$

b) Theo công thức tìm kích thước mẫu khi kiểm định một phía:

$$n \geq \left[\frac{2f(1-f)}{\Delta^2}(u_\alpha + u_\beta)^2\right]$$

với $\beta = 0,1 \rightarrow u_{0,1} = 0,46$

$$n \geq \left[\frac{2.0,625.0,375}{(0,65 - 0,6)^2}(1,645 + 0,46)^2\right] = 830,8 \Rightarrow n = 831$$

Như vậy, để giảm β từ 0,3372 xuống còn 0,1 thì phải điều tra thêm 431 khách hàng nữa.

2. Trường hợp mẫu nhỏ

Nếu kích thước mẫu nhỏ và không thể dùng quy luật chuẩn để kiểm định thì từ mục 6 chương VI ta đã biết lúc đó

tần suất mẫu f sẽ phân phối theo quy luật nhị thức với biểu thức xác suất là:

$$P_n(x) = C_n^x p^x (1-p)^{n-x} \quad f = 0, \frac{1}{n}, \frac{2}{n} \ldots 1 \qquad (8.56)$$

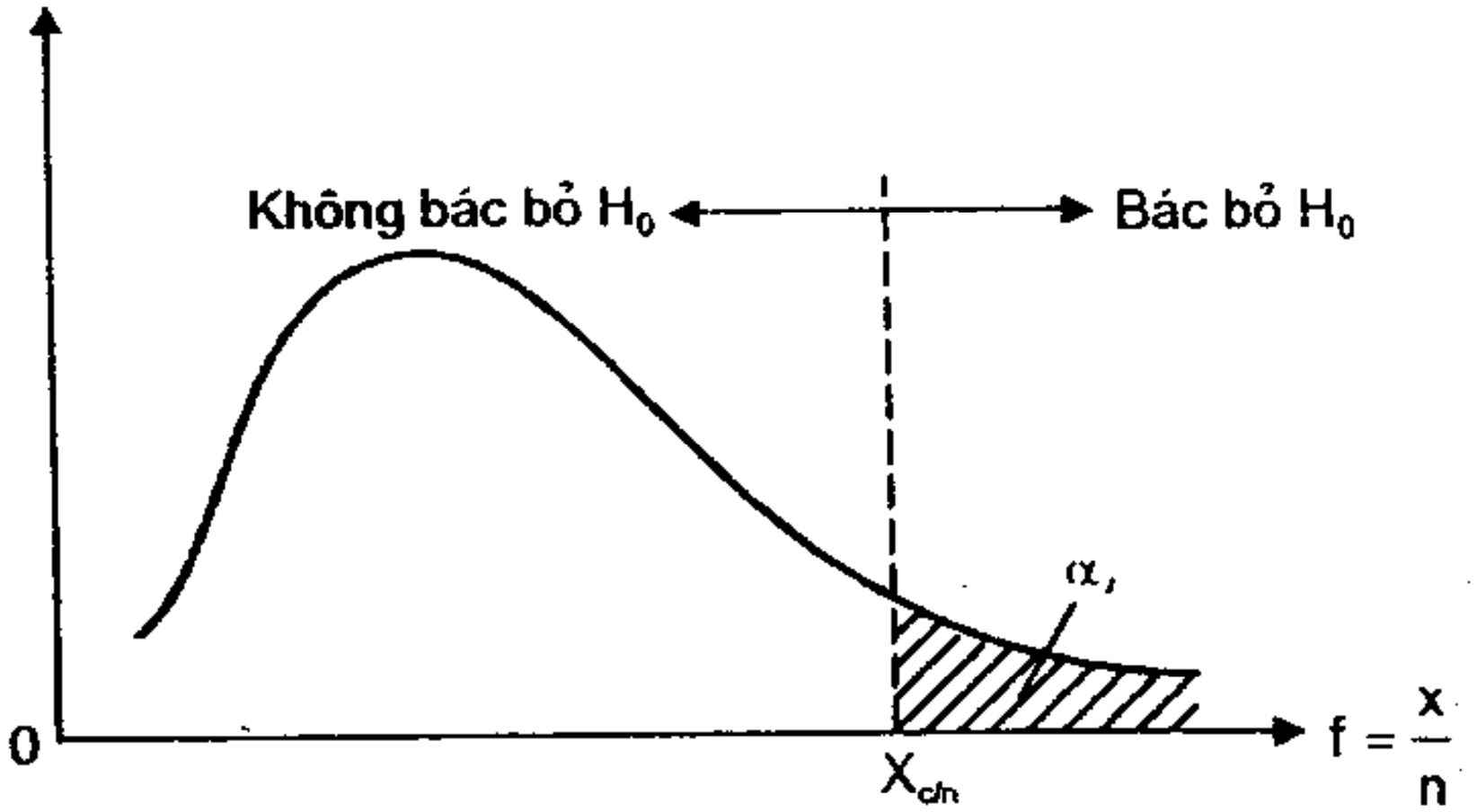
Với điều kiện giả thuyết đối H_1 là đúng thì biểu thức xác suất của f có dạng:

$$P_n(x) = C_n^x p_0^x (1-p_0)^{n-x}$$

Lúc đó tùy thuộc vào giả thuyết đối H_1, miền bác bỏ giả thuyết với mức ý nghĩa α được xác định như sau:

a) H_0: $p = p_0$; H_1: $p > p_0$

Miền bác bỏ bên phải sẽ được mô tả trên hình vẽ như sau (hình 8.3a)



Hình 8.3a

Trong đó $\frac{x_c}{n}$ là giá trị tới hạn. Ở đây, ta dùng dạng liên tục chỉ nhằm mục đích minh họa. Từ đó:

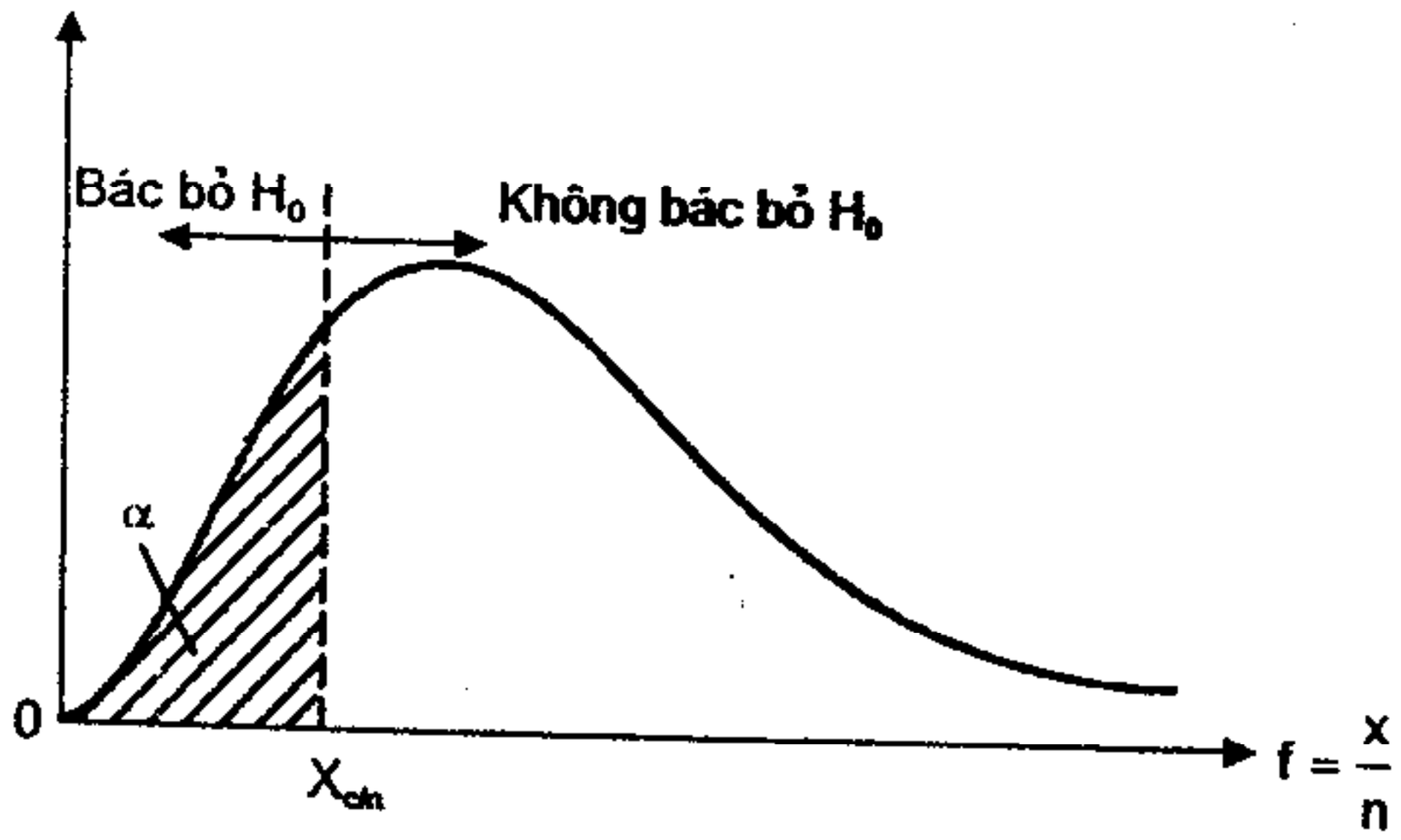
$$P(G \in W_\alpha / H_0) = \sum_{\frac{x}{n} \le \frac{x_c}{n}}^{1} C_n^x p_0^x (1-p_0)^{n-x}$$

$$= \sum_{x \le x_c}^{n} C_n^x p_0^x (1-p_0)^{n-x} = \alpha \qquad (8.57)$$

Vậy giả thuyết H_0 bị bác bỏ nếu $x \ge x_0$. Với n và p_0 cho trước có thể tìm được giá trị nguyên nhỏ nhất x_0 (bằng cách tra bảng phân phối nhị thức) đảm bảo xác suất mắc sai lầm loại một $\le \alpha$.

b) $H_0: p = p_0$; $H_1: p < p_0$

Trên đồ thị miền bác bỏ có dạng sau (hình 8.3b).



Hình 8.3b

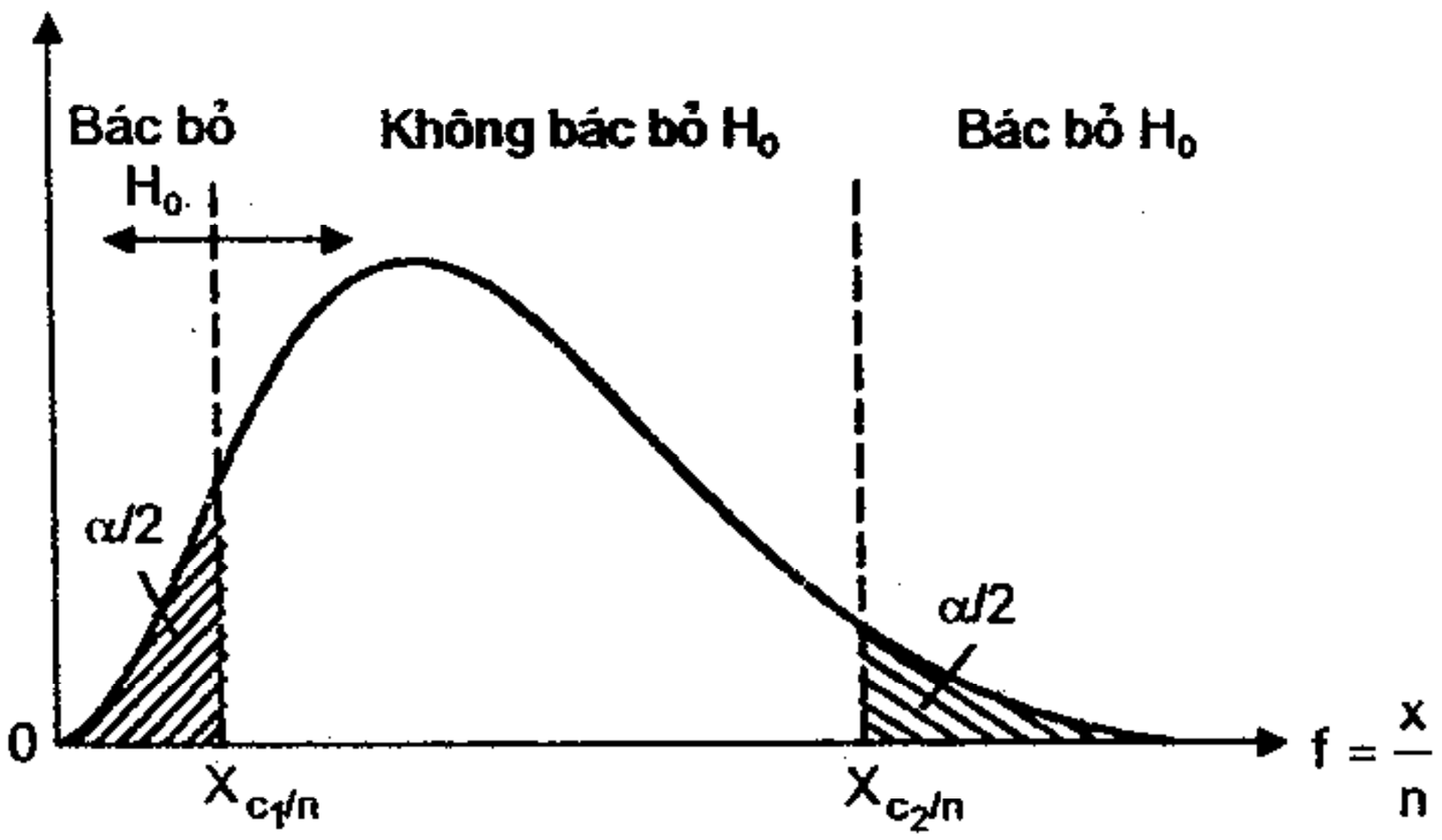
Lúc đó giá trị tới hạn là $x \le x_c$ với x_c là giá trị nguyên lớn nhất thỏa mãn:

$$P(G \in W_\alpha / H_0) = \sum_{x=0}^{x_c} C_n^x p_0^x (1-p_0)^{n-x} = \alpha \qquad (8.58)$$

c) H_0: $p = p_0$; H_1: $p \neq p_0$

Trên đồ thị miền bác bỏ có dạng hình 8.3c.

Lúc đó giá trị tới hạn là giá trị $x \leq x_{c_1}$ với x_{c_1} là giá trị nguyên lớn nhất và giá trị $x \geq x_{c_2}$ với x_{c_2} là giá trị nguyên nhỏ nhất thỏa mãn:



Hình 8.3c

$$P(G \in W_\alpha / H_0) = \sum_{x=0}^{x_{c_1}} C_n^x p_0^x (1-p_0)^{n-x} + \sum_{x=x_{c_2}}^{n} C_n^x p_0^x (1-p_0)^{n-x}$$

$$= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \qquad (8.59)$$

Thí dụ 14. Một máy gia công một loại chi tiết có tỷ lệ phế phẩm là 10%. Kiểm tra ngẫu nhiên 15 chi tiết thấy có 6 phế phẩm. Vậy, với mức ý nghĩa 0,05 có thể cho rằng tỷ lệ phế phẩm của máy đó đã tăng lên không?

Giải: Gọi p là tỷ lệ phế phẩm của máy đó. Cặp giả thuyết thống kê là:

$$H_0: p = 0,1;\ H_1: p > 0,1$$

Vì mẫu nhỏ và

$$\frac{\left|\sqrt{\begin{matrix}0,1\\0,9\end{matrix}} - \sqrt{\begin{matrix}0,9\\0,1\end{matrix}}\right|}{\sqrt{15}} = 0,69 > 0,3$$

nên ta dùng công thức (8.57) để kiểm định. Miền bác bỏ được xác định theo bảng sau với $n = 15$ và $p = 0,1$ (xem phụ lục 1).

x	$P_n(X)$	$P(X \geq x)$	x	$P_n(X)$	$P(X \geq x)$
0	0,2059	1	6	0,0019	0,0022
1	0,3432	0,7941	7	0,0003	0,0003
2	0,2669	0,4509	8	0,0000	0
3	0,1285	0,1840	.	.	.
4	0,0428	0,0555	.	.	.
5	0,0105	0,0127	15	0,0000	0

Với $x = 4 \Rightarrow P(x \geq 4) = 0,0555$

Với $x = 5 \Rightarrow P(x \geq 5) = 0,0127$

Để đảm bảo $\alpha \leq 0,05$ thì $x_\alpha = 5$.

Do x trong mẫu bằng $6 > 5$ nên bác bỏ H_0, thừa nhận H_1 tức là thừa nhận tỷ lệ phế phẩm của máy đó có tăng lên.

2.5. Kiểm định giả thuyết về hai tham số p của hai biến ngẫu nhiên phân phối A (p)

Giả sử có hai tổng thể nghiên cứu, trong đó các biến ngẫu nhiên X_1 và X_2 cùng phân phối không - một với các

tham số tương ứng là p_1 và p_2. Nếu p_1 và p_2 chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau, ta đưa ra giả thuyết thống kê:

$$H_0: p_1 = p_2$$

Để kiểm định giả thuyết trên, từ các tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước n_1 và n_2.

$$W_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$$

$$W_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$$

Và chọn lập tiêu chuẩn kiểm định là thống kê:

$$G = U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Từ mục 6 chương VI ta đã biết thống kê U phân phối xấp xỉ chuẩn hóa nếu $n_1 > 30$ và $n_2 > 30$

Nếu giả thuyết H_0 là đúng ($p_1 = p_2 = p$) thì tiêu chuẩn kiểm định trở thành:

$$G = U = \frac{f_1 - f_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

thông thường p chưa biết nên được thay bằng ước lượng của nó là:

$$\bar{f} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

Như vậy, ta có tiêu chuẩn kiểm định:

$$G = U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (8.60)$$

phân phối xấp xỉ N(0,1) nếu $n_1 > 30$ và $n_2 > 30$. Do đó tùy thuộc vào giả thuyết đối H_1, các miền bác bỏ mức α được xác định như sau:

a) H_0: $p_1 = p_2$; H_1: $p_1 > p_2$

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}};\ U > u_\alpha \right\} \qquad (8.61)$$

b) H_0: $p_1 = p_2$; H_1: $p_1 < p_2$

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}};\ U < -u_\alpha \right\} \qquad (8.62)$$

c) H_0: $p_1 = p_2$; H_1: $p_1 \neq p_2$

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}};\ |U| > u_{\alpha/2} \right\} \qquad (8.63)$$

Với hai mẫu cụ thể ta tính được các giá trị cụ thể của f_1, f_2 và $\bar{f}$ và giá trị quan sát U_{qs} của tiêu chuẩn kiểm định và so sánh với W_α để kết luận.

Thí dụ 15: Kiểm tra ngẫu nhiên các sản phẩm cùng loại do hai nhà máy sản xuất thu được các số liệu sau:

Bảng 8.4

Nhà máy	Số sản phẩm được kiểm tra	Số phế phẩm
A	$n_1 = 1000$	$x_1 = 20$
B	$n_2 = 900$	$x_2 = 30$

Với mức ý nghĩa $\alpha = 0,05$ có thể coi tỷ lệ phế phẩm của hai nhà máy là như nhau hay không?

Giải. Gọi p_1 và p_2 tương ứng là tỷ lệ phế phẩm của hai nhà máy A và B. Như vậy, đây là bài toán kiểm định cặp giả thuyết H_0: $p_1 = p_2$ và H_1: $p_1 \neq p_2$ với n_1 và n_2 khá lớn.

Theo công thức (8.63) tiêu chuẩn kiểm định là:

$$U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{1000} + \frac{1}{900}\right)}}$$

Do $\alpha = 0,05 \rightarrow u_{\alpha/2} = u_{0,025} = 1,96$ nên miền bác bỏ là $(-\infty; -1,96)$ và $(1,96; +\infty)$

Với hai mẫu cụ thể ta tìm được:

$$f_1 = \frac{20}{1000} = 0,02; \quad f_2 = \frac{30}{900} = 0,033$$

$$\bar{f} = \frac{20 + 30}{1000 + 900} = 0,0263$$

$$U_{qs} = \frac{0,02 - 0,033}{\sqrt{0,0263.0,9737\left(\frac{1}{1000} + \frac{1}{900}\right)}} = -1,81$$

$U_{qs} \notin W_\alpha$ vậy chưa có cơ sở để bác bỏ H_0, tức là có thể coi tỷ lệ phế phẩm ở hai nhà máy là như nhau.

Dùng Stata ta có thể ước lượng khoảng tin cậy mức 95% cho tỷ lệ phế phẩm của hai nhà máy A và B và kiểm định với mức ý nghĩa 0,05 xem, chẳng hạn, tỷ lệ phế phẩm của nhà máy A có bằng 0,025 hay không. Kết quả thu được như sau:

```
cii 1000 20

                                                    --Binomial Exact--
Variable |   Obs        Mean    Std.Err.     [95% Conf. Interval]
---------+--------------------------------------------------------
         |  1000          .02    .0044272    .0122583    .0307199

cii 900 30

                                                    --Binomial Exact--
Variable |   Obs        Mean    Std.Err.     [95% Conf. Interval]
---------+--------------------------------------------------------
         |  1000          .02    .0044272    .0122583    .0307199

bitesti 1000 20 0.025

         N  Observed k  Expected k  Assumed p  Observed p
------------------------------------------------------------
      1000          20          25    0.02500     0.02000

Pr(k >= 20)                 = 0.869568 (one-sided test)
Pr(k <= 20)                 = 0.182210 (one-sided test)
Pr(k <= 20 or k >= 30)      = 0. 361544 (two-sided test)
```

Việc mở rộng thủ tục kiểm định cho k (k > 2) tỷ lệ p của k tổng thể phân phối không - một sẽ được tiến hành bằng phương pháp kiểm định phi tham số.

2.6. Kiểm định giả thuyết về phương sai của biến ngẫu nhiên phân phối chuẩn

Giả sử trong tổng thể biến ngẫu nhiên gốc X phân phối $N(\mu, \sigma^2)$ với σ^2 chưa biết song có cơ sở để giả thiết rằng giá trị của nó bằng σ_0^2. Người ta đưa giả thuyết thống kê H_0: $\sigma^2 = \sigma_0^2$.

Để kiểm định giả thuyết trên từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$W = (X_1, X_2, ..., X_n)$$

và chọn tiêu chuẩn kiểm định là thống kê:

$$G = \chi^2 = \frac{(n-1)S^2}{c_0^2} \qquad (8.64)$$

Nếu giả thuyết H_0 đúng thì theo mục §6 Chương VI thống kê χ^2 phân phối theo quy luật "khi bình phương" với (n – 1) bậc tự do. Do đó với mức ý nghĩa α cho trước và tùy thuộc vào dạng của giả thuyết đối H_1 miền bác bỏ W_α được xây dựng theo các trường hợp sau:

a) H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 > \sigma_0^2$

$$W_\alpha = \left\{\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}; \chi^2 > \chi_\alpha^{2(n-1)}\right\} \qquad (8.65)$$

b) H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 < \sigma_0^2$;

$$W_\alpha = \left\{\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}; \chi^2 < \chi_{1-\alpha}^{2(n-1)}\right\} \qquad (8.66)$$

c) H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 \neq \sigma_0^2$

$$W_\alpha = \left\{\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}; \chi^2 < \chi_{1-\alpha/2}^{2(n-1)} \text{ hoặc } \chi^2 > \chi_{\alpha/2}^{2(n-1)}\right\} \qquad (8.67)$$

Với mẫu cụ thể ta tìm được giá trị cụ thể s^2 và giá trị quan sát χ^2_{qs} của tiêu chuẩn kiểm định, so sánh với W_α và kết luận.

Thí dụ 16: Để kiểm tra độ chính xác của một máy người ta đo ngẫu nhiên kích thước của 15 chi tiết do máy đó sản xuất và tính được $s^2 = 14,6$. Với mức ý nghĩa $\alpha = 0,01$ hãy kết luận máy móc có hoạt động bình thường không, biết rằng kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn có dung sai theo thiết kế là $\sigma^2 = 12$.

Giải: Gọi X là kích thước chi tiết, theo giả thiết X phân phối chuẩn. Vậy đây là bài toán kiểm định cặp giả thuyết:

$$H_0: \sigma^2 = 12$$

$$H_1: \sigma^2 > 12$$

Từ (8.65) tiêu chuẩn kiểm định có dạng:

$$\chi^2 = \frac{14.S^2}{12}$$

$$\text{Do } \alpha = 0,01 \Rightarrow \chi^{2(14)}_{\alpha} = \chi^{2(14)}_{0,01} = 29,14$$

nên miền bác bỏ có dạng $(29,14; +\infty)$

Với mẫu cụ thể ta có giá trị quan sát của tiêu chuẩn kiểm định:

$$\chi^2_{qs} = \frac{14.14,6}{12} = 17,033$$

$\chi^2_{qs} \notin W_\alpha$, vậy chưa có cơ sở để bác bỏ H_0, hay có thể nói máy móc vẫn làm việc bình thường.

2.7. Kiểm định giả thuyết về sự bằng nhau của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn

Giả sử có hai tổng thể nghiên cứu trong đó biến ngẫu nhiên gốc X_1 trong tổng thể thứ nhất phân phối $N(\mu_1; \sigma_1^2)$ và biến ngẫu nhiên gốc X_2 trong tổng thể thứ hai phân phối $N(\mu_2; \sigma_2^2)$. Nếu σ_1^2 và σ_2^2 chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau thì ta đưa ra giả thuyết thống kê H_0: $\sigma_1^2 = \sigma_2^2$.

Để kiểm định giả thuyết trên từ hai tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước tương ứng là n_1 và n_2:

$$W_1 = (X_{11}, X_{12},..., X_{1n_1})$$

$$W_2 = (X_{21}, X_{22},..., X_{2n_2})$$

và chọn tiêu chuẩn kiểm định là thống kê:

$$G = F = \frac{S_1^2.\sigma_2^2}{S_2^2.\sigma_1^2} \qquad (8.68)$$

nếu $S_1^2 > S_2^2$. Từ mục 6 Chương VI ta biết thống kê F phân phối theo quy luật Fisher-Snedecor với $(n_1 - 1)$ và $(n_2 - 1)$ bậc tự do.

Nếu giả thuyết H_0 đúng thì tiêu chuẩn kiểm định có dạng:

$$F = \frac{S_1^2}{S_2^2} \qquad (S_1^2 > S_2^2)$$

và vẫn phân phối $F(n_1 - 1, n_2 - 1)$. Do đó với mức ý nghĩa α cho trước và tùy thuộc vào dạng của giả thuyết đối H_1 có thể xây dựng được các miền bác bỏ W_α tương ứng sau:

a) H_0: $\sigma_1^2 = \sigma_0^2$; H_1: $\sigma_1^2 > \sigma_2^2$

$$W_\alpha = \left\{ F = \frac{S_1^2}{S_2^2};\ F > f_\alpha^{(n_1-1,\, n_2-1)} \right\} \qquad (8.69)$$

b) H_0: $\sigma_1^2 = \sigma_0^2$; H_1: $\sigma_1^2 \neq \sigma_2^2$

$$W_\alpha = \left\{ F = \frac{S_1^2}{S_2^2};\ F < f_{1-\alpha/2}^{(n_1-1,\, n_2-1)} \text{ hoặc } F > f_{\alpha/2}^{(n_1-1,\, n_2-1)} \right\} \qquad (8.70)$$

Với hai mẫu cụ thể tìm được s_1^2, s_2^2 và giá trị quan sát F_{qs} của tiêu chuẩn kiểm định, so sánh với W_α và kết luận.

Thí dụ 17: Có hai giống lúa có năng suất trung bình xấp xỉ như nhau song mức độ phân tán về năng suất có thể khác nhau. Để kiểm tra điều đó người ta gặt mẫu tại 2 vùng trồng hai giống lúa đó và thu được kết quả sau:

Bảng 8.5

Giống lúa	Số điểm gặt	Phương sai
A	$n_1 = 41$	$s_1^2 = 11{,}41$
B	$n_2 = 30$	$s_2^2 = 6{,}52$

Với mức ý nghĩa $\alpha = 0{,}05$ hãy kết luận về vấn đề trên, biết năng suất lúa là biến ngẫu nhiên phân phối chuẩn.

Giải. Gọi X_1 và X_2 là năng suất của 2 giống lúa A và B, X_1 và X_2 phân phối chuẩn. Đây là bài toán kiểm định cặp giả thuyết:

$$H_0: \sigma_1^2 = \sigma_2^2;\ H_1: \sigma_1^2 \neq \sigma_2^2$$

Theo (8.70) tiêu chuẩn kiểm định có dạng $F = \frac{S_1^2}{S_2^2}$

Do $\alpha = 0,05$ nên $f_{\alpha/2}^{(40,29)} = f_{0,025}^{(40,29)} = 2,03$;

$$f_{1-\alpha/2}^{(40,29)} = \frac{1}{f_{\alpha/2}^{(29,40)}} = \frac{1}{1,94} = 0,52$$

Vậy miền bác bỏ là (0; 0,52) và (2,03; +∞)

Với mẫu cụ thể ta có:

$$F_{qs} = \frac{11,41}{6,52} = 1,75$$

$F_{qs} \notin W_\alpha$ vậy chưa có cơ sở bác bỏ H_0 hay độ phân tán của năng suất hai giống lúa trên là như nhau. Sau đây là kết quả giải bằng Stata:

```
sdtesti      41. 3.38          30. 2.55
Two - sample test of variance          x: Number of obs = 41
                                       y: Number of obs = 30
```

Variable	Mean	Std. Err.	t	P > |t|	[95% Conf. Interval]
x	.	.5278673	.	.	. .
y	.	.4655642	.	.	. .
combined	.	.3630034	.	.	. .

```
Ho: sd(x) = sd(y)
F Observed = F  = F(29, 40) = 1.757
F Lower tail = F_L = F(29,40) = 0.569
F Upper tail = F_U = F(29,40) = 1.757
Ha: s1 < s2          Ha: s1 ~= s2                    Ha: s1 > s2
P<F = 0.9510         P<F_L+P>F_U=0.1072              P>F=0.0490
```

2.8. Kiểm định K phương sai của K biến ngẫu nhiên phân phối chuẩn

Giả sử có k tổng thể nghiên cứu, trong đó các biến ngẫu nhiên $X_1, X_2,..., X_k$ cùng phân phối chuẩn. Từ các tổng thể rút ra k mẫu độc lập kích thước $n_1, n_2,...,n_k$ và tìm được các phương sai mẫu tương ứng $S_1^2, S_2^2,..., S_k^2$.

Với mức ý nghĩa α phải kiểm định cặp giả thuyết:

H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

H_1: Có ít nhất hai phương sai khác nhau.

Như vậy, cần kiểm định xem sự khác biệt của các phương sai mẫu là có ý nghĩa hay không

Để chọn lập tiêu chuẩn kiểm định ta xét hai trường hợp sau:

1. Nếu các kích thước mẫu $n_1, n_2..., n_k$ khác nhau

Lúc đó nếu ký hiệu $\overline{S}^2$ là trung bình số học của các phương sai mẫu:

$$\overline{S}^2 = \frac{\sum_{i=1}^{n} h_i S_i^2}{h}$$

trong đó $h = \sum_{i=1}^{k} h_i$ với $h_i = n_i - 1$ tức là số bậc tự do của phương sai mẫu S_i^2 thì tiêu chuẩn kiểm định là thống kê Bartlett sau đây:

$$B = \frac{V}{C} \qquad (8.71)$$

trong đó $V = 2{,}303[h.\lg \overline{S}^2 - \sum_{i=1}^{u} h_i.\lg S_i^2]$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{u}\frac{1}{h_i} - \frac{1}{h}\right]$$

Với điều kiện giả thuyết H_0 là đúng thì thống kê B phân phối xấp xỉ khi bình phương với k – 1 bậc tự do nếu mọi $h_i > 2$ tức là mọi $n_i > 3$.

Lúc đó miền bác bỏ mức α được xác định bằng biểu thức:

$$W_\alpha = \left\{B = \frac{V}{C}; B > \chi_\alpha^{2(k-1)}\right\} \qquad (8.72)$$

Thí dụ 18: Từ 4 mẫu độc lập kích thước $n_1 = 10$, $n_2 = 12$, $n_3 = 15$, $n_4 = 16$ được rút ra từ các tổng thể phân phối chuẩn tìm được các phương sai mẫu tương ứng bằng 0,25; 0,4; 0,36; 0,46. Với mức ý nghĩa 0,05 hãy kiểm định xem phương sai của các tổng thể có bằng nhau hay không.

Giải. Cặp giả thuyết có dạng

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

H_1: Có ít nhất 2 phương sai khác nhau.

Theo công thức kiểm định (8.72) ta có với

$$\alpha = 0{,}05 \rightarrow \chi_\alpha^{2(k-1)} = \chi_{0{,}05}^{2(3)} = 7{,}8$$

Vậy miền bác bỏ là (7,8; +∞)

Để tìm B_{qs} ta lập bảng tính sau:

Mẫu i	Kích thước n_i	Số bậc tự do h_i	Phương sai s_i^2	$h_i s_i^2$	$\lg s_i^2$	$h_i \lg s_i^2$	$\frac{1}{h_i}$
1	10	9	0,25	2,25	−1,3979	−12,5811	0,11
2	13	12	0,4	4,8	−19,2252	−19,2252	0,08
3	15	14	0,36	5,04	−21,7822	−21,7822	0,07
4	16	15	0,46	6,9	−24,942	−24,942	0,07
Σ				18,99		−78,53	0,32

Từ đó

$$\bar{s}^2 = \frac{\sum h_i s_i^2}{h} = \frac{18,99}{50} = 0,3798$$

$$\Rightarrow \lg \bar{s}^2 = -1,5795$$

$$V = 2,303\,[h \lg \bar{s}^2 - \sum h_i \lg s_i^2]$$

$$= 2,303[50(-1,5795) + 78,5350] = -1,01$$

$$C = 1 + \frac{1}{3(4-1)}\left[0,32 - \frac{1}{50}\right] = 1,03$$

$$\Rightarrow B = -0,98$$

$B \notin W_\alpha$ nên chưa thể bác bỏ H_0, tức là các phương sai của tổng thể có thể xem như bằng nhau với mức ý nghĩa 0,05.

Sau đây là kết quả giải bằng Stata cho thí dụ A

```
oneway x v, tabulate
```

Summary of Thu nhap ca nhan

v	Mean	Std. Dev.	Freq.
1	1649.73	47.374897	100
2	1655.99	47.408284	100
3	1628.12	48.399699	100
Total	1644.6133	49.050351	300

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	42763.8867	2	21381.9433	9.39	0.0001
Within groups	676611.26	297	2278.15239		
Total	719375.147	299	2405.93695		

Bartlett's test for equal variances: chi2(2) = 0.0585

Prob>chi2 = 0.971

Nếu chấp nhận H_0 thì có thể lấy $\bar{s}^2$ để ước lượng phương sai đồng đều của các tổng thể. Chẳng hạn, trong bài toán trên có thể ước lượng phương sai của các tổng thể là:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 \approx \bar{s}^2 = 0{,}3798$$

2. Nếu các kích thước mẫu $n_1 = n_2 = \ldots = n_k = n$

Trong trường hợp này ta có thể sử dụng tiêu chuẩn kiểm định Bartlett đã xét ở trên, tuy nhiên do phân phối xác suất của tiêu chuẩn Bartlett chỉ là xấp xỉ nên kết quả kiểm định không thật đáng tin cậy. Với trường hợp kích thước mẫu bằng nhau thì ta dùng tiêu chuẩn kiểm định Cochran sẽ cho kết quả tin cậy hơn. Tiêu chuẩn Cochran là thống kê:

$$G = \frac{S^2_{max}}{S_1^2 + S_2^2 + ... + S_k^2} \qquad (8.73)$$

tức là tỉ số giữa phương sai mẫu lớn nhất và tổng các phương sai mẫu. Phân phối xác suất của thống kê G chỉ phụ thuộc vào số bậc tự do (n – 1) và số lượng mẫu k. Lúc đó, miền bác bỏ mức α được xác định bằng biểu thức:

$$W_\alpha = \left\{ G = \frac{S^2_{max}}{S_1^2 + S_2^2 + ... + S_k^2}; G > g_\alpha^{[n-1,k]} \right\} \qquad (8.74)$$

các giá trị tới hạn $g_\alpha^{[n-1,k]}$ được tìm trong bảng tính sẵn (Phụ lục 13). Với k mẫu cụ thể tìm được G_{qs} so sánh với W_α và kết luận.

Thí dụ 19. Từ 4 mẫu độc lập kích thước như nhau và bằng 17 rút ra từ các tổng thể phân phối chuẩn tìm được các phương sai mẫu tương ứng là 0,26; 0,36; 0,40; 0,42.

a) Với mức ý nghĩa 0,05 hãy kiểm định sự bằng nhau của các phương sai tổng thể.

b) Hãy ước lượng các phương sai tổng thể nếu có thể cho rằng chúng bằng nhau.

Giải.

a) Ta tìm giá trị quan sát của tiêu chuẩn Cochran:

$$G_{qs} = \frac{0,42}{0,26 + 0,36 + 0,40 + 0,42} = 0,2917$$

Với $\alpha = 0,05$, số bậc tự do $17 - 1 = 16$ và số lượng mẫu là 4 ta tìm được giá trị tới hạn là:

$$g_{\alpha}^{[n-1,k]} = g_{0,05}^{(16,4)} = 0,4366$$

vậy miền bác bỏ là $(0,4366 ; +\infty)$

Do $G_{qs} \notin W_{\alpha}$ nên chưa thể bác bỏ H_0 hay có thể cho rằng các phương sai tổng thể bằng nhau.

b) Vì ta thừa nhận H_0 nên phương sai đồng đều của tổng thể có thể ước lượng bằng trung bình số học của các phương sai mẫu.

$$\sigma^2 \approx \bar{s}^2 = \frac{0,26 + 0,36 + 0,40 + 0,42}{4} = 0,36$$

§3. KIỂM ĐỊNH PHI THAM SỐ

Ở mục trước ta đã xét các kiểm định tham số. Đặc điểm chung của các kiểm định này là phải dựa trên một số điều kiện nào đó liên quan đến tổng thể nghiên cứu, chẳng hạn kiểm định T và kiểm định F phải dựa trên giả thiết là tổng thể phân phối chuẩn. Khi các giả thuyết thống kê phụ thuộc vào các giả thiết về tổng thể và các tham số của nó, ta có các kiểm định tham số.

Trong thực tế có nhiều trường hợp ta không đưa ra bất kỳ giả thiết nào liên quan đến tham số của tổng thể hoặc dạng phân phối xác suất của tổng thể. Lúc đó, các kiểm định

được xét ở mục trước không thể áp dụng được. Các kiểm định trong điều kiện như vậy gọi là **kiểm định phi tham số.**

Như vậy *kiểm định phi tham số* **là các thủ tục thống kê** để kiểm định giả thuyết khi không có được các giả thiết liên quan đến tham số của tổng thể hay dạng phân phối xác suất của tổng thể.

Sau đây ta sẽ xem xét một số kiểm định phi tham số quan trọng nhất.

3.1. Kiểm định khi bình phương

Kiểm định khi bình phương dựa trên việc sử dụng tiêu chuẩn χ^2 của K.Pearson là một trong những kiểm định phi tham số quan trọng nhất. Nó thường được sử dụng để giải quyết các bài toán sau đây trong thực tế.

1. Kiểm định giả thuyết về tính độc lập của hai dấu hiệu định tính

Giả sử cần nghiên cứu đồng thời hai dấu hiệu định tính A và B trên cùng một tổng thể. Dấu hiệu A có các phạm trù là $A_1, A_2,..., A_l$ còn dấu hiệu B có các phạm trù là $B_1, B_2,..., B_k$. Nếu có cơ sở để giả thiết rằng A và B độc lập ta đưa ra cặp giả thuyết thống kê sau đây:

H_0: A và B độc lập

H_1: A và B phụ thuộc

Để kiểm định giả thuyết trên, từ tổng thể lập mẫu kích thước n và trình bày các số liệu mẫu dưới dạng *báng tiếp liên* sau đây:

A \ B	B_1	B_2	...	B_j	...	B_k	Tổng số n_i
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_1
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_i
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
A_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	n_h
Tổng số m_j	m_1	m_2	...	m_j	...	m_k	$\sum = n$

Trong đó:

n là kích thước mẫu;

n_i là tổng các tần số ứng với dấu hiệu thành phần A_i;

m_j là tổng các tần số ứng với dấu hiệu thành phần B_j;

n_{ij} là tần số ứng với các phần tử đồng thời mang dấu hiệu A_i và B_j.

Với n khá lớn thì theo định nghĩa thống kê về xác suất ta có:

$$P(A_iB_j) \approx \frac{n_{ij}}{n} \qquad i = \overline{1,h}; \qquad j = \overline{1,k}$$

$$P(A_i) \approx \frac{n_i}{n} \qquad i = \overline{1,h}$$

$$P(B_j) \approx \frac{n_j}{n} \qquad j = \overline{1,k}$$

Nếu giả thiết H_0 đúng tức là A và B độc lập thì các dấu hiệu thành phần cũng độc lập nên:

$$P(A_iB_j) = P(A_i).P(B_j)$$

Tức là: $\frac{n_{ij}}{n} = \frac{n_i}{n}.\frac{m_j}{n}$ $\quad i = \overline{1,h};\quad j = \overline{1,k}$

Vì thế tiêu chuẩn kiểm định giả thuyết có thể được chọn là thống kê sau:

$$G = \chi^2 = n\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{\left(\frac{n_{ij}}{n} - \frac{n_i}{n}.\frac{m_j}{n}\right)^2}{\frac{n_i}{n}\frac{m_j}{n}}$$

hay $$\chi^2 = n\left[\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i.m_j} - 1\right] \qquad (8.75)$$

Nếu giả thuyết H_0 đúng và n khá lớn thì ta có thể coi thống kê χ^2 phân phối theo quy luật "khi bình phương" với (h – 1) (k – 1) bậc tự do. Vì vậy, với mức ý nghĩa α miền bác bỏ của H_0 là:

$$W_\alpha = \left\{\chi^2 = n\left[\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i.m_j} - 1\right];\ \chi^2 > \chi_\alpha^{2(h-1)(k-1)}\right\} \qquad (8.76)$$

Dựa vào mẫu cụ thể tính được giá trị quan sát χ^2_{qs}, so sánh với W_α và kết luận.

Thí dụ 1. Nghiên cứu sự ảnh hưởng của hoàn cảnh gia đình đối với tình trạng phạm tội của trẻ em ở tuổi vị thành niên qua điều tra ngẫu nhiên 148 em nhỏ thu được kết quả sau:

Bảng 8.6

Hoàn cảnh gia đình / Tình trạng phạm tội	Bố hoặc mẹ đã chết	Bố mẹ li hôn	Còn cả bố mẹ	Tổng cộng n_i
Không phạm tội	20	25	18	63
Phạm tội	29	43	13	85
Tổng cộng m_j	49	68	31	$\sum = 148$

Với mức ý nghĩa = 0,05 có thể kết luận hoàn cảnh gia đình độc lập với tình trạng phạm tội của trẻ em hay không?

Giải. Đây là bài toán kiểm định tính độc lập của hai dấu hiệu định tính với cặp giả thuyết là H_0: Tình trạng phạm tội độc lập với hoàn cảnh gia đình; H_1: Tình trạng phạm tội phụ thuộc hoàn cảnh gia đình.

Tiêu chuẩn kiểm định là:

$$\chi^2 = 148\left[\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i.m_j} - 1\right]$$

Do $\alpha = 0,05 \Rightarrow \chi_{\alpha}^{2(h-1)(k-1)} = \chi_{0,05}^{2(2-1)(3-1)} = 5,991$

Qua mẫu cụ thể tính được $\chi_{qs}^2 = 4,0552$

Kết quả giải bằng Stata như sau:

.tab pt gd, chi2

Tình trạng phạm tội	Hoàn cảnh gia đình			Total
	1	2	3	
1	20	25	18	63
2	29	43	13	85
Total	49	68	31	148

Pearson chi2 (2) = 4.0433 Pr = 0.132

$\chi^2_{qs} \notin W_\alpha$, vậy chưa có cơ sở để bác bỏ H_0, tức là có thể xem hoàn cảnh gia đình và tình trạng phạm tội là độc lập nhau.

Phương pháp kiểm định khi bình phương vừa xét ở trên cho phép ta kết luận xem hai dấu hiệu định tính A và B có độc lập với nhau hay không. Nếu giả thuyết H_0 bị bác bỏ tức là A và B phụ thuộc (tương quan) thì sẽ nảy sinh nhu cầu đánh giá mức độ chặt chẽ của sự phụ thuộc đó. Vấn đề đánh giá mức độ chặt chẽ của sự phụ thuộc giữa hai dấu hiệu định tính, giữa hai hoặc nhiều dấu hiệu định lượng sẽ được giải quyết bằng phương pháp phân tích tương quan ở chương X.

2. Kiểm định giả thuyết về k tham số p của k biến ngẫu nhiên phân phối không - một

Khi mở rộng việc so sánh hai tham số p của phân phối A(p) cho trường hợp có nhiều hơn hai tổng thể thì có thể sử dụng thủ tục kiểm định khi bình phương như sau:

Giả sử có k tổng thể nghiên cứu trong đó các biến ngẫu nhiên $X_1, X_2,..., X_k$ cùng phân phối không - một với các tham số tương ứng là $p_1, p_2,..., p_k$. Nếu có cơ sở để giả thiết rằng giá trị của chúng bằng nhau thì ta đưa ra cặp giả thuyết sau:

H_0: $p_1 = p_2 = ... = p_k = p$

H_1: Có ít nhất hai giá trị khác nhau.

Để kiểm định cặp giả thuyết trên từ các tổng thể lập k mẫu ngẫu nhiên độc lập kích thước tương ứng là $n_1, n_2,..., n_k$ và giả thiết các số liệu mẫu được cho dưới dạng bảng sau (Bảng 8.7).

Bảng 8.7

Mẫu	Số lần xuất hiện biến cố	Số lần không xuất hiện biến cố	Tổng số n_i
1	X_{11}	X_{12}	n_1
.	.	.	.
i	X_{i1}	X_{i2}	n_i
.	.	.	.
k	X_{k1}	X_{k2}	n_k
Tổng số m_j	m_1	m_2	$\sum n_i = n$

Lúc đó nếu $n_1, n_2, ..., n_k$ đều lớn hơn 50 thì các thống kê:

$$U_i = \frac{X_{i1} - n_i p_i}{\sqrt{n_i p_i (1 - p_i)}} \qquad i = \overline{1,k}$$

sẽ phân phối xấp xỉ N(0.1). Vì vậy, thống kê:

$$\chi^2 = \sum_{i=1}^{k} U_i^2 = \sum_{i=1}^{k} \frac{(X_{i1} - n_i p_i)^2}{n_i p_i (1 - p_i)} \qquad (8.77)$$

sẽ phân phối khi bình phương với k bậc tự do. Nếu giả thuyết H_0 là đúng ($p_1 = p_2 = ... = p_k = p$) thì do p thường chưa biết nên có thể thay bằng ước lượng của nó là:

$$\bar{f} = \frac{X_{11} + X_{21} + ... + X_{k1}}{n_1 + n_2 + ... + n_k}$$

và tiêu chuẩn kiểm định (8.77) trở thành:

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_{i1} - n_i \bar{f})^2}{n_i \bar{f}(1 - \bar{f})} \qquad (8.78)$$

và χ^2 sẽ phân phối khi bình phương với (k – 1) bậc tự do. Mặt khác, có thể biến đổi (8.78) thành biểu thức tương đương sau đây:

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{2}\frac{X_{ij}^2}{n_i m_j} - 1\right] \qquad (8.79)$$

Do đó miền bác bỏ mức α được xác định bằng biểu thức:

$$W_\alpha = \left\{\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{2}\frac{X_{ij}^2}{n_i m_j} - 1\right];\ \chi^2 > \chi_\alpha^{2(k-1)}\right\} \qquad (8.80)$$

Từ các tổng thể lập k mẫu cụ thể và tìm được χ_{qs}^2, so sánh với W_α và kết luận.

Cần nhấn mạnh rằng để sử dụng tiêu chuẩn kiểm định trên thì phải đáp ứng các yêu cầu sau:

- Các mẫu điều tra là ngẫu nhiên và độc lập nhau.
- Kích thước mỗi mẫu tối thiểu là 50.
- Trong mỗi mẫu số lần xuất hiện biến cố ít nhất là 5.

Thí dụ 2. Trước khi đưa ra thị trường một loại sản phẩm với màu sơn mới, người ta muốn xét xem các lứa tuổi khác nhau phản ứng như thế nào với màu sắc sản phẩm đó. Một cuộc điều tra theo những nhóm lứa tuổi khác nhau về màu sắc sản phẩm mà doanh nghiệp sản xuất thu được kết quả sau:

	Thích	Không thích	∑
Dưới 25 tuổi	104	21	125
Từ 25 đến 35	122	28	150
Từ 36 đến 45	68	32	100
Từ 46 trở lên	41	34	75
∑	335	115	450

Với mức ý nghĩa $\alpha = 0,05$ có thể cho rằng tỷ lệ những người thích màu sắc đó của sản phẩm là như nhau đối với mọi lứa tuổi hay không?

Giải. Cặp giả thuyết thống kê:

H_0: $p_1 = p_2 = p_3 = p_4 = p$

H_1: Có ít nhất hai p_i khác nhau

Với $\alpha = 0,05 \rightarrow \chi^{2(k-1)}_{\alpha} = \chi^{2(3)}_{0,05} = 7,815$

Vậy miền bác bỏ là $(7,815; +\infty)$

Để tính giá trị quan sát của tiêu chuẩn kiểm định ta lập bảng tính sau:

Mẫu	Thích	Không thích	n_i
1	104 0,258	21 0,031	125
2	122 0,296	28 0,045	150
3	68 0,138	32 0,089	100
4	41 0,067	34 0,134	75
m_j	335	115	n = 450

Tại mỗi ô ta tính giá trị $\frac{x_{ij}^2}{n_i m_j}$ và ghi vào góc ô

Lúc đó $\sum_{i=1}^{4}\sum_{j=1}^{2}\frac{x_{ij}^2}{n_i m_j} = 0{,}258 + \ldots + 0{,}134 = 1{,}058$

Từ đó $\chi^2_{qs} = 450[1{,}058 - 1] = 26{,}1$

$\chi^2_{qs} \in W_\alpha$. Vậy với mức ý nghĩa 0,05 từ các mẫu đã cho bác bỏ H_0, thừa nhận H_1 tức là tỷ lệ khách hàng thích màu sản phẩm nói trên là khác nhau theo các lứa tuổi.

Việc tìm β, P-value cũng được tiến hành giống như ở các mục trước.

Chẳng hạn P-value = $P\left[\chi^2 > \left|\chi^2_{qs}\right|\right] = P[\chi^2 > 26{,}1]$

Với số bậc tự do bằng 3. Giá trị lớn nhất là 16,27 do đó P-value < 0,001.

Giải bằng Stata cho kết quả sau:

.tabi 104 122 68 41\21 28 32 34

row	col 1	2	3	4	Total
1	104	122	68	41	35
2	21	28	32	34	115
Total	125	150	100	75	450

Pearson chi2 (3) = 26.3821 Pr = 0.000

3. Kiểm định giả thuyết về quy luật phân phối xác suất của biến ngẫu nhiên

Giả sử chưa biết quy luật phân phối xác suất của biến ngẫu nhiên gốc X trong tổng thể nghiên cứu song có cơ sở để

giả thiết rằng X phân phối theo một quy luật A nào đó. Lúc đó đưa ra cặp giả thuyết thống kê sau:

H_0: X phân phối theo quy luật A

H_1: X không phân phối theo quy luật A

Để kiểm định cặp giả thuyết trên tiêu chuẩn khi bình phương được sử dụng như sau:

a) Nếu X là biến ngẫu nhiên rời rạc

Từ tổng thể rút ra một mẫu kích thước n trong đó biến ngẫu nhiên X có bảng phân phối tần số thực nghiệm sau đây:

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

$$\sum_{i=1}^{k} n_i = n$$

Nếu giả thuyết H_0 là đúng thì có thể tính được các xác suất lý thuyết để X nhận các giá trị tương ứng:

$$P_i = P(X = x_i) \qquad i = \overline{1,k}$$

Từ đó tần số lý thuyết của phân phối xác suất sẽ là:

$$n'_i = np \qquad (i = \overline{1,k})$$

và bảng phân phối tần số lý thuyết có dạng:

x_i	x_1	x_2	...	x_k
n'_i	n'_1	n'_2	...	n'_k

$$\sum_{i=1}^{k} n'_i = n$$

lúc đó tiêu chuẩn kiểm định được chọn là thống kê:

$$G = \chi^2 = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} \qquad (8.81)$$

Đây là dạng tương đương của (8.75)

Biến ngẫu nhiên χ^2 nói trên phân phối "khi bình phương"

với (k – r – 1) bậc tự do trong đó r là số tham số cần ước lượng của quy luật cần kiểm định. Các tham số này được ước lượng bằng phương pháp hợp lý tối đa. Chẳng hạn, nếu quy luật cần kiểm định là quy luật nhị thức hay Poisson thì r = 1.

Với mức ý nghĩa bằng α cho trước miền bác bỏ W_α được xác định bằng biểu thức:

$$W_\alpha = \left\{ \chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'};\ \chi^2 > \chi_\alpha^{2(k-r-1)} \right\} \qquad (8.82)$$

Với một mẫu cụ thể tính được giá trị quan sát của χ^2 và so sánh với miền bác bỏ W_α để kết luận.

- Nếu $\chi^2_{qs} \in W_\alpha$ thì bác bỏ giả thuyết về dạng phân phối A của biến ngẫu nhiên X.

- Nếu $\chi^2_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết về dạng phân phối A của X.

Thí dụ 3. Số lời gọi đến một trạm điện thoại (X) trong một phút được cho trong bảng sau:

Số lời gọi x_i	0	1	2	3	4	5	≥ 6
Số phút tương ứng n_i	17	22	26	20	11	2	2

Với mức ý nghĩa α = 0,01 có thể coi X phân phối theo quy luật Poisson được không?

Giải. Cặp giả thuyết thống kê là:

H_0: X phân phối theo quy luật Poisson

H_1: X không phân phối theo quy luật Poisson.

Từ mẫu trên tính được $\bar{x} = 2$ là ước lượng hợp lý tối đa của λ trong phân phối Poisson. Để tính giá trị quan sát χ^2 ta lập bảng tính toán sau:

x_i	n_i	$p_i = \frac{e^{-2}2^{x_i}}{x_i!}$	$n'_i = np_i$	$\frac{(n_i - n'_i)^2}{n'_i}$
0	17	0,1353	13,53	0,89
1	22	0,2707	27,07	0,95
2	26	0,2707	27,07	0,04
3	20	0,1804	18,04	0,21
4	11	0,0902	9,02	0,43
5	2	0,0361	3,61	0,72
≥ 6	2	0,0166	1,66	0,07
	n = 100	$\sum$ = 1.000		$\chi^2 = 3,31$

Do $\alpha = 0,01 \rightarrow \chi_\alpha^{2(k-r-1)} = \chi_{0,01}^{2(7-1-1)} = 15,1$ vậy miền bác bỏ là $(15,1; +\infty)$. $\chi_{qs}^2 \notin W_\alpha$ nên chưa có cơ sở bác bỏ H_0 hay có thể coi X phân phối theo quy luật Poisson với $\lambda = 2$.

Một bài toán kiểm định về quy luật phân phối xác suất rời rạc khá quan trọng trong thực thế là bài toán kiểm định sự phù hợp của quy luật đa thức. Ta sẽ xét chi tiết hơn bài toán này.

Quy luật đa thức là sự mở rộng của quy luật nhị thức khi thỏa mãn các điều kiện sau:

- Tiến hành n phép thử độc lập.

- Trong mỗi phép thử có thể xảy ra k biến cố $A_1, A_2, ..., A_k$ tạo nên một nhóm đầy đủ các biến cố.

- Xác suất xảy ra biến cố A_i $(i = \overline{1,k})$ trong mỗi phép thử đều bằng p_i $(i = \overline{1,k})$. Lúc đó, xác suất để trong n phép thử nói trên biến cố A_i $(i = \overline{1,k})$ xuất hiện tương ứng n_i lần $(\sum_{i=1}^{n} n_i = n)$ được tính bằng công thức:

$$P_n(n_1,n_2,...,n_k) = \frac{n!}{n_1!n_2!...n_k!} P_1^{n_1} P_2^{n_2} ... P_k^{n_k} \qquad (8.83)$$

Lúc đó, biến ngẫu nhiên X chỉ số lần xuất hiện các biến cố A_1, A_2,..., A_k trong n phép thử nói trên sẽ phân phối theo quy luật đa thức với các tham số là n và p_1, p_2,..., p_k.

Bài toán kiểm định sự phù hợp của quy luật đa thức được phát biểu như sau: Giả sử trong tổng thể nghiên cứu biến ngẫu nhiên X phân phối đa thức song chưa biết các tham số p_1, p_2,..., p_k. Nếu có cơ sở để giả thiết rằng giá trị của chúng tương ứng bằng p_1^0, p_2^0,..., p_k^0 người ta đưa ra cặp giả thuyết:

H_0: $p_i = p_i^0$ $(i = \overline{1,k})$.

H_1: Có ít nhất một xác suất khác với giá trị giả thuyết.

Để kiểm định giả thuyết trên, từ tổng thể rút ra một mẫu kích thước m và giả sử trong mẫu các biến cố A_1, A_2,..., A_k xuất hiện tương ứng n_1, n_2,..., n_k lần.

Tiêu chuẩn kiểm định vẫn là thống kê χ^2 trong biểu thức (8.81).

$$G = \chi^2 = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$$

trong đó n'_i là tần số lý thuyết tương ứng được tính với điều kiện giả thuyết H_0 là đúng, tức là:

$$n'_i = np_i^0$$

Do đó với mức ý nghĩa α miền bác bỏ W_α vẫn được xác định bằng biểu thức (8.82) và thủ tục kiểm định cũng tiến hành như ở các mục trước.

Thí dụ 4. Một công ty dược phẩm cho rằng loại thuốc cảm cúm mà họ đang bán có sự biến động khá rõ về nhu cầu

theo mùa vụ. Họ đánh giá rằng trong tổng khối lượng thuốc bán hàng năm thì 40% được bán vào mùa đông, 40% vào mùa xuân, 10% vào mùa hè và 10% vào mùa thu. Để đánh giá lại điều đó người ta thống kê 1000 lô thuốc đã tiêu thụ trong năm và thu được kết quả sau:

Mùa	Tần số
Đông	374
Xuân	292
Hè	169
Thu	165

Với mức ý nghĩa 0,05 hãy kiểm định các tỷ lệ tiêu thụ đã nói ở trên.

Giải. Ta có n = 1000 và k = 4

Cặp giả thuyết có dạng:

H_0: $p_1 = 0,4$; $p_2 = 0,4$; $p_3 = 0,1$; $p_4 = 0,1$

H_1: Có ít nhất một xác suất khác với giá trị giả thuyết.

Để tìm giá trị quan sát của tiêu chuẩn kiểm định ta lập bảng tính sau:

Mùa	Tần số thực nghiệm n_i	Tần số lý thuyết n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
Đông	374	400	1,69
Xuân	292	400	29,16
Hè	169	100	47,61
Thu	165	100	42,25
$\sum$	n = 1000	n = 1000	120,71 = χ^2_{qs}

Với $\alpha = 0,05 \rightarrow \chi_{\alpha}^{2(k-1)} = \chi_{0,05}^{2(3)} = 7,815$

Vậy miền bác bỏ là (7,815; +∞)

Do $\chi_{qs}^2 \in W_\alpha$ nên bác bỏ H_0, thừa nhận H_1 tức là tỷ lệ tiêu thụ thuốc không đúng như đã giả định.

Thí dụ 5. Giả sử với số liệu của bài trước ta muốn kiểm định xem có sự khác biệt hay không giữa các mùa về tỷ lệ thuốc tiêu thụ. Lúc đó cặp giả thuyết có dạng:

H_0: $p_1 = p_2 = p_3 = p_4 = 0,25$

H_1: Có ít nhất một xác suất khác 0,25

Ta có bảng tính sau:

Mùa	n_i	n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
Đông	374	250	61,504
Xuân	292	250	7,056
Hè	169	250	26,244
Thu	165	250	28,9
			$\chi_{qs}^2 = 120,704$

$\chi_{qs}^2 \in W_\alpha$ nên bác bỏ H_0.

b) Nếu X là biến ngẫu nhiên liên tục

Từ tổng thể rút ra một mẫu kích thước n và giả sử các số liệu mẫu được ghép lớp như sau:

$x_{i-1} - x_i$	$x_0 - x_1$	$x_1 - x_2$	...	$x_{k-1} - x_k$
n_i	n_1	n_2	...	n_k

Lúc đó các xác suất lý thuyết p_i chính là xác suất để biến ngẫu nhiên X nhận giá trị trong khoảng $(x_{i-1}; x_i)$ nếu giả thuyết H_0 là đúng:

$$p_i = p(x_{i-1} < X < x_i) \qquad (i = \overline{1,k}) \tag{8.84}$$

còn các tần số lý thuyết vẫn được tính bằng công thức:

$$n_i' = np_i \qquad (i = \overline{1,k})$$

Từ đó thủ tục kiểm định được tiến hành giống như ở các mục trước.

Sau đây ta sẽ xét trường hợp quan trọng hơn cả trong thực tế là kiểm định giả thuyết về dạng phân phối chuẩn của biến ngẫu nhiên. Trong trường hợp phân phối chuẩn công thức (8.84) có dạng cụ thể sau:

$$p_i = p(x_{i-1} < X < x_i) = \phi_0\left(\frac{x_i - \mu}{\sigma}\right) - \phi_0\left(\frac{x_{i-1} - \mu}{\sigma}\right) (i = \overline{1,k})$$

thông thường ta chưa biết μ và σ nên chúng được thay bằng các ước lượng hợp lý tối đa tương ứng là $\bar{x}$ và $\sqrt{MS}$. Lúc đó công thức trở thành:

$$p_i \approx \phi_0\left(\frac{x_i - \bar{x}}{\sqrt{MS}}\right) - \phi_0\left(\frac{x_{i-1} - \bar{x}}{\sqrt{MS}}\right)$$

Thí dụ 6. Gặt ngẫu nhiên 200 thửa ruộng của một vùng thu được các số liệu sau:

Bảng 8.8

Năng suất (Tạ/ha)	Số thửa ruộng tương ứng
4 - 6	15
6 - 8	26
8 - 10	25
10 - 12	30
12 - 14	26
14 - 16	21
16 - 18	24
18 - 20	20
20 - 22	13
	n = 200

Với mức ý nghĩa $\alpha = 0,05$ có thể coi năng suất lúa của vùng đó phân phối theo quy luật chuẩn được không?

Giải. Cặp giả thuyết thống kê:

H_0: Năng suất lúa X phân phối chuẩn

H_1: Năng suất lúa X không phân phối chuẩn

Qua mẫu cụ thể trên, ước lượng hợp lý tối đa của μ là $\bar{x} = 12,63$, của σ^2 là $MS = 22,04$ và $\sqrt{MS} = 4,695$.

Để tính các xác suất $p_i = P(x_i \le X \le x_{i+1})$ ta áp dụng công thức của quy luật chuẩn:

$$P(x_i \le X \le x_{i+1}) = \phi_0\left(\frac{x_{i+1} - \mu}{\sigma}\right) - \phi_0\left(\frac{x_i - \mu}{\sigma}\right)$$

hay
$$p_i \approx \phi_0\left(\frac{x_{i+1} - \bar{x}}{\sqrt{MS}}\right) - \phi_0\left(\frac{x_i - \bar{x}}{\sqrt{MS}}\right)$$

Với khoảng thứ nhất $(x_1 - x_2)$ ta thay bằng $(-\infty; x_2)$ và khoảng cuối cùng $(x_k - x_{k+1})$ thay bằng $(x_k; +\infty)$ để hợp của tất cả các khoảng tạo thành toàn bộ trục số. Lập bảng tính toán sau:

$x_i - x_{i+1}$	$U_i = \frac{x_i - \bar{x}}{\sqrt{MS}}$	$U_{i+1} = \frac{x_{i-1} - \bar{x}}{\sqrt{MS}}$	$\phi_0 (U_i)$	$\phi_0 (U_{i+1})$	$p_i = \phi_0(U_{i+1}) - \phi_0 (U_i)$	$n'_i = np_i$
$-\infty - 6$	$-\infty$	–1,41	–0,5	–0,4207	0,0793	15,86
6 – 8	–1,41	–0,99	–0,4207	–0,3389	0,0818	16,36
8 – 10	–0,99	–0,156	–0,3389	–0,2123	0,1266	25,32
10 – 12	–0,156	–0,13	–0,2123	–0,0517	0,1606	32,12
12 – 14	–0,13	0,29	–0,0517	0,1141	0,1658	33,16
14 – 16	0,29	0,72	0,1141	0,2642	0,1501	30,02
16 – 18	0,72	1,14	0,2642	0,3729	0,0197	21,74
18 – 20	1,14	1,57	0,3729	0,4418	0,0689	13,78
20 – $+\infty$	1,57	$+\infty$	0,4418	0,5	0,0582	11,6
					$\sum = 1,00$	$\sum = 200$

Giá trị quan sát $\chi^2_{qs} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} = 13,32$

Do $\alpha = 0,05 \Rightarrow \chi^{2(k-r-1)}_{\alpha} = \chi^{2(6)}_{0,05} = 12.6$

Vậy miền bác bỏ là (12,6; $+\infty$). $\chi^2_{qs} \in W_\alpha \Rightarrow$ Bác bỏ H_0, thừa nhận H_1 tức là thừa nhận X không tuân theo quy luật phân phối chuẩn.

Chú ý rằng tiêu chuẩn χ^2 của Pearson chỉ áp dụng được khi kích thước mẫu đủ lớn ($n > 50$) và các tần số tương ứng với mỗi giá trị hay mỗi khoảng giá trị của mẫu cũng phải đủ lớn ($n_i \geq 5; \forall i$). Vì vậy, nếu các tần số thực nghiệm n_i quá nhỏ ($n_i < 5$) thì phải ghép các giá trị hay các khoảng giá trị của mẫu lại để tăng giá trị của tần số thực nghiệm lên.

Nếu các điều kiện trên không thỏa mãn thì phải áp dụng các phương pháp kiểm định khác mới thu được kết luận đáng tin cậy.

3.2. Một số kiểm định khác về quy luật phân phối xác suất

Ở mục trước ta đã xét việc sử dụng tiêu chuẩn khi bình phương để kiểm định giả thuyết về dạng phân phối xác suất của biến ngẫu nhiên. Sau đây sẽ trình bày thêm một số thủ tục kiểm định khác về dạng phân phối xác suất đặc biệt là liên quan đến dạng phân phối chuẩn của tổng thể nghiên cứu.

1. Tiêu chuẩn phù hợp của Kolmogorov

Tiêu chuẩn Kolmogorov áp dụng được đối với mọi phân phối liên tục song các tham số của quy luật phân phối lý thuyết trong tổng thể giả định đã biết trước chứ không xác định dựa vào các số liệu mẫu. Giả sử biến ngẫu nhiên X trong tổng thể giả thiết phân phối theo một quy luật lý thuyết liên tục nào đó với hàm phân bố xác suất $F(x)$. Từ tổng thể lập mẫu kích thước n và tìm được hàm phân bố thực

nghiệm $F^*(X_i)$ tại mỗi giá trị X_i của mẫu $(i = \overline{1,n})$. Lúc đó tiêu chuẩn kiểm định giả thuyết được chọn là thống kê:

$$D = \max_{X_i} |F^*(X_i) - F(X_i)| \qquad (8.85)$$

Khi $n \to \infty$ thì không tùy thuộc vào dạng của F(x) biến ngẫu nhiên $\lambda = \sqrt{n}D$ luôn luôn hội tụ về phân phối Kolmogorov có hàm phân bố xác suất như sau:

$$K(x) = \begin{cases} \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2x^2} & \text{với } x > 0 \\ 0 & \text{với } x \le 0 \end{cases}$$

do đó với mức ý nghĩa α cho trước có thể tìm được giá trị tới hạn λ_α thỏa mãn điều kiện:

$$P(\lambda > \lambda_\alpha) = \alpha$$

Các giá trị của λ_α được tính sẵn thành bảng (Phụ lục 14). Như vậy, miền bác bỏ giả thuyết được xác định bằng biểu thức:

$$W_\alpha = \left\{ \lambda = \sqrt{n} \max_{X_i} |F^*(X_i) - F(X_i)|;\ \lambda > \lambda_\alpha \right\} \qquad (8.86)$$

Thủ tục kiểm định cũng được tiến hành như đã làm ở các mục trước.

Thí dụ 7. Kết quả đo kích thước của 1000 chi tiết được cho dưới dạng bảng phân phối thực nghiệm ghép lớp như sau: (x_i là giá trị giữa lớp).

i	x_i	n_i	i	x_i	n_i
1	98,0	21	6	100,5	201
2	98,5	47	7	101,0	142
3	99,0	87	8	101,5	97
4	99,5	158	9	102,0	41
5	100,0	181	10	102,5	25

Dùng tiêu chuẩn Kolmogorov hãy kiểm định giả thuyết cho rằng kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn với kỳ vọng toán $\mu = 100,25$ và độ lệch chuẩn $\sigma = 1$. Mức ý nghĩa được chọn là 0,05.

Giải. Hàm phân bố xác suất lý thuyết của phân phối chuẩn trong trường hợp này có thể viết dưới dạng:

$$F(x) = \frac{1}{2} + \phi_0(x - \mu)$$

Còn hàm phân bố thực nghiệm $F^*(x_i)$ có thể tính theo công thức:

$$F^*(x_i) = \frac{1}{1000}\left[\sum_{j=1}^{i-1} n_j + 0,5n_i\right]$$

Với mỗi giá trị x_i ta tìm hiệu $F^*(x_i) - F(x_i)$ và tìm giá trị lớn nhất về giá trị tuyệt đối. Quá trình tính toán được cho trong bảng sau:

Bảng 8.9

i	$x_i - \mu$	$\phi_0(x_i - \mu)$	$F(x_i)$	$F^*(x_i)$	$F^*(x_i) - F(x_i)$
1	–2,25	–0,4877	0,0123	0,0105	0,0018
2	–1,75	–0,4599	0,0401	0,0445	0,0044
3	–1,25	–0,3944	0,1056	0,1115	0,0059
4	–0,75	–0,2734	0,2266	0,2340	0,0074
5	–0,25	–0,0987	0,4013	0,4035	0,0022
6	0,25	0,0987	0,5987	0,5945	0,0042
7	0,75	0,2734	0,7734	0,7660	0,0074
8	1,25	0,3944	0,8944	0,8855	0,0089
9	1,75	0,4599	0,9599	0,9545	0,0054
10	2,25	0,4877	0,9877	0,9875	0,0002

Từ bảng tính toán tìm được

$$D_{qs} = 0,0089$$

do đó $\lambda_{qs} = \sqrt{n}\, D_{qs} = \sqrt{1000}\,.0,0089 = 0,281$

Với $\alpha = 0,05$ giá trị tới hạn $\lambda_\alpha = \lambda_{0,05} = 1,358$ vậy miền bác bỏ giả thuyết là $(1,358; +\infty)$. $\lambda_{qs} \notin W_\alpha$ do đó với mức ý nghĩa 0,05 chưa có cơ sở bác bỏ giả thuyết về dạng phân phối chuẩn của kích thước chi tiết.

Chú ý rằng để có thể áp dụng tiêu chuẩn phù hợp của Kolmogorov thì kích thước mẫu cũng phải đủ lớn ($n \geq 40$).

Một biến tướng của tiêu chuẩn kiểm định trên là tiêu chuẩn Kolmogorov - Smirnov. Tiêu chuẩn này được sử dụng để kiểm định xem hai mẫu kích thước n_1 và n_2 có cùng được rút ra từ một tổng thể nghiên cứu hay không. Để kiểm định điều đó người ta sử dụng thống kê:

$$D_{n_1, n_2} = \max_x \left| F_1^*(x) - F_2^*(x) \right| \qquad (8.87)$$

Trong đó: $F_1^*(x)$ và $F_2^*(x)$ là các hàm phân bố thực nghiệm của hai mẫu tương ứng. Từ đó tiêu chuẩn kiểm định được chọn là:

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1, n_2} \qquad (8.88)$$

và với mức ý nghĩa α miền bác bỏ giả thuyết được xác định bằng biểu thức:

$$W_\alpha = \left\{ \lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max_x \left| F_1^*(x) - F_2^*(x) \right| ; \lambda > \lambda_\alpha \right\} \qquad (8.89)$$

Thủ tục kiểm định cũng được tiến hành giống như trên.

Thí dụ 8. Có hai nhóm chi tiết cùng loại do hai máy sản xuất, mỗi nhóm gồm 60 chi tiết được đem đo và thu được kết quả cho trong bảng sau (Bảng 8.10).

Bảng 8.10

x_i	Số giá trị $x \le x_i$		$F_1^*(x)$	$F_2^*(x)$	$\lvert F_1^*(x) - F_2^*(x) \rvert$
	Nhóm I	Nhóm II			
71,95	1	0	0,0167	0	0,0167
72,11	1	1	0,0167	0,0167	0,0000
72,12	1	2	0,0167	0,0333	0,0167
72,14	3	2	0,0500	0,0333	0,0167
...	...	...	...	...	...
72,53	43	50	0,7167	0,8333	0,1167
72,54	46	52	0,7667	0,8667	0,1000
72,55	46	56	0,7667	0,9333	0,1667
72,56	49	57	0,8167	0,9500	0,1333
72,58	51	57	0,8500	0,9500	0,1000
72,60	52	58	0,8667	0,9667	0,1000
...	...	...	...	...	...
72,69	57	60	0,9500	1,0000	0,0500
72,70	58	60	0,9667	1,0000	0,0333
72,72	59	60	0,9833	1,0000	0,0167
72,73	60	60	1,0000	1,0000	0,0000

Hãy dùng tiêu chuẩn Kolmogorov - Smirnov để kiểm định với mức ý nghĩa 0,1 xem hai mẫu trên có thuộc cùng một tổng thể hay không, tức là hai máy có sản xuất các chi

tiết với kích thước theo cùng một quy luật phân phối xác suất hay không.

Giải. Sắp xếp kích thước của các chi tiết thuộc cả hai nhóm theo thứ tự tăng dần và tìm các hàm phân bố thực nghiệm $F_1^*(x)$ và $F_2^*(x)$ từ đó xác định hiệu của chúng về giá trị tuyệt đối (xem bảng).

Từ đó: $D_{n_1.n_2} = 0{,}1667$

và $$\lambda_{qs} = \sqrt{\frac{60.60}{60+60}}\,0{,}1667 = 0{,}9130$$

với $\alpha = 0{,}1 \rightarrow \lambda_\alpha = 1{,}224$ vậy miền bác bỏ giả thuyết là $(1{,}224; +\infty)$. Do $\lambda_{qs} \notin W_\alpha$ nên chưa có cơ sở để bác bỏ giả thuyết H_0. Nói cách khác phân phối xác suất của kích thước các chi tiết do hai máy đó sản xuất có thể coi là như nhau.

2. Kiểm định Lilliefors về dạng phân phối chuẩn

Khi sử dụng tiêu chuẩn phù hợp của Kolmogorov ta giả thiết rằng kỳ vọng toán và phương sai của phân phối lý thuyết được xác định bên ngoài mẫu và đã biết. Nếu điều kiện này không thỏa mãn thì việc kiểm định sẽ kém chính xác.

H.W Lilliefors đã áp dụng kiểm định Kolmogorov vào trường hợp phân phối chuẩn khi kỳ vọng toán và phương sai của tổng thể chưa biết. Lúc đó sẽ dùng trung bình mẫu và phương sai mẫu để thay thế cho kỳ vọng toán và phương sai tổng thể khi tính giá trị của hàm phân bố xác suất lý thuyết. Còn tiêu chuẩn kiểm định cũng vẫn là cực đại của sai lệch giữa hàm phân bố thực nghiệm và lý thuyết tại mỗi giá trị của mẫu. Sau đó giá trị này được so sánh trực tiếp với giá trị

tới hạn L_α được cho trong phụ lục 15 tương ứng với mức ý nghĩa α và kích thước mẫu n để kết luận.

Thí dụ 9. Theo dõi thời gian thực hiện một hóa đơn cho khách hàng (đơn vị tính là 10 giây) thu được kết quả sau: 51, 50, 45, 53, 46, 49, 47. Với mức ý nghĩa 0,05 có thể cho rằng thời gian thực hiện xong một hóa đơn phân phối chuẩn hay không.

Giải. Sắp xếp các giá trị của X theo trình tự tăng dần và tiến hành tính toán ta có bảng sau (Bảng 8.11).

Bảng 8.11

i	x_i	x_i^2	$u_i = \frac{x_i - \bar{x}}{s}$	$F^*_{(x_i)} = \frac{i}{7}$	$F_{(x_i)} = P(x < x_i)$	$\lvert F^*_{(x_i)} - F_{(x_i)} \rvert$
1	45	2025	–1,40	0,1429	0,0808	0,0621
2	46	2116	–1,02	0,2857	0,1539	0,1318
3	47	2209	–0,65	0,4286	0,2578	0,1708
4	49	2401	0,11	0,5714	0,5438	0,0276
5	50	2500	0,48	0,7143	0,6844	0,0299
6	51	2601	0,86	0,8571	0,8051	0,0520
7	53	2809	1,61	1,0000	1,0000	0,0000

$\bar{x} = 48,7143$

$s = 2,6573$

Từ đó:

$$D_{qs} = \max_{x_i} \left| F^*_{(x_i)} - F_{(x_i)} \right| = 0,1708$$

với n = 7; α = 0,05 từ phụ lục 15 tìm được $L_\alpha = L_{0,05} = 0,3$ vậy miền bác bỏ là (0,3; +∞). Do $D_{qs} \notin W_\alpha$ nên chưa có cơ sở bác bỏ H_0 hay có thể cho rằng thời gian thực hiện một hóa đơn phân phối chuẩn.

Các kiểm định trên đều có thể thực hiện bằng phần mềm Stata. Sau đây là một kết quả kiểm định phân phối chuẩn theo kiểm định Kolmogorov thu được bằng phần mềm Stata

.ksmirnov x3 = normprob ((x3 – 1628.12)/48.3997)

One - Sample Kolmogorov - Smirnov test against theoretical distribution

normprob ((x3 – 1628.12)/48.3997)

Smaller group	D	P-value	Corrected
x3:	0.0698	0.377	
Cumulative	-0.0603	0.483	
Combined K - S	0.0698	0.714	0.670

3. Kiểm định Jarque - Bera về dạng phân phối chuẩn

Một kiểm định khác về dạng phân phối chuẩn, đặc biệt hay được dùng trong phân tích hồi quy để kiểm định dạng phân phối chuẩn của các phần dư là kiểm định Jarque - Bera. Trong kiểm định này người ta sử dụng hệ số bất đối xứng a_3 của mẫu và hệ số nhọn a_4 của mẫu để kiểm định.

Tiêu chuẩn kiểm định được chọn là thống kê:

$$JB = n\left[\frac{a_3^2}{6} + \frac{(a_4 - 3)^2}{24}\right] \qquad (8.90)$$

Nếu giả thuyết H_0 về dạng phân phối chuẩn đúng thì với kích thước mẫu đủ lớn thống kê JB sẽ phân phối xấp xỉ khi

bình phương với 2 bậc tự do vì vậy với mức ý nghĩa α miền bác bỏ được xác định bằng biểu thức:

$$W_\alpha = \left\{ JB = n\left[\frac{a_3^2}{6} + \frac{(a_4 - 3)^2}{24}; JB > X_\alpha^{2(2)}\right]\right\} \qquad (8.91)$$

Thủ tục kiểm định cũng tiến hành giống như ở các mục trước.

Thí dụ 10. Với các số liệu đã cho trong thí dụ A, với mức ý nghĩa 0,05 hãy kiểm định xem thu nhập hàng năm của dân cư vùng 3 có phân phối chuẩn hay không.

Giải. Từ các số liệu của vùng 3 ta tìm được:

$\bar{x} = 1628,12;$ $s^2 = 2342,531$

$a_3 = 0,0724;$ $a_4 = 2,777$

Từ đó:

$$JB_{qs} = 100\left[\frac{0,0724^2}{6} + \frac{(2,777-3)^2}{24}\right] = 0,29$$

Với $\alpha = 0,05$ tra bảng được $X_{0,05}^{2(2)} = 5,991$ vậy miền bác bỏ là $(5,991; +\infty)$

Do $JB_{qs} \notin W_\alpha$ nên chưa thể bác bỏ giả thiết là thu nhập của dân cư vùng 3 phân phối chuẩn.

Kết quả giải bằng Stata như sau:

```
.sktest x3
```

Skewness/Kurtosis tests for Normality

		joint		
Variable	Pr (Skewness)	Pr (Kurtosis)	adj chi - sq (2)	Pr (chi - sq)
x3	0.753	0.825	0.15	0.9289

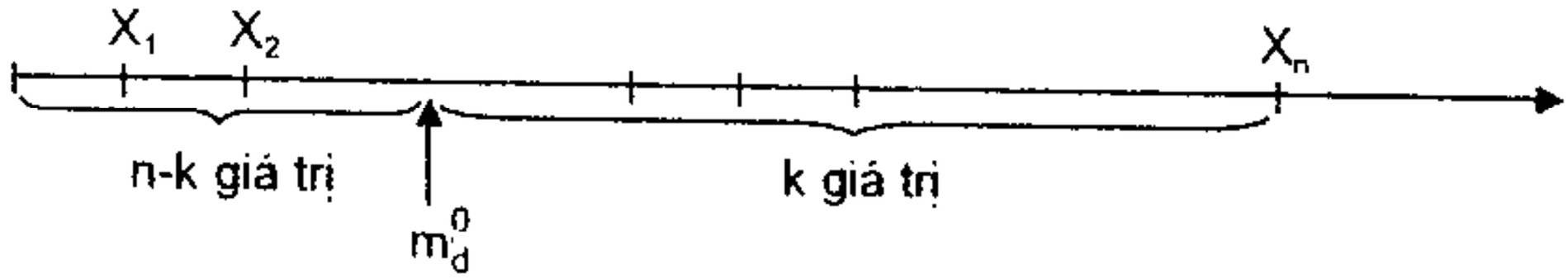
3.3. Kiểm định theo dấu

Kiểm định theo dấu là một loại kiểm định phi tham số được dùng trong các quyết định kinh doanh. Nó thường được sử dụng để kiểm định giả thuyết về giá trị trung vị của tổng thể vì như đã trình bày ở mục 2.8 chương VII, trong nhiều trường hợp thực tế trung vị lại tỏ ra hữu dụng hơn trung bình. Sau đây ta xét hai ứng dụng của kiểm định theo dấu đối với giá trị trung vị.

1. Dùng kiểm định theo dấu để kiểm định về một trung vị của tổng thể

Giả sử trong tổng thể nghiên cứu, giá trị của trung vị của biến ngẫu nhiên chưa biết song có cơ sở để giả thiết rằng nó bằng m_d^0. Lúc đó, ta đưa ra giả thuyết thống kê: H_0: $m_d = m_d^0$.

Để kiểm định giả thuyết trên, từ tổng thể lập mẫu ngẫu nhiên kích thước n và sắp xếp các giá trị mẫu theo thứ tự tăng dần X_1 X_2... X_n. Giả sử trên trục số các giá trị mẫu có phân phối như sau:



và như trên hình, giả sử có k giá trị mẫu lớn hơn hoặc bằng giá trị giả thuyết m_d^0. Nếu ta gán cho các giá trị lớn hơn hoặc bằng m_d^0 dấu +, còn các giá trị nhỏ hơn m_d^0 dấu – thì k chính là số giá trị mang dấu + trên mẫu.

Như đã phân tích ở mục 2.8 Chương VII, xác suất để mỗi giá trị của mẫu lớn hơn hoặc bằng m_d đều bằng 0,5 do đó nếu

giả thuyết H_0 là đúng thì k là biến ngẫu nhiên phân phối nhị thức. Vì vậy, có thể dùng các phương thức sau để kiểm định giả thuyết H_0:

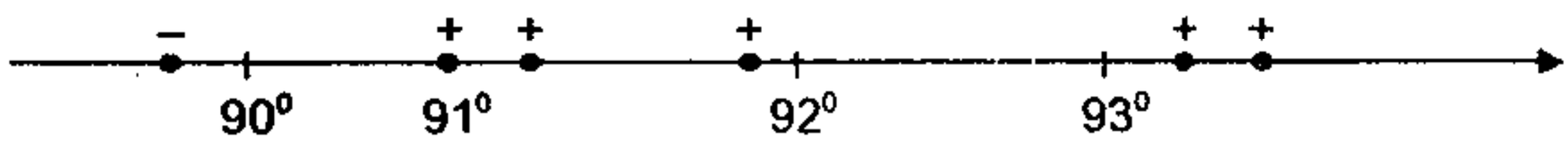
- Do khoảng tin cậy cũng có thể dùng để kiểm định giả thuyết H_0 (xem mục 2.1 Chương VIII) do đó xuất phát từ trung vị x_d của mẫu có thể xây dựng khoảng tin cậy mức $1 - \alpha$. Từ đó bác bỏ H_0 với mức ý nghĩa α nếu m_d^0 rơi ra ngoài khoảng tin cậy.

- Dùng công thức Bernoulli để tìm P-value theo công thức

$$\text{P-value} = P(x \geq k)$$

và dựa vào P-value để kết luận.

Thí dụ 11. Khi tia sáng của máy phân cực chiếu qua loại đường lactoza α thì góc quay của kim đồng hồ là 90^0. Một kỹ sư hóa công nghiệp lấy ngẫu nhiên 6 mẫu đường từ một loại đường mới nhập về đem đo và thu được kết quả mô tả trên trục số như sau:



Tìm P-value để kiểm định giả thuyết là loại đường nhập về đúng là đường lactoza α.

Giải. Theo bảng phân phối nhị thức với $n = 6$; $p = 0,5$ tìm được:

$$\text{P-value} = P(X \geq 5) = p_6(5) + p_6(6) = 0,11 = 11\%$$

Vậy nếu lấy mức ý nghĩa là 0,05 thì chưa thể bác bỏ H_0 hay có thể thừa nhận loại đường nhập về đúng là lactoza α.

Sau đây là kết quả giải bằng Stata:

```
.bitesti 6 1 0. 5
```

N	Observed k	Expected k	Assumed p	Observed p
6	1	3	0.50000	0,16667

```
Pr (k >= 1)                 = 0.984375 (one - sided test)
Pr (k <= 1)                 = 0.109375 (one - sided test)
Pr (k <= 1 or k >= 5)       = 0.218750 (two - sided test)
```

2. Dùng kiểm định theo dấu để kiểm định giả thuyết về hai trung vị của hai tổng thể khi hai mẫu gồm các giá trị theo cặp

Khi kiểm định sự bằng nhau của hai trung vị trong trường hợp hai mẫu rút ra có các giá trị tương ứng theo từng cặp thì có thể đưa bài toán về trường hợp đã xét ở trên bằng cách thiết lập sự sai lệch của từng cặp giá trị:

$$D_i = X_{1i} - X_{2i} \qquad i = \overline{1,n}$$

Lúc đó coi như ta chỉ có một mẫu gồm các giá trị D_1 D_2... D_n và việc kiểm định sự bằng nhau của hai trung vị tương đương với kiểm định trung vị của các sai lệch bằng không: H_0: $m_d = 0$.

Lúc đó xác suất để với mỗi giá trị của mẫu các sai lệch nhận giá trị dương hoặc âm bằng 0,5 và cũng như trên, xác suất để số quan sát X của mẫu nhận giá trị dương sẽ phân phối nhị thức với các tham số là n (kích thước mẫu) và p = 0,5. Từ đó tìm được P-value và kết luận.

Thí dụ 12. Tám người tình nguyện kiểm tra khả năng hít thở của phổi trước và sau khi áp dụng một phương pháp mới để điều trị bệnh hen. Kết quả thu được như sau:

X_1 (trước điều trị)	X_2 (sau điều trị)	$D_i = X_{1i} - X_{2i}$
750	850	+100
860	880	+20
950	930	–20
839	860	+ 30
750	800	+ 50
680	740	+ 60
720	760	+ 40
810	800	–10

Hãy tìm P–value để kiểm định giả thuyết cho rằng phương pháp điều trị này không mang lại hiệu quả.

Giải. Phương pháp điều trị không có hiệu quả tương đương với giả thuyết trung vị của các sai lệch bằng không H_0: $m_d = 0$.

Theo bảng phân phối nhị thức với n = 8; p = 0,5 tìm được

$$P\text{–value} = P(d \geq 6) = 0{,}145 = 14{,}5\%$$

Vì P–value khá lớn nên không thể bác bỏ H_0 ở mức ý nghĩa 0,05, thậm chí ở mức 0,1.

Sau đây là kết quả giải bằng Stata

```
.signtest x1 = x2
Sign test
        sign |   observed      expected
-------------+------------------------
    positive |          2             4
    negative |          6             4
        zero |          0             0
         all |          8             8
```

One · sided tests:

H_0: median of x1 – x2 = 0 vs. Ha: median of **x1 – x2 > 0**

Pr (#positive > = 2)

= Binomial (n = 8, x >= 2, p = 0.5) = 0.9648

H_0: median of x1 – x2 = 0 vs. Ha: median of **x1 – x2 < 0**

Pr (#positive > = 6)

= Binomial (n = 8, x >= 6, p = 0.5) = 0.1445

Two – sided test:

H_0: median of x1 – x2 = 0 vs. Ha: median of **x1 – x2 ~ = 0**

Pr (#positive > = 6 or # negative >= 6)

= min (1, 2* Binomial (n = 8, x >= 6, p = 0.5)) = 0.2891

3.4. Kiểm định tổng hạng của Wilcoxon về hai kỳ vọng toán của hai biến ngẫu nhiên

Như đã trình bày ở mục 2, để kiểm định sự bằng nhau của hai kỳ vọng toán của 2 biến ngẫu nhiên trong tổng thể nghiên cứu thì ta luôn giả thiết rằng các dấu hiệu nghiên cứu trong tổng thể phân phối chuẩn. Nếu chúng không phân phối chuẩn thì các mẫu điều tra phải đủ lớn để có thể áp dụng định lý giới hạn trung tâm. Song nếu các kích thước mẫu điều tra không đủ lớn thì không thể áp dụng các phương pháp kiểm định đã xét. Ở phần này ta đưa ra giải pháp cho trường hợp này thông qua kiểm định phi tham số bằng tổng các hạng của Wilcoxon (còn gọi là kiểm định Mann - Whitney). Nó dựa trên các giả thiết sau:

- Các biến ngẫu nhiên X_1 và X_2 trong hai tổng thể nghiên cứu có thể phân phối theo một quy luật bất kì không nhất thiết là quy luật chuẩn.

- Các mẫu ngẫu nhiên rút ra từ hai tổng thể phải độc lập nhau song có thể có kích thước tùy ý.

Lúc đó, việc kiểm định sự bằng nhau của hai kỳ vọng toán tương đương với việc kiểm định giả thuyết

H_0: Hai tổng thể có phân phối giống nhau.

Với các giả thuyết đối tương ứng

H_1: Tổng thể thứ nhất có phân phối lệch sang phải so với tổng thể thứ hai.

H_1: Tổng thể thứ nhất có phân phối lệch sang trái so với tổng thể thứ hai.

H_1: Hai tổng thể 1 và 2 có các tham số đặc trưng vị trí hoàn toàn khác nhau.

Nếu giả thuyết H_0 là đúng thì hai mẫu độc lập được rút ra từ hai tổng thể đó sẽ tương đồng với nhau. Để đo mức độ tương đồng này có thể sử dụng các hạng kết hợp (từ nhỏ đến lớn) của các giá trị của mẫu kết hợp lại từ hai mẫu điều tra và kiểm tra tổng các hạng của các giá trị của mẫu thứ nhất (hoặc của mẫu thứ hai).

(Nhắc lại rằng xếp hạng các giá trị của một biến là việc gán số thứ tự cho các giá trị đó theo trình tự tăng dần của chúng).

Nếu giả thuyết H_0 là đúng, tức là hai tổng thể là như nhau thì tổng các hạng của mẫu sẽ tỷ lệ thuận với kích thước mẫu. Nếu ký hiệu T là tổng các hạng của mẫu thứ nhất thì về trực quan có thể thấy nếu T quá nhỏ hoặc quá lớn sẽ là chứng cứ để ta bác bỏ H_0.

Nếu H_0 đúng thì thống kê T sẽ có phân phối mẫu với kỳ vọng toán và phương sai như sau:

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{8.92}$$

$$\sigma_T^2 = \frac{n_1 n_2}{12}(n_1 + n_2 + 1) \tag{8.93}$$

Hơn nữa nếu cả n_1 và n_2 đều lớn hơn 10 thì T sẽ phân phối xấp xỉ chuẩn.

Kiểm định tổng các hạng Wilcoxon giả thiết rằng biến X trong tổng thể là liên tục, như vậy sẽ không có bất kỳ hai giá trị nào bằng nhau. Trong thực tế, khi điều tra mẫu có thể gặp trường hợp hai hoặc nhiều giá trị mẫu bằng nhau. Lúc đó mỗi giá trị của nhóm giá trị bằng nhau sẽ được gán cho hạng bằng trung bình số học của các hạng trong nhóm. Chẳng hạn nếu hai giá trị bằng nhau nằm ở hạng thứ ba và thứ tư thì mỗi giá trị được gán hạng bằng 3,5 và hạng của giá trị tiếp theo sẽ là 5. Khi có các giá trị mẫu bằng nhau thì công thức của phương sai là:

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left[n_1 + n_2 + 1 - \frac{\sum_j t_j(t_j^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right] \tag{8.94}$$

Trong đó: t_j là tần số của các hạng ghép nhóm trong nhóm thứ j. Chú ý rằng nếu mọi giá trị của mẫu đều khác nhau thì từ (8.94) ta lại thu được (8.93).

Thủ tục kiểm định tổng các hạng của Wilcoxon bao gồm hai trường hợp sau:

α) Nếu $n_1 \leq 10$ và $n_2 \leq 10$

Lúc đó thống kê T là tổng các hạng của mẫu thứ nhất. Với mức ý nghĩa α cho trước dùng phụ lục 11 để tìm các giá

trị tới hạn T_L và T_u và tùy thuộc vào giả thuyết đối H_1 các miền bác bỏ được xác định như sau:

a) H_0: Tổng thể 1 và 2 có phân phối giống nhau

H_1: Tổng thể 1 có phân phối lệch sang phải so với tổng thể thứ hai

$$W_\alpha = \{T; T > T_u\} \quad (8.95)$$

b) H_0: Tổng thể 1 và 2 có phân phối giống nhau

H_1: Tổng thể 1 có phân phối lệch sang trái so với tổng thể thứ hai

$$W_\alpha = \{T; T < T_L\} \quad (8.96)$$

c) H_0: Tổng thể 1 và 2 có phân phối giống nhau.

H_1: Hai tổng thể có tham số đặc trưng vị trí hoàn toàn khác nhau.

$$W_\alpha = \{T; T < T_L \text{ hoặc } T > T_u\} \quad (8.97)$$

β) Nếu $n_1 > 10$ và $n_2 > 10$

Lúc đó tiêu chuẩn kiểm định có dạng:

$$G = U = \frac{T - \mu_T}{\sigma_T} \quad (8.98)$$

và W_α tương ứng với 3 cặp giả thuyết trên là:

$$\text{a. } W_\alpha = \left\{U = \frac{T - \mu_T}{\sigma_T}; U > u_\alpha\right\} \quad (8.99)$$

$$\text{b. } W_\alpha = \left\{U = \frac{T - \mu_T}{\sigma_T}; U < -u_\alpha\right\} \quad (8.100)$$

$$\text{c. } W_\alpha = \left\{U = \frac{T - \mu_T}{\sigma_T}; |U| > u_{\alpha/2}\right\} \quad (8.101)$$

Như vậy, kiểm định tổng các hạng của Wilcoxon gián tiếp phản ánh quan hệ của μ_1 và μ_2.

Nếu phân phối của tổng thể 1 lệch sang phải so với tổng thể 2 thì có cơ sở để kết luận rằng $\mu_1 > \mu_2$. Các kết luận tương tự có thể đưa ra với hai trường hợp còn lại.

Thí dụ 13. Các kỹ sư môi trường quan tâm đến việc là dự án nạo vét một cái hồ trong thành phố có thực sự mang lại hiệu quả hay không. Để làm điều đó, trước và sau khi nạo vét người ta đã lấy 12 mẫu nước mỗi lần và đo lượng oxygen không tan (tính bằng ppm) trong các mẫu nước và thu được kết quả sau:

Trước khi nạo vét		Sau khi nạo vét	
11	11,6	10,2	10,8
11,2	11,7	10,3	10,8
11,2	11,8	10,4	10,9
11,2	11,9	10,6	11,1
11,4	11,9	10,6	11,1
11,5	12,1	10,7	11,3

Với mức ý nghĩa 0,05 hãy kiểm định cặp giả thuyết sau:

H_0: Phân phối của lượng oxygen không tan trước và sau khi nạo vét là như nhau.

H_1: Phân phối lượng oxygen trước khi nạo vét lệch sang phải so với phân phối sau khi nạo vét.

Giải. Trước hết ta xác định hạng của phân phối kết hợp của 24 giá trị mẫu theo chiều tăng dần và sau đó tính tổng hạng của mẫu thứ nhất. Ta thu được kết quả sau:

Trước khi nạo vét		Sau khi nạo vét	
Giá trị	Hạng	Giá trị	Hạng
11	10	10,2	1
11,2	14	10,3	2
11,2	14	10,4	3
11,2	14	10,6	4,5
11,4	17	10,6	4,5
11,5	18	10,7	6
11,6	19	10,8	7,5
11,7	20	10,8	7,5
11,8	21	10,9	9
11,9	22,5	1,1	11,5
11,9	22,5	11,1	11,5
12,1	24	11,3	16
	T = 216		

Vì $n_1 > 10$ và $n_2 > 10$ ta dùng tiêu chuẩn kiểm định (8.98). Sau khi ghép nhóm các hạng giống nhau ta còn lại 18 nhóm với các tần số như sau:

Hạng	Tần số t_j	Hạng	Tần số t_j
1	1	14	3
2	1	16	1
3	1	17	1
4,5	2	18	1
6	1	19	1
7,5	2	20	1
9	1	21	1
10	2	22,5	2
11,5	2	24	1

Từ đó

$$\mu_T = \frac{n_1(n_1+n_2+1)}{2} = \frac{12(12+12+1)}{2} = 150$$

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left[(n_1+n_2+1) - \frac{\sum_j t_j(t_j^2-1)}{(n_1+n_2)(n_1+n_2-1)}\right]$$

$$= \frac{12.12}{12}\left[(12+12+1) - \frac{6+6+24+6}{(12+12)(12+12+1)}\right] = 298,956$$

$\sigma_T = 17,29$

Từ đó $T_{qs} = \frac{T-\mu_T}{\sigma_T} = \frac{216-150}{17,29} = 3,82$

Với $\alpha = 0,05 \rightarrow u_\alpha = u_{0,05} = 1,645$

Vậy miền bác bỏ là $(1,645; +\infty)$

Do $T_{qs} \in W_\alpha$ nên với mức ý nghĩa 0,05 bác bỏ H_0 thừa nhận H_1 tức là thừa nhận phân phối của lượng oxygen không tan trước khi nạo vét lệch sang phải so với phân phối này sau khi nạo vét. Nó cũng chứng tỏ rằng lượng oxygen không tan trung bình trong 1 ml nước hồ trước khi nạo vét lớn hơn so với sau nạo vét hay dự án vét hồ là thực sự có hiệu quả.

Giải bằng Stata cho kết quả sau:

.ranksum x1, by (x2)

Two – sample Wilcoxon rank – sum (Mann – Whitney) test

x2	obs	rank sum	expected
1	12	216	150
2	12	84	150
combined	24	300	300

unadjusted variance	300.00
adjustment for ties	– 1.04
adjusted variance	298.96

H_0: x1 (x2 = =1) = x1 (x2 = = 2)

z = 3.817

Prob > | z | = 0.0001

Phương pháp kiểm định dựa trên việc sử dụng hạng của các giá trị mẫu thay cho bản thân các giá trị đó còn có thể sử dụng đối với các dấu hiệu định tính được đo theo thang thứ bậc hay khoảng cách, khi sự khác biệt về giá trị mẫu không có nhiều ý nghĩa tính toán.

3.5. Kiểm định tổng hạng theo dấu của Wilcoxon

Phương pháp kiểm định theo từng cặp giá trị phụ thuộc xét ở mục 2 chỉ sử dụng được khi các biến ngẫu nhiên trong tổng thể phân phối chuẩn. Khi giả thiết này bị vi phạm thì có thể thay thế bằng phương pháp kiểm định phi tham số thông qua tổng hạng theo dấu Wilcoxon. Nó là một biến tướng của kiểm định tổng hạng đã xét ở mục 3.4.

Nếu kí hiệu:

n là tổng số các cặp giá trị của hai mẫu điều tra có hiệu số D_i khác không.

T_+ là tổng các hạng mang dấu dương

(nếu không có hạng nào mang dấu dương thì $T_+ = 0$).

T_- là tổng các hạng mang dấu âm (nếu không có hạng nào mang dấu âm thì $T_- = 0$)

$T = \min (T_+, |T_-|)$

$$\mu_T = \frac{n(n+1)}{4} \tag{8.102}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{8.103}$$

Nếu có nhiều giá trị D_i cùng dấu và cùng giá trị thì được ghép lại thành một nhóm và mỗi giá trị được gán cho hạng bằng trung bình các hạng của nhóm đó. Lúc đó phương sai của T được tính bằng công thức.

$$\sigma_T^2 = \frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_j t_j(t_j-1)(t_j+1)\right] \tag{8.104}$$

Nếu mọi nhóm đều có tần số bằng 1 thì ta lại thu được công thức (8.103).

Việc kiểm định được tiến hành theo các giả thuyết sau:

H_0: Phân phối của các hiệu số D_i là đối xứng qua giá trị 0.

a) H_1: Các hiệu số D_i có phân phối lệch sang phải điểm 0

b) H_1: Các hiệu số D_i có phân phối lệch sang trái điểm 0

c) H_1: Các hiệu số D_i hoặc lệch phải, hoặc lệch trái so với điểm 0

Lúc đó thủ tục kiểm định được tiến hành theo hai trường hợp sau:

α) Nếu $n \leq 50$

Tiêu chuẩn kiểm định là thống kê T được xác định tùy thuộc vào giả thuyết đối H_1 như sau:

a) $T = |T_-|$ (8.105)

b) $T = T_+$ (8.106)

c) $T = \min [T_+, |T_-|]$ (8.107)

Nguyên tắc kiểm định như sau: Với mức ý nghĩa α (với kiểm định một phía có thể bằng 0,05; 0,025; 0,01; 0,005 và kiểm định hai phía có thể bằng 0,1; 0,05; 0,02; 0,01) và số lượng các giá trị D_i khác không bằng n, bác bỏ H_0 nếu giá trị của T nhỏ hơn hoặc bằng giá trị tới hạn được cho trong phụ lục 12.

β) Nếu n > 50

Tiêu chuẩn kiểm định là:

$$U = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \qquad (8.108)$$

Với các giả thuyết đối H_1 ở trường hợp a và b thì bác bỏ H_0 nếu $U_{qs} < -u_\alpha$ còn với trường hợp c thì bác bỏ H_0 nếu $U_{qs} < -u_{\alpha/2}$

Thí dụ 14. Để thử nghiệm năng suất của hai giống lúa người ta chọn ngẫu nhiên 10 thửa ruộng, mỗi thửa lại chia đôi, một phần trồng giống lúa A, một phần trồng giống lúa B. Sau khi thu hoạch được kết quả sau (đơn vị kg/sào).

Thửa ruộng	Giống lúa A	Giống lúa B	Hiệu số
1	312	346	–34
2	333	372	–39
3	356	392	–36
4	316	351	–35
5	310	330	–20
6	352	364	–12
7	389	375	14
8	313	315	–2
9	316	327	–11
10	346	378	–32

Với mức ý nghĩa 0,05 hãy kiểm định xem phân phối của năng suất hai giống lúa có khác nhau hay không.

Giải. Trước hết ta xếp hạng từ thấp đến cao giá trị tuyệt đối của các hiệu số D_i, sau đó gán cho các hạng dấu tương ứng của chúng. Ta thu được bảng sau:

Thửa ruộng	Hạng của hiệu số $\|X_{1i} - X_{2i}\|$	Hạng có dấu
1	7	–7
2	10	–10
3	9	–9
4	8	–8
5	5	–5
6	3	–3
7	4	4
8	1	–1
9	2	–2
10	6	–6

Từ đó

$T_{+qs} = 4$

$T_{-qs} = -51$

$T_{qs} = \min [T_+, |T_-|] = 4$

Để kiểm định hai phía với $n = 10$ và $\alpha = 0,01$, tra bảng tìm được giá trị tới hạn là 8.

Do $T_{qs} \leq 8$ nên với mức ý nghĩa 0,05 bác bỏ H_0 và thừa nhận H_1 tức là phân phối của năng suất hai giống lúa là khác

nhau. Điều đó cho thấy rằng năng suất lúa trung bình cũng khác nhau theo hướng giống lúa A thấp hơn giống lúa B vì phân phối năng suất có xu hướng lệch trái.

Giải bằng Stata cho kết quả sau:

.signrank x1 = x2

Wilcoxon signed – rank test

sign	obs	sum ranks	expected
positive	1	4	27.5
negative	9	51	27.5
zero	0	0	0
all	10	55	55

unadjusted variance 96.25

adjustment for ties 0.00

adjustment for zeros 0.00

adjusted variance 96.25

H_0: x1 = x2

z = – 2.395

Prob > |z| = 0.0166

3.6. Kiểm định Kruskal - Wallis về k kỳ vọng toán

Khi cần so sánh k (k > 2) kỳ vọng toán của k biến ngẫu nhiên không phân phối chuẩn thì có thể áp dụng kiểm định Kruskal - Wallis sau đây. Nó là sự mở rộng tương ứng của kiểm định tổng hạng của Wilcoxon.

Giả sử có k tổng thể nghiên cứu, trong đó các biến ngẫu

nhiên X_1 X_2... X_k phân phối theo một quy luật chưa biết, hoặc đã biết song không phải là phân phối chuẩn. Từ các tổng thể trên rút ra k mẫu độc lập kích thước tương ứng là n_1 n_2... n_k. Lúc đó việc kiểm định sự bằng nhau của k kỳ vọng toán tương đương với việc kiểm định cặp giả thuyết.

H_0: k phân phối trong các tổng thể là giống nhau.

H_1: Có ít nhất hai phân phối khác nhau.

Tiêu chuẩn kiểm định được chọn là thống kê

$$H = \frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{T_i^2}{n_i} - 3(n_T+1) \qquad (8.109)$$

Trong đó:

n_i là kích thước mẫu thứ i $(i = \overline{1,k})$

n_T là tổng kích thước của k mẫu

T_i là tổng các hạng của mẫu thứ i sau khi tất cả các mẫu đã được kết hợp lại và xếp hạng chung.

Với điều kiện giả thuyết H_0 là đúng thì thống kê H phân phối khi bình phương với số bậc tự do bằng k – 1.

Do đó với mức ý nghĩa α miền bác bỏ được xác định bằng biểu thức.

$$W_\alpha = \left\{H = \frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{T_i^2}{n_i} - 3(n_T+1);\ H > \chi_\alpha^{2(k-1)}\right\} \qquad (8.110)$$

Từ k mẫu cụ thể tìm được H_{qs}, so sánh với W_α và kết luận.

Thí dụ 15. Ba nhóm học sinh trung học, mỗi nhóm 10 người thuộc các chuyên ban A, B và D được chọn một cách

ngẫu nhiên để kiểm tra kiến thức về văn phạm tiếng Việt. Kết quả điểm kiểm tra như sau:

Nhóm chuyên ban A	Nhóm chuyên ban B	Nhóm chuyên ban D
32	28	32
32	21	30
26	15	30
26	15	29
22	14	26
20	14	23
19	14	20
16	11	19
14	9	18
14	8	12

Với mức ý nghĩa 0,05 hãy kiểm định xem học sinh theo học các chuyên ban A, B và D có khác nhau về kiến thức văn phạm tiếng Việt hay không.

Giải. Ta có cặp giả thuyết sau:

H_0 - Không có sự khác biệt giữa học sinh các chuyên ban về kiến thức văn phạm tiếng Việt.

H_1 - Có ít nhất 1 chuyên ban khác biệt với các chuyên ban khác.

Để tìm H_{qs} trước hết phải xếp hạng 30 kết quả thu được theo trình tự từ thấp đến cao. Ta thu được kết quả xếp hạng như sau:

A	Hạng	B	Hạng	D	Hạng
32	29	28	24	32	29
32	29	21	18	30	26,5
26	22	15	10,5	30	26,5
26	22	15	10,5	29	25
22	19	14	7	26	22
20	16,5	14	7	23	20
19	14,5	14	7	20	16,5
16	12	11	3	19	14,5
14	7	9	2	18	13
14	7	8	1	12	4
Σ	178		90		197

Từ đó

$$H_{qs} = \frac{12}{30(30+1)}\left[\frac{178^2}{10}+\frac{90^2}{10}+\frac{197^2}{10}\right]-3(30+1)=8,4$$

Nếu các mẫu có nhiều giá trị trùng nhau thì có thể tính H' thay cho H theo công thức:

$$H' = \frac{H}{1-\sum_{j=1}^{h}(t_j^3-t_j)/(n_T^3-n_T)} \qquad (8.111)$$

trong đó t_j là số giá trị giống nhau của nhóm thứ j.

Chẳng hạn với thí dụ trên ta xếp lại các giá trị của ba mẫu theo hạng và tần số tương ứng

Hạng	Tần số	Hạng	Tần số
1	1	16,5	2
2	1	18	1
3	1	19	1
4	1	20	1
7	5	22	3
10,5	2	24	1
12	1	25	1
13	1	26,5	2
14,5	2	29	3

lúc đó

$$\frac{\sum_j (t_j^3 - t_j)}{n_T^3 - n_T} = \frac{1}{30^3 - 30}[(5^3 - 5) + (2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (3^3 - 3) + (2^3 - 2) + (3^3 - 3)]$$

$$= \frac{192}{26970} = 0{,}0071$$

Từ đó

$$H'_{qs} = \frac{H}{1 - 0{,}0071} = \frac{8{,}4}{0{,}9929} = 8{,}46$$

như vậy là hai kết quả xấp xỉ nhau.

Với $\alpha = 0{,}05 \rightarrow \chi_{\alpha}^{2(k-1)} = \chi_{0{,}05}^{2(2)} = 5{,}991$

Vậy miền bác bỏ là $(5{,}991; +\infty)$

Do $H'_{qs} \in W_\alpha$ nên bác bỏ H_0, thừa nhận H_1 tức là thừa nhận có ít nhất một nhóm học sinh có kiến thức về văn phạm tiếng Việt cao hơn các nhóm khác.

Giải bằng Stata cho kết quả sau:

.kwallis x, by (ban)

Test: Equality of populations (Kruskal - Wallis Test)

ban	_Obs	_RankSum
A	10	178.00
B	10	90.00
D	10	197.00

chi – squared = 8.410 with 2 d.f.

probability = 0.0149

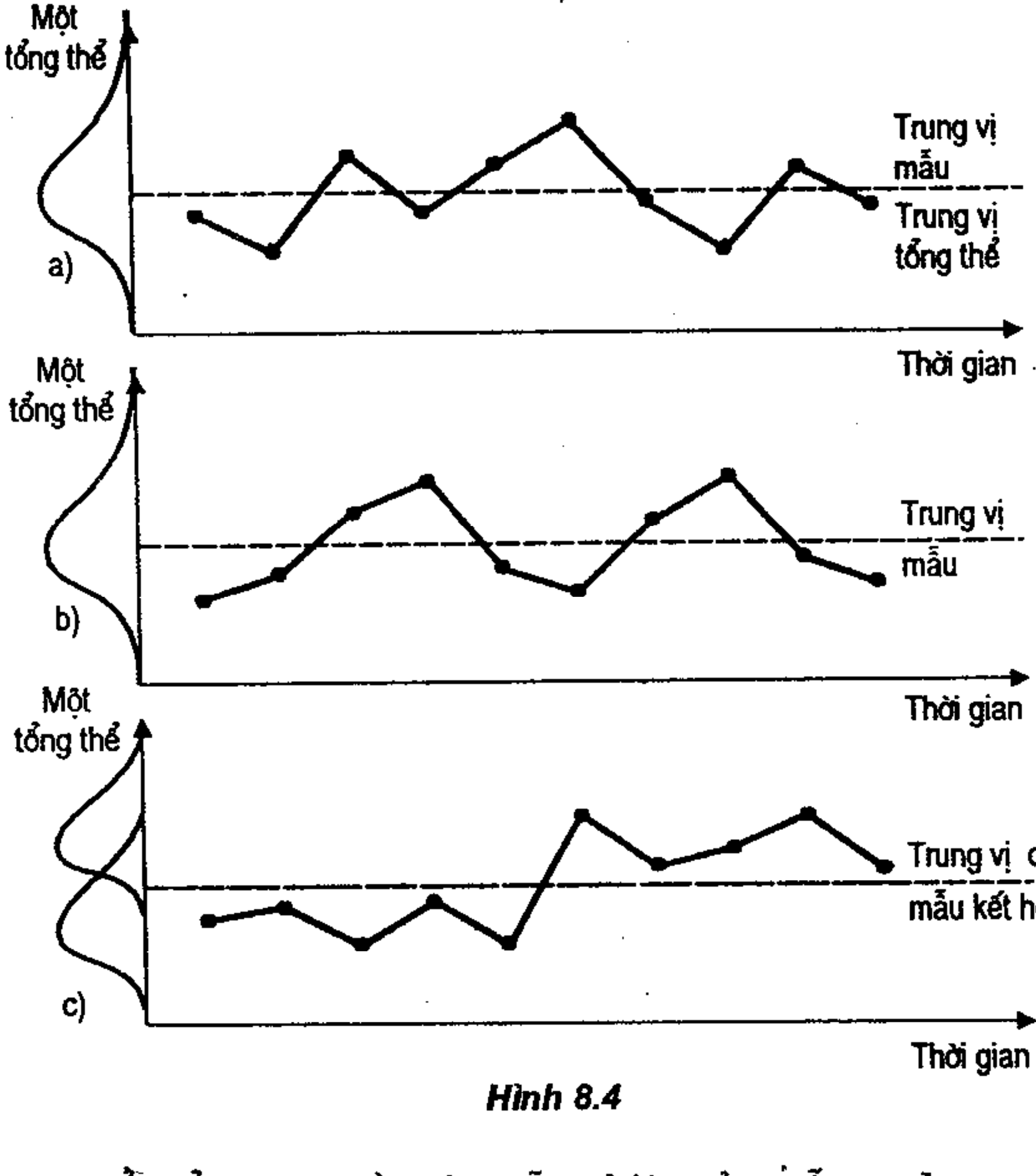
3.7. Kiểm định đoạn mạch

1. Kiểm định một mẫu

Một trong các giả thuyết cơ bản mà ta đưa ra ở các mục trước là giả thuyết mẫu được rút ra từ tổng thể một cách ngẫu nhiên. Phần này ta sẽ xét một phương pháp kiểm định tính ngẫu nhiên của mẫu gọi là phương pháp kiểm định đoạn mạch.

Theo định nghĩa mẫu ngẫu nhiên là tập hợp các biến ngẫu nhiên độc lập được xây dựng từ biến X trong tổng thể. Như vậy thì các phần tử của mẫu phải được lấy ra một cách độc lập từ một tổng thể chung. Chẳng hạn nếu giá trị mẫu được mô tả theo trật tự thời gian thì đồ thị của nó có dạng như hình 8.4a. Mặt khác nếu các giá trị mẫu tương quan với nhau thì đồ thị của chúng sẽ chao đảo như ở hình 8.4b. Còn

nếu mẫu được rút ra từ hai tổng thể khác nhau thì đồ thị của giá trị mẫu sẽ có dạng như hình 8.4c.



Hình 8.4

Để kiểm định về tính ngẫu nhiên của mẫu ta đưa ra giả thuyết thống kê

H_0: Mẫu được lập một cách ngẫu nhiên

H_1: Mẫu được lập một cách phi ngẫu nhiên

Chú ý rằng nếu H_0 là đúng tức là mẫu được tạo lập một cách ngẫu nhiên thì các giá trị nhỏ hơn trung vị (ký hiệu là T) và các giá trị lớn hơn trung vị (ký hiệu là C) phải phân phối một cách hoàn toàn ngẫu nhiên xung quanh trung vị.

Còn nếu H_0 sai thì sự phân phối đó sẽ mang tính quy luật rõ rệt. Chẳng hạn theo hình 8.4a ta có dãy phân phối trong quan hệ so sánh với trung vị như sau:

TT CT CC TT CT (*)

còn với hình 8.4b thì dãy phân phối là

TT CC TT CC TT (**)

Từ đó ta có định nghĩa sau: Mỗi đoạn mạch là một dãy liên tiếp các ký hiệu giống nhau mà trước và sau nó hoặc không có ký hiệu hoặc ký hiệu khác.

Như vậy, trong dãy (*) ta có số đoạn mạch R = 7. Còn trong dãy (**) ta có số đoạn mạch R = 5.

Giả sử kích thước mẫu là n (nếu n lẻ thì đường trung vị sẽ đi qua điểm trung vị, lúc đó điểm này sẽ không được tính là T hay C. Lúc đó n là số lượng các quan sát còn lại).

Nếu H_o là đúng thì số lượng của đoạn mạch R sẽ có phân phối xấp xỉ chuẩn với kỳ vọng toán

$$E(R) \approx \frac{n}{2} + 1 \qquad (8.112)$$

và sai số chuẩn

$$Se(R) \approx \frac{\sqrt{n-1}}{2} \qquad (8.113)$$

Từ đó tiêu chuẩn kiểm định được chọn là

$$U = \frac{R - E(R)}{Se(R)} \qquad (8.114)$$

và miền bác bỏ mức α được xác định bằng biểu thức

$$W_\alpha = \left\{ U = \frac{R - E(R)}{Se(R)};\ U < -u_\alpha \right\} \qquad (8.115)$$

Từ mẫu cụ thể xác định được R và tìm giá trị quan sát U_{qs}, so sánh nó với W_α và kết luận.

Thí dụ 16. Với các số liệu mẫu được cho trong hình 8.4.a hãy kiểm định giả thuyết về tính ngẫu nhiên của mẫu điều tra với mức ý nghĩa 0,05.

Giải. Từ hình 8.4.a ta tìm được

$$n = 10;\ R = 7$$

$$E(R) \approx \frac{10}{2} + 1 = 6$$

$$Se(R) \approx \frac{\sqrt{10-1}}{2} = 1{,}5$$

Từ đó

$$U_{qs} = \frac{7-6}{1{,}5} = 0{,}67$$

với $\alpha = 0{,}05 \rightarrow u_\alpha = u_{0.05} = 1{,}645$

Vậy miền bác bỏ là $(-\infty; -1{,}645)$

Do $U_{qs} = 0{,}67 \notin W_\alpha$ nên chưa thể bác bỏ H_0 tức là có thể cho rằng mẫu được rút ra một cách ngẫu nhiên.

2. *Kiểm định Wald - Wolfowitz*

Kiểm định Wald - Wolfowitz là sự mở rộng kiểm định đoạn mạch để xác định xem hai tổng thể có các phân phối xác suất như nhau hay không. Trong trường hợp này cặp giả thuyết thống kê là

H_0: Hai tổng thể nghiên cứu có cùng một phân phối xác suất

H_1: Hai tổng thể nghiên cứu có các phân phối xác suất khác nhau.

Như vậy đây là phương án kiểm định phi tham số để thay cho kiểm định T về sự bằng nhau của hai trung bình tổng thể. Để tiến hành kiểm định này chỉ yêu cầu hai mẫu rút ra từ hai tổng thể là độc lập và hoàn toàn ngẫu nhiên. Giả sử điều tra hai mẫu như vậy với kích thước tương ứng là n_1 và n_2 và các giá trị của mẫu được sắp xếp đồng thời theo trình tự tăng dần thành một dãy gồm $n = n_1 + n_2$ kết quả. Sau đó gọi R là tổng số đoạn mạch của dãy kết quả đó thì ta có thể áp dụng thủ tục kiểm định các đoạn mạch đã trình bày ở mục trước đối với R.

Thí dụ 17. Người quản lý một cửa hàng muốn kiểm định xem hiệu quả làm việc của hai nhân viên bán hàng có như nhau hay không. Để làm điều đó người quản lý đã thống kê số sản phẩm mà mỗi nhân viên đã bán được trong một số ngày và thu được kết quả sau:

Nhân viên A: 35 44 39 50 48 29 60 75 49 66

Nhân viên B: 17 23 13 24 33 21 18 16 32

Hãy tìm P-value để kiểm định xem hoạt động của nhân viên A có thực sự hiệu quả hơn nhân viên B hay không.

Giải. Ta có $n_1 = 10$ và $n_2 = 9$. Xếp các giá trị của cả hai mẫu thành một dãy kết quả theo trình tự tăng dần thu được dãy sau:

B B B B B B B A B B A A A A A A A A A

Như vậy tổng số đoạn mạch R = 4. Từ đó tìm được

$$E(R) = 10,47$$

$$\sigma_R = 2,1$$

và $$U_{qs} = \frac{4 - 10,47}{2,1} = -3,08$$

Tra bảng tìm được P-value ≈ 0.

Như vậy có thể bác bỏ giả thuyết là hiệu quả làm việc của hai nhân viên A và B là như nhau.

Các ký hiệu và công thức cơ bản

* H_0 - Giả thuyết gốc

* H_1 - Giả thuyết đối

* G - Tiêu chuẩn kiểm định

* G_{qs} - Giá trị quan sát của tiêu chuẩn kiểm định

* α - Mức ý nghĩa của kiểm định và là xác suất mắc sai lầm loại một

$$\alpha = P(G \in W_\alpha / H_0)$$

* β - Xác suất mắc sai lầm loại hai

$$\beta = P(G \notin W_\alpha / H_1)$$

* $1 - \beta$ - lực kiểm định

$$1 - \beta = P(G \in W_\alpha / H_1)$$

* W_α - miền bác bỏ giả thuyết. Nếu $G_{qs} \in W_\alpha \rightarrow$ Bác bỏ H_0

* P-value = P ($G > G_{qs}$) nếu là miền bác bỏ bên phải

P-value = P ($G < G_{qs}$) nếu là miền bác bỏ bên trái

P-value = P ($G > |G_{qs}|$) nếu là miền bác bỏ hai phía.

* Kiểm định U về μ (σ đã biết)

H_0: $\mu = \mu_0$;

Tiêu chuẩn kiểm định

$$U = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} \sim N(0,1)$$

$\beta = P[U < u_\alpha - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})}]$ nếu là miền bác bỏ một phía

$\beta = P[U < u_{\alpha/2} - \frac{|\mu_0 - \mu_1|}{Se(\bar{x})}]$ nếu là miền bác bỏ hai phía

$$n \geq \left[\frac{\sigma^2(u_\alpha + u_\beta)^2}{\Delta^2}\right]$$ nếu là miền bác bỏ một phía

$$n \geq \left[\frac{\sigma^2(u_{\alpha/2} + u_\beta)^2}{\Delta^2}\right]$$ nếu là miền bác bỏ hai phía

* Kiểm định T về μ (σ chưa biết)

$$H_0: \mu = \mu_0$$

Tiêu chuẩn kiểm định

$$T = \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \sim T(n-1)$$

$\beta = P[T < t_{\alpha}^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(\overline{x})}]$ nếu là kiểm định một phía

$\beta = P[T < t_{\alpha/2}^{(n-1)} - \frac{|\mu_0 - \mu_1|}{Se(\overline{x})}]$ nếu là kiểm định hai phía

$n \geq \left[\frac{S^2}{\Delta^2}(t_{\alpha}^{(m-1)} + t_{\beta}^{(m-1)}\right]$ nếu là kiểm định một phía

$n \geq \left[\frac{S^2}{\Delta^2}(t_{\alpha/2}^{(m-1)} + t_{\beta}^{(m-1)}\right]$ nếu là kiểm định hai phía

* Kiểm định hai kỳ vọng toán: $H_0: \mu_1 = \mu_2$

+ Trường hợp đã biết σ_1^2 và σ_2^2

Tiêu chuẩn kiểm định

$$U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

+ Trường hợp chưa biết σ_1^2 và σ_2^2 ($\sigma_1^2 = \sigma_2^2$)

Tiêu chuẩn kiểm định

$$T = \frac{(\overline{x}_1 - \overline{x}_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2)$$

+ Trường hợp chưa biết σ_1^2 và σ_2^2 $(\sigma_1^2 \neq \sigma_2^2)$

Tiêu chuẩn kiểm định

$$T = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T(k)$$

+ Trường hợp hai mẫu phụ thuộc theo từng cặp. Tiêu chuẩn kiểm định

$$T = \frac{\bar{D}\sqrt{n}}{S_D} \sim T(n-1)$$

* Kiểm định về P; H_0: $p = p_0$

Tiêu chuẩn kiểm định

$$U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

* Kiểm định về hai tham số p; H_0: $p_1 = p_2$

Tiêu chuẩn kiểm định

$$U = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

trong đó $\bar{f} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$

* Kiểm định về σ^2, H_0: $\sigma^2 = \sigma_0^2$

Tiêu chuẩn kiểm định

$$\chi_2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

* Kiểm định về hai tham số σ^2; H_0: $\sigma_1^2 = \sigma_2^2$ tiêu chuẩn kiểm định

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

* Kiểm định k phương sai

$$H_0: \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$

+ Trường hợp kích thước mẫu khác nhau

Tiêu chuẩn kiểm định Bartlett

$$B = \frac{V}{C} \sim \chi^2(k-1)$$

trong đó $V = 2{,}303\left[h.\lg \overline{S}^2 - \sum_{i=1}^{k} h_i \lg S_i^2\right]$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k} \frac{1}{h_i} - \frac{1}{h}\right]$$

+ Trường hợp kích thước mẫu bằng nhau

Tiêu chuẩn kiểm định Cochran:

$$G = \frac{S_{max}^2}{S_1^2 + S_2^2 + ... + S_k^2}$$

* Kiểm định χ^2 về tính độc lập của 2 dấu hiệu định tính

H_0: A và B độc lập

Tiêu chuẩn kiểm định

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{X_{ij}^2}{n_i.m_j} - 1\right] \sim \chi^2[(k-1)(h-1)]$$

* Kiểm định χ^2 về k tham số p.

$$H_0: p_1 = p_2 = \ldots = p_k$$

Tiêu chuẩn kiểm định

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{2}\frac{X_{ij}^2}{n_i.m_j} - 1\right] \sim \chi^2(k-1)$$

* Kiểm định χ^2 về quy luật phân phối xác suất

H_0: X phân phối theo quy luật A

Tiêu chuẩn kiểm định

$$\chi^2 = \sum_{i=1}^{k}\frac{(n_i - n'_i)^2}{n'_i} \sim \chi^2(k-r-1)$$

* Kiểm định Kolmogorov về quy luật phân phối xác suất

Tiêu chuẩn kiểm định:

$$\lambda = \sqrt{n}D = \sqrt{n}.\max_{X_i}\left|F^*(X_i) - F(X_i)\right|$$

* Kiểm định Jarque - Bera về phân phối chuẩn

H_0: X phân phối chuẩn

Tiêu chuẩn kiểm định

$$JB = n\left[\frac{a_3^2}{6} + \frac{(a_4 - 3)^2}{24}\right] \sim \chi^2(2)$$

*Kiểm định tổng hạng Wilcoxon (mẫu lớn)

H_0: Hai tổng thể có phân phối giống nhau

Tiêu chuẩn kiểm định

$$U = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$$

* Kiểm định tổng hạng theo dấu Wilcoxon (mẫu lớn)

H_0: Các hiệu số D_i đối xứng qua giá trị 0.

Tiêu chuẩn kiểm định

$$U = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$$

* Kiểm định Kruskal - Wallis về k kỳ vọng toán

H_0: k phân phối trong các tổng thể là giống nhau.

Tiêu chuẩn kiểm định

$$H = \frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n_T+1) \sim \chi^2(k-1)$$

* Kiểm định đoạn mạch

+ Trường hợp một mẫu

H_0: Mẫu được lập một cách ngẫu nhiên

Tiêu chuẩn kiểm định

$$U = \frac{R - E(R)}{Se(R)} \sim N(0,1)$$

+ Trường hợp hai mẫu

H_0: Hai tổng thể nghiên cứu có cùng phân phối xác suất

Tiêu chuẩn kiểm định Wald - Wolfowitz

$$U = \frac{R - E(R)}{Se(R)} \sim N(0,1)$$

Câu hỏi ôn tập

1. Khi kiểm định một giả thuyết thống kê, nếu kết luận giá trị của thống kê mẫu khác biệt một cách có ý nghĩa so với giá trị giả thuyết thì điều đó có nghĩa là gì ?

2. Sai lầm loại một có ý nghĩa như thế nào? Nó ảnh hưởng như thế nào đến giá trị tới hạn của kiểm định?

3. Khi tăng kích thước của mẫu lên thì điều đó sẽ ảnh hưởng như thế nào đến:

a) Xác suất mắc sai lầm loại I ?

b) Giá trị tới hạn ?

Hãy mô tả điều đó bằng đồ thị.

4. Sai lầm loại II là gì? Quan hệ của nó đối với sai lầm loại I như thế nào?

5. Phát biểu sau đây là đúng hay sai

P (sai lầm loại I) + P (sai lầm loại II) = 1

6. Tại sao khi P-value có giá trị càng nhỏ thì ta lại càng nghiêng về việc bác bỏ giả thuyết H_0.

7. Độ lệch chuẩn của mẫu có ảnh hưởng như thế nào đến giá trị tới hạn của kiểm định?

8. Hãy phát biểu cặp giả thuyết thống kê cho các tình huống sau đây:

a) Một nhà sản xuất kẹo tuyên bố là trọng lượng trung bình của mỗi gói kẹo là 500 gram. Kiểm tra ngẫu nhiên 500 gói kẹo tìm được trọng lượng trung bình mỗi gói là 450 gram và độ lệch chuẩn là 100 gram.

b) Một nhà sản xuất tủ lạnh tuyên bố là tỷ lệ tủ lạnh phải bảo hành khi sử dụng không vượt quá 3%. Theo dõi ngẫu nhiên 170 tủ lạnh đã bán ra thấy có 12 chiếc phải bảo hành.

c) Trước chiến dịch quảng cáo, điều tra ngẫu nhiên 10 tuần tìm được doanh số trung bình đối với một loại mỹ phẩm là 35 triệu/tuần và độ lệch chuẩn là 3 triệu. Sau chiến dịch quảng cáo, theo dõi ngẫu nhiên 12 tuần tìm được doanh số trung bình là 37 triệu/tuần và độ lệch chuẩn là 4 triệu.

d) Ở địa phương A, xét nghiệm ngẫu nhiên 500 người thấy 50 người có ký sinh trùng sốt rét. Ở địa phương B, xét nghiệm ngẫu nhiên 1000 người thì thấy 120 người có ký sinh trùng sốt rét.

9. Phân biệt sự khác nhau giữa kiểm định tham số và kiểm định phi tham số.

10. Phân biệt sự khác nhau giữa giả thuyết đơn và giả thuyết kép. Cho ví dụ.

11. Tại sao khi giá trị quan sát của tiêu chuẩn kiểm định không thuộc vào miền bác bỏ giả thuyết thì ta lại chỉ có thể nói rằng chưa có cơ sở để bác bỏ giả thuyết H_0.

Chương IX

PHÂN TÍCH PHƯƠNG SAI

§1. ĐẶT VẤN ĐỀ

Giả sử chúng ta cần nghiên cứu, phân tích sự biến động của một tổng thể thông qua một biến ngẫu nhiên (chỉ tiêu nghiên cứu). Thông thường, như đã xét ở các chương trước, chúng ta có các bài toán ước lượng, kiểm định giả thuyết về giá trị tham số và quy luật phân phối xác suất của biến ngẫu nhiên đó. Tuy nhiên, phương pháp kiểm định, so sánh tham số đó chỉ sử dụng được khi sự biến động của biến ngẫu nhiên đó chỉ chịu tác động của một nhân tố và cũng ở một hoặc tối đa hai mức cố định. Nếu sự biến động đó được tạo nên do nhiều nhân tố cùng tác động hoặc do một nhân tố nhưng ở nhiều mức độ khác nhau thì phải tiến hành phân tích phương sai mới thấy được vai trò ảnh hưởng của từng mức nhân tố, hoặc từng nhân tố cũng như sự ảnh hưởng tổng hợp của các nhân tố (nếu có) trong việc tạo nên sự biến đổi, sự sai khác đó.

Nhân tố ở đây được hiểu là các yếu tố, điều kiện khách quan (thời tiết, khí hậu, thiên tai...) hoặc chủ quan (phương pháp thí nghiệm, giống cây trồng, máy móc sử dụng...) có tác động trực tiếp đến sự biến động của biến ngẫu nhiên mà ta

nghiên cứu. Ví dụ, chúng ta cần phân tích sự biến động của năng suất một loại cây trồng (X) dưới tác động của hai nhân tố: Nơi trồng (ký hiệu là F) và giống (ký hiệu là G). Ta có X là một biến ngẫu nhiên định lượng và thông thường F và G là các biến (ngẫu nhiên hoặc không ngẫu nhiên) định tính. Do đó các giá trị của nhân tố (trong thống kê người ta thường gán cho chúng bằng những số nguyên 1, 2...) chỉ có ý nghĩa phân loại và trong phân tích phương sai chúng ta gọi đó là các mức của nhân tố. Ví dụ, F có p giá trị ($F = \overline{1,p}$) tức là F có p mức khác nhau. F = 1 có nghĩa là cây trồng được trồng ở nơi được đánh số là 1, F = 2 là nơi được đánh số là 2... ta cũng giả thiết rằng G có q giá trị (q mức) tức là loại cây trồng trên có q giống khác nhau được đem ra trồng thí nghiệm. Ta có thể mô tả bằng sơ đồ sau đây:

$$X = (\text{Năng suất cây trồng}) \xrightarrow{\text{Nhân tố tác động}} \{X_{ij}\}\ i = \overline{1,p}$$

$$F = (\text{Nơi trồng}) \qquad j = \overline{1,q}$$

$$G = (\text{Giống})$$

Trong đó X_{ij} = (Năng suất cây trồng ở địa điểm i và của giống j).

Nếu chúng ta chỉ xét một nhân tố G có 2 mức (q = 2) thì chúng ta có bài toán so sánh năng suất trung bình của hai giống cây trồng (xem chương VIII). Với q > 2 và chúng ta lại giả thiết rằng điều kiện địa lý của các lô ruộng trồng các loại giống cây trồng nói trên là như nhau, phương pháp canh tác ở những lô ruộng đó cũng như nhau..., nói chung các điều kiện khác đều giống nhau (điều này trên thực tế rất ít khi xảy ra) thì chúng ta có bài toán so sánh năng suất trung bình của q giống cây trồng khác nhau...

Tùy theo *số lượng nhân tố* và *điều kiện tác động* của chúng lên chỉ tiêu nghiên cứu (biến ngẫu nhiên) mà người ta có những bài toán phân tích phương sai khác nhau cùng với những mô hình phân tích phương sai khác nhau. Tệp số liệu dùng để minh họa cho các mô hình phân tích phương sai sẽ được trình bày dưới đây có tên gọi là XSTK9_1. Trong tệp số liệu này F có 4 mức (p = 4) và G có 2 mức (q = 2), sau đây là nội dung tệp số liệu. (Xem bảng 9.1).

Bảng 9.1

Contains data from C:\WINSTATA\XSTK9_1.DTA

obs: 13

vars: 3 15 Aug 1998 10:54

size: 130 (84.4% of memory free)

1. x	float %9.0g	năng suất
2. f	byte %8.0g	nơi trồng
3. g	byte %8.0g	giống

sorted by:

• list

	x	f	g
1.	1.38	1	1
2.	1.38	1	1
3.	1.42	1	2
4.	1.42	1	2
5.	1.41	2	1

6.	1.42	2	2
7.	1.44	2	1
8.	1.45	2	2
9.	1.32	3	1
10.	1.33	3	1
11.	1.34	3	2
12.	1.31	4	2
13.	1.33	4	1

§2. MÔ HÌNH PHÂN TÍCH PHƯƠNG SAI MỘT NHÂN TỐ

Xét biến ngẫu nhiên X tuân theo quy luật phân phối chuẩn $N(\mu, \sigma^2)$ và một nhân tố F tác động lên X có p mức khác nhau. Như vậy, ứng với mỗi mức nhân tố i ta có biến ngẫu nhiên X_i và chúng cũng tuân theo quy luật phân phối chuẩn $N(\mu_i, \sigma_i^2)$ (ví dụ như X_i là năng suất cây trồng ở nơi thứ i). Nếu tiến hành quan sát X_i bằng cách lấy một mẫu ngẫu nhiên kích thước n_i:

$(X_{1i}, X_{2i}, ..., X_{ki}, ..., X_{n_i i})$ $(i = \overline{1, p})$ thì ta có thể viết:

$$X_{ki} = \mu + \alpha_i + U_{ki} \qquad k = \overline{1, n_i} \tag{9.1}$$

Trong đó α_i đặc trưng cho sự khác biệt về giá trị trung bình μ của biến ngẫu nhiên X dưới tác động của nhân tố F ở mức i và U_{ki} là các sai số ngẫu nhiên giả thiết là độc lập với nhau, cùng tuân theo quy luật phân bố chuẩn $N(0, \sigma_u^2)$.

Mô hình (9.1) được gọi là mô hình phân tích phương sai một nhân tố. Để tiến hành phân tích phương sai ta xét cặp giả thuyết sau đây:

H_0: ($\alpha_i = 0\ \forall i, i = \overline{1,p}$)

H_1: (Tồn tại ít nhất một $\alpha_i \neq 0$)

Cặp giả thuyết trên tương đương với cặp giả thuyết sau đây:

H_0: ($\mu_1 = \mu_2 = ... = \mu_p$)

H_1: (Tồn tại ít nhất một cặp ($i \neq i'$) sao cho $\mu_i \neq M_i'$)

Bảng số liệu dùng để tính toán phân tích phương sai được xây dựng như sau:

Nhân tố F / STT quan sát	1	2	...	i	...	p
1	x_{11}	x_{12}	.	x_{1i}	.	x_{1p}
2	x_{21}	x_{22}	.	x_{2i}	.	x_{2p}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n_i	n_1	n_2		n_i		n_p

Ví dụ 9.1: Với số liệu đã cho ở bảng 9.1 ta có bảng số liệu dùng để phân tích phương sai trường hợp một nhân tố như sau:

F / STT quan sát	1	2	3	4
1	1,38	1,41	1,32	1,31
2	1,38	1,42	1,33	1,33
3	1,42	1,44	1,34	
4	1,42	1,45		
n_i	$n_1 = 4$	$n_2 = 4$	$n_3 = 3$	$n_4 = 2$

Phương pháp phân tích phương sai về cơ bản dựa trên cơ sở tính toán và phân tích một số đặc trưng mẫu sau đây.

- TSS (Total sum of squares) là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát X_{ki} và giá trị trung bình mẫu chung của chúng. Ký hiệu:

$$\overline{X}_i = \left(\sum_{k=1}^{n_i} X_{ki}\right) / n_i \qquad i = \overline{1,p}$$

$$\overline{\overline{X}} = \left(\sum_i \sum_k X_{ki}\right) / n \quad n = \sum_i n_i, \quad i = \overline{1,p}, k = \overline{1,n_i}$$

Ta có:

$$TSS = \sum_i \sum_k (X_{ki} - \overline{\overline{X}})^2 = \sum_i Q_i - \left[\sum_i T_i\right]^2 / n \qquad (9.2)$$

Trong đó: $Q_i = \sum_k X_{ki}^2$; $T_i = \sum_k X_{ki}$

- MSS (Model sum of squares) là tổng bình phương các sai lệch giữa các giá trị trung bình mẫu của các nhóm quan sát (phân theo mức nhân tố i) và trung bình mẫu chung.

$$MSS = \sum_i (\overline{X}_i - \overline{\overline{X}})^2 n_i = \sum_i (T_i^2 / n_i) - \left[\sum_i T_i\right]^2 / n \qquad (9.3)$$

- RSS (Residual sum of squares) là tổng của tất cả các tổng bình phương các sai lệch của các giá trị X_{ki} so với trung bình mẫu của từng nhóm:

$$RSS = \sum_i \left(\sum_k (X_{ki} - \overline{X}_i)^2\right) = \sum_i Q_i - \sum_i (T_i^2 / n_i) \qquad (9.4)$$

Dễ dàng thấy rằng TSS = MSS + RSS và số bậc tự do của

TSS là (n – 1) do đó số bậc tự do của MSS và RSS tương ứng là (p – 1) và (n – p).

Như vậy, ta thấy toàn bộ sự khác biệt của các giá trị X_{ki} so với $\overline{X}$ được chia làm hai phần: Một phần được tạo nên bởi nhân tố trong mô hình (MSS) và phần còn lại là do bản thân sự khác biệt giữa các giá trị quan sát tạo ra (RSS) (Đó là sự khác biệt hoàn toàn do ngẫu nhiên mà có).

Trở lại ví dụ 9.1, ta lập bảng tính toán sau đây:

Bảng 9.2

STT quan sát \ F	1		2		3		4		Kết quả
	x_{k1}	x_{k1}^2	x_{k2}	x_{k2}^2	x_{k3}	x_{k3}^2	x_{k4}	x_{k4}^2	
1	1,38	1,9044	1,41	1,9881	1,32	1,7424	1,31	1,7161	$n_1 = n_2 = 4$
2	1,38	1,9044	1,42	2,0164	1,33	1,7689	1,33	1,7689	$n_3 = 3, n_4 = 2$
3	1,42	2,0164	1,44	2,0736	1,34	1,7956			$n = \sum n_i = 13$
4	1,42	2,0164	1,45	2,1025					
$T_i = \sum_k X_{ki}$	5,6		5,72		3,99		2,64		$\sum_i T_i = 17,95$
$Q_i = \sum_k X_{ki}^2$		7,7416		8,1806		5,3069		3,485	$\sum_i Q_i = 24,8141$
$\frac{T_i^2}{n_i}$	7,84		8,1796		5,3067		3,4848		$\sum_i \frac{T_i^2}{n_i} = 24,8111$

Từ kết quả tính toán trên và các công thức (9.2) (9.3) (9.4) ta có:

$$TSS = \sum Q_i - \left[\sum T_i\right]^2 / n = 0{,}029292$$

$$MSS = \sum (T_i^2 / n_i) - \left[\sum T_i\right]^2 / n = 0{,}026292$$

$$RSS = TSS - MSS = 0{,}003$$

Để kiểm định cặp giả thuyết của mô hình (9.1):

H_0: $(\alpha_i = 0\ \forall i,\ i = \overline{1,p})$

H_1: (Tồn tại ít nhất một $\alpha_i \neq 0$)

ta sử dụng tiêu chuẩn sau đây:

$$F = \frac{MSS/(p-1)}{RSS/(n-p)}$$

Nếu H_0 đúng, F tuân theo quy luật Fisher – Snedecor với bậc tự do là (p – 1) và (n – p). Ta có miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{F = \frac{MSS/(p-1)}{RSS/(n-p)}; F > f_\alpha(p-1, n-p)\right\}$$

Trong đó: $f_\alpha(p-1, n-p)$ là giá trị tới hạn mức α của phân phối $F(p-1, n-p)$.

Đối với ví dụ 9.1 ta có:

$$F_{qs} = \frac{0{,}026292/(4-1)}{0{,}003/(13-4)} = \frac{0{,}08764}{0{,}000333} = 26{,}31$$

Chọn $\alpha = 0{,}05$, tra bảng ta được $f_{0{,}05}(3,9) = 3{,}86 \Rightarrow F_{qs} \in W_\alpha$

Vậy ta bác bỏ giả thiết về sự bằng nhau của các giá trị năng suất trung bình của giống cây trồng nói trên được trồng ở bốn nơi khác nhau. Nói một cách khác năng suất cây trồng sẽ khác nhau có ý nghĩa nếu được trồng ở bốn nơi có điều

kiện môi trường khác nhau nói trên. Tuy nhiên, như chúng ta đã nhận xét từ đầu, năng suất cây trồng không thể chỉ chịu tác động của một yếu tố về địa lý (nơi trồng) mà còn chịu tác động của nhiều nhân tố khác nữa, ví dụ như: Giống, phương pháp canh tác... Vậy nếu chỉ xét 1 nhân tố địa lý (F) thì mức độ tác động của nó đối với sự biến động của năng suất cây trồng (X) là bao nhiêu? Để trả lời câu hỏi này chúng ta xét biểu thức sau đây:

$$TSS = MSS + RSS$$

Chia cả hai vế cho TSS ta được:

$$1 = 100\% = \frac{MSS}{TSS} + \frac{RSS}{TSS}$$

$$\text{Đặt } R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Như vậy ta thấy ý nghĩa của R^2 chính là tỷ lệ hay số phần trăm chiếm trong tổng số 100% của toàn bộ sự sai lệch của X_{ki} so với giá trị trung bình của chúng. R^2 được sử dụng để đo mức độ ảnh hưởng của các nhân tố chứa trong mô hình đối với sự biến động của các giá trị của biến ngẫu nhiên X xung quanh giá trị trung bình của nó. R^2 được gọi là *hệ số xác định* của mô hình phân tích phương sai, đó cũng chính là mức độ thích hợp của mô hình. R^2 càng lớn mô hình càng thích hợp, càng giải thích được nhiều hơn sự biến động của các giá trị của biến ngẫu nhiên X dưới tác động của các nhân tố có trong mô hình.

Trở lại ví dụ đã nêu trên, chúng ta có:

$$R^2 = \frac{MSS}{TSS} = \frac{0{,}026292}{0{,}029292} = 0{,}8976$$

Điều này có nghĩa là nhân tố nơi trồng (F) ảnh hưởng tới 89,76% đến sự biến động của năng suất cây trồng (89,76% sự khác biệt về năng suất trung bình ở các nơi trồng khác nhau được giải thích bởi yếu tố địa lý - nhân tố F).

Tất cả những kết quả phân tích trên có thể thực hiện một cách đơn giản bởi lệnh anova của Stata. Sau đây là kết quả giải bài toán trên (mô hình 9.1) bằng máy tính để các bạn tham khảo và so sánh. Chú ý: So sánh kết quả tính bằng tay sẽ có một chút sai số không đáng kể do cách làm tròn số.

```
anova x f

Number of obs = 13      R-squared = 0.8976

Root MSE = .018257 Adj R-squared = 0.8634
```

Source	Partial SS	df	MS	F	Prob > F
Model	.026292296	3	.008764099	26.29	0.0001
f	.026292296	3	.008764099	26.29	0.0001
Residual	.003000004	9	.000333334		
Total	.029292299	12	.002441025		

§3. MÔ HÌNH PHÂN TÍCH PHƯƠNG SAI HAI NHÂN TỐ

Giả sử chúng ta xét biến ngẫu nhiên X có phân phối chuẩn và hai nhân tố F, G. Nhân tố F có p mức $(i = \overline{1,p})$ và nhân tố G có q mức $(j = \overline{1,q})$. Chúng ta có hai mô hình phân tích phương sai sau đây:

3.1. Mô hình hai nhân tố tác động riêng rẽ

Trong mô hình này chúng ta đã giả thiết rằng F và G cùng tác động lên X, nhưng một cách riêng rẽ. Tức là giữa F và G không có mối quan hệ cùng tác động đồng thời lên X. Ta có mô hình sau đây:

$$X_{ij} = \mu + \alpha_i + \beta_j + U_{ij} \qquad i = \overline{1,p};\ j = \overline{1,q}$$

Trong đó: α_i, β_j là các hằng số đặc trưng cho sự khác biệt về giá trị trung bình μ của X dưới tác động của hai nhân tố F và G tương ứng ở mức i và j.

U_{ij} là các sai số ngẫu nhiên độc lập với nhau và cùng tuân theo quy luật phân bố chuẩn $N(0, \sigma_u^2)$.

Với mô hình loại này ta có bảng số liệu phân tích phương sai như sau:

F \ G	1	2	...	j	...	q
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1q}
.	.	.		.		.
.	.	.		.		.
i	X_{i1}	X_{i2}		X_{ij}		X_{iq}
.	.	.		.		.
.	.	.		.		.
p	X_{p1}	X_{p2}		X_{pj}		X_{pq}

Dễ dàng nhận thấy chúng ta có pq nhóm, trong mỗi một nhóm chỉ có một quan sát. Nếu trong một nhóm có nhiều quan sát ta lấy giá trị trung bình của nhóm đó làm đại diện.

Ký hiệu:

$$\overline{X} = (\sum_i \sum_j X_{ij})/pq$$

$$\overline{X}_i = (\sum_j X_{ij})/q = \frac{X_{i\bullet}}{q}$$

$$\overline{X}_j = (\sum_i X_{ij})/p = \frac{X_{j\bullet}}{p}$$

Ta có tổng bình phương các sai lệch chung của toàn bộ quan sát so với giá trị trung bình mẫu là:

$$TSS = \sum_i \sum_j (X_{ij} - \overline{X})^2 = \sum_{i,j} X_{ij}^2 - pq(\overline{X})^2 \qquad (9.5)$$

Vế phải có thể biến đổi như sau:

$$\sum_i \sum_j (X_{ij} - \overline{X})^2 = \sum_i q(\overline{X}_i - \overline{X})^2 + \sum_j p(\overline{X}_j - \overline{X})^2$$
$$+ \sum_i \sum_j (X_{ij} - \overline{X}_i - \overline{X}_j + \overline{X})^2$$

Trong đó tổng bình phương các sai lệch mà trung bình mẫu các nhóm so với trung bình mẫu chung (tức là sai lệch do việc phân chia thành các nhóm bởi hai nhân tố F và G) là:

$$MSS = \sum_i q(\overline{X}_i - \overline{X})^2 + \sum_j p(\overline{X}_j - \overline{X})^2$$

$$= \left(\frac{\sum X_{i\bullet}^2}{q} - \frac{(\sum X_{ij})^2}{pq} \right) + \left(\frac{\sum X_{j\bullet}^2}{p} - \frac{(\sum X_{ij})^2}{pq} \right) \qquad (9.6)$$

$$= MSS_F + MSS_G$$

Trong đó: MSS_F là tổng bình phương các sai lệch về trung bình của các nhóm được tạo nên bởi nhân tố F và MSS_G là tổng bình phương các sai lệch về trung bình của các nhóm được tạo nên bởi nhân tố G.

Phần còn lại chính là tổng bình phương các sai lệch được tạo nên bởi sai số ngẫu nhiên:

$$RSS = \sum_i \sum_j (X_{ij} - \overline{X}_i - \overline{X}_j + \overline{X})^2 =$$

$$= TSS - MSS_F - MSS_G \qquad (9.7)$$

Bậc tự do của TSS là (qp – 1), của MSS_F là (p – 1), của MSS_G là (q – 1), do đó bậc tự do của RSS là:

$$pq - 1 - (p - 1) - (q - 1) = (p - 1)(q - 1).$$

Ví dụ 9.2:

Tệp số liệu cho ví dụ này có nội dung trong bảng 9.3.

Bảng 9.3

```
• d
Contains data from C:\WINSTATA\XSTK9_2.DTA
obs:          12
vars:          3                                  15 Aug 1998
10: 37
-------------------------------------------------------------------------------
size:         84 (86.2% of memory free)
1.y           byte %8.0g       chi phi hoc hanh
2. f          byte %8.0g       thanh phan gia dinh
3.g           byte %8.0g       vung
-------------------------------------------------------------------------------
```

Sorted by:

• list

	x	f	g
1.	22	1	1
2.	17	2	1
3.	28	3	1
4.	34	4	1
5.	21	1	2
6.	15	2	2
7.	27	3	2
8.	33	4	2
9.	31	1	3
10	19	2	3
11.	30	3	3
12.	35	4	3

Trong đó: X là chi phí học hành bình quân cho một đứa trẻ đi học của một gia đình. Hai nhân tố, F là thành phần gia đình (F = 1: Công nhân, F = 2: Nông dân, F = 3: Trí thức; F = 4: Buôn bán) và G là vùng, nơi cư trú (G = 1: Bắc, G = 2: Trung ; G = 3: Nam). Từ tập số liệu này ta có bảng số liệu phân tích phương sai hai nhân tố tác động riêng rẽ như sau:

F \ G	1	2	3
1	22	21	31
2	17	15	19
3	28	27	30
4	34	33	35

Để giải bài toán trên, chúng ta lập bảng tính sau đây:

G \ F	1	2	3	4	$X_{j\bullet}$	$X_{j\bullet}^2$
1	22	17	28	34	101	10201
2	21	15	27	33	96	9216
3	31	19	30	35	115	13225
$X_{i\bullet}$	74	51	85	102	$\sum X_{ij} = 312$	$\sum X_{j\bullet}^2 = 32642$
$\sum_j X_{ij}^2$	1886	875	2413	3470	$\sum X_{ij}^2 = 8644$	
$X_{i\bullet}^2$	5476	2601	7225	10404	$\sum X_{i\bullet}^2 = 25706$	

Từ những kết quả tính được ở bảng trên thay vào công thức (9.5), (9.6), (9.7) ta có:

$$MSS_F = \frac{25706}{3} - \frac{(312)^2}{12} = 456,66$$

$$MSS_G = \frac{32642}{4} - \frac{(312)^2}{12} = 48,5$$

$$TSS = 8644 - \frac{(312)^2}{12} = 532$$

Vậy ta suy ra: RSS = 532 – 456,66 – 48,5 = 26,84

Kiểm định giả thuyết của mô hình

a. Cặp giả thuyết:

$$H_0^F: (\alpha_i = 0 \quad \forall i,\ i = \overline{1,p})$$

$$H_1^F: (\text{Có ít nhất một } \alpha_i \neq 0)$$

Nếu H_0^F đúng thì $F = \frac{MSS_F/(p-1)}{RSS/(p-1)(q-1)}$ sẽ tuân theo quy luật Fisher – Snedecor với bậc tự do là (p – 1) và (p – 1)(q – 1)

Do đó miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{MSS_F/(p-1)}{RSS/(p-1)(q-1)}; F > f_\alpha[(p-1),(p-1)(q-1)] \right\}$$

Đối với ví dụ 9.2 thì:

$$F_{qs} = \frac{456{,}66/(4-1)}{26{,}84/(4-1)(3-1)} = 34{,}04$$

Chọn $\alpha = 0{,}05$ tra bảng ta được $f_{0{,}05}(3,6) = 4{,}76 \rightarrow F_{qs} \in W_\alpha$.

Ta bác bỏ giả thuyết H_0^F, có nghĩa là thành phần gia đình (nhân tố F) có ảnh hưởng đến mức chi tiêu cho học hành của một gia đình.

b. Cặp giả thuyết:

$$H_0^G: (\beta_j = 0 \quad \forall j, j = \overline{1,q})$$

$$H_1^G: (\text{Có ít nhất một } \beta_j \neq 0)$$

Tương tự ta có miền bác bỏ:

$$W_\alpha = \left\{ F = \frac{MSS_G/(q-1)}{RSS/(p-1)(q-1)}; F > f_\alpha[(q-1),(p-1)(q-1)] \right\}$$

Đối với ví dụ 9.2 thì:

$$F_{qs} = \frac{48{,}5/(3-1)}{26{,}84/6} = 5{,}42$$

Tra bảng ta có $f_{0.05}(2,6) = 5{,}14 \rightarrow F_{qs} \in W_\alpha$ do đó ta bác bỏ giả thuyết H_0^G. Điều này có nghĩa là chúng ta có thể cho rằng

nhân tố vùng cũng có ảnh hưởng đến mức chi tiêu cho học hành của một gia đình.

Sau đây là kết quả giải bài toán phân tích phương sai hai nhân tố tác động riêng rẽ trên cơ sở tệp số liệu XSTK 9_2 đã nêu trên bằng Stata.

```
anova y f g
```

Number of obs = 12 R – squared = 0.9496

Root MSE = 2.11476 Adj R – squared = 0.9075

Source	Partial SS	df	MS	F	Prob > F
Model	505.166667	5	101.033333	22.59	0.0008
f	456.666667	3	152.222222	34.04	0.0004
g	48.50	2	24.25	5.42	0.0452
Residual	26.8333333	6	4.47222222		
Total	532.00	11	48.3636364		

Cũng như đối với mô hình phân tích phương sai một nhân tố, ta có hệ số xác định của mô hình phân tích phương sai hai nhân tố tác động riêng rẽ là R^2 được xác định như sau:

$$R^2 = \frac{MSS}{TSS} = \frac{MSS_F + MSS_G}{TSS}$$

Theo ví dụ 10.2 thì:

$$R^2 = \frac{456,66 + 48,5}{532} = \frac{505,15}{532} = 0,9495$$

Có nghĩa là mô hình đưa ra đã giải thích được 94,95% sự biến động của giá trị chi phí cho học hành trung bình của một gia đình dưới tác động của hai nhân tố là thành phần gia

đình (F) và vùng cư trú (G). Trong đó F đóng góp 85,83% và G đóng góp 9,12% vào sự biến động của X.

3.2. Mô hình phân tích phương sai hai nhân tố tác động tổng hợp

Trong mô hình này ở mỗi một nhóm tương ứng với một cặp mức nhân tố (i, j) ta cần phải có số quan sát lớn hơn 1. Giả sử ở tất cả các nhóm ta đều có số quan sát là như nhau và bằng m (còn gọi là mẫu cân bằng) thì ta có mô hình hai nhân tố tác động tổng hợp như sau:

$$X_{kij} = \mu + \alpha_i + \beta_j + \delta_{ij} + U_{kij}$$
$$k = \overline{1,m} \; ; i = \overline{1,p} \; ; j = \overline{1,q}$$

Trong đó: α_i, β_j là các hằng số đặc trưng cho sự khác biệt về giá trị trung bình μ của X do tác động của hai nhân tố F và G, còn δ_{ij} là hằng số đặc trưng cho sự khác biệt đó nhưng được gây nên bởi tác động tổng hợp của hai nhân tố (F, G) ở mức (i, j). Ta cũng giả thiết $X \sim N(\mu, \sigma_x^2)$ và các sai số ngẫu nhiên U_{kij} độc lập với nhau, cũng tuân theo quy luật phân bố chuẩn $N(0, \sigma_u^2)$.

Ký hiệu:

$$\overline{X} = \left(\sum_{k,i,j} X_{kij}\right) / mqp$$

$$\overline{X}_i = \left(\sum_{k,j} X_{kij}\right) / mq = X_{i\bullet} / mq$$

$$\overline{\overline{X}}_j = \left(\sum_{k,i} X_{kij}\right) / mp = X_{j\bullet} / mp$$

$$\overline{X}_{ij} = \left(\sum_{k} X_{kij}\right) / m = X_{ij\bullet} / m$$

Xét tương tự như mục 3.1 ta có:

$$TSS = \sum_{k,i,j}(X_{kij} - \bar{X})^2 = \sum_{k,i,j} X_{kij}^2 - mpq(\bar{X})^2$$

$$= \sum X_{kij}^2 - \frac{(\sum X_{kij})^2}{mpq}$$

$$MSS_F = \sum_i mq(\bar{X}_i - \bar{X})^2 = \sum_i mq(\bar{X}_i)^2 - mpq(\bar{X})^2$$

$$= \frac{\sum X_{i\bullet}^2}{mq} - \frac{(\sum X_{kij})^2}{mpq}$$

$$MSS_G = \sum_j mp(\bar{X}_j - \bar{X})^2 = \sum_j mp(\bar{X}_j)^2 - mpq(\bar{X})^2$$

$$= \frac{\sum X_{j\bullet}^2}{mp} - \frac{(\sum X_{kij})^2}{mpq}$$

$$RSS = \sum_{k,i,j}(X_{kij} - \bar{X}_{ij})^2 = \sum_{k,i,j} X_{kij}^2 - m\sum_{i,j}(\bar{X}_{ij})^2$$

$$= \sum X_{kij}^2 - \frac{\sum X_{ij\bullet}^2}{m} \qquad (9.8)$$

$$MSS_{F\times G} = TSS - MSS_F - MSS_G - RSS$$

Bậc tự do của TSS là (mpq – 1), của MSS_F là (p – 1), của MSS_G là (q – 1), RSS là pq(m – 1) và của $MSS_{F\times G}$ là (p – 1) (q – 1)

+ Phân tích sự tác động của hai nhân tố F và G tương đương với việc kiểm định các cặp giả thuyết sau đây:

a. H_0^F: ($\alpha_i = 0 \quad \forall i, i = \overline{1,p}$)

H_1^F: (Có ít nhất một $\alpha_i \neq 0$)

Miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{MSS_F/(p-1)}{RSS/pq(m-1)}; F > f_\alpha[p-1, pq(m-1)] \right\}$$

b. H_0^G : ($\beta_j = 0 \quad \forall j, j = \overline{1,q}$)

H_1^G : (Có ít nhất một $\beta_j \neq 0$)

$$W_\alpha = \left\{ F = \frac{MSS_F/(q-1)}{RSS/pq(m-1)}; F > f_\alpha[q-1, pq(m-1)] \right\}$$

c. H_0^{F*G} : ($\delta_{ij} = 0 \quad \forall i, j : i = \overline{1,p};\ j = \overline{1,q}$)

H_1^{F*G} : (Tồn tại ít nhất một cặp (i, j): $\delta_{ij} \neq 0$)

$$W_\alpha = \left\{ F = \frac{MSS_{F*G}/(p-1)(q-1)}{RSS/pq(m-1)}; F > f_\alpha[(p-1)(q-1), pq(m-1)] \right\}$$

+ Tương tự như phần 3.1 chúng ta cũng có hệ số xác định của mô hình trong trường hợp này là:

$$R^2 = \frac{MSS}{TSS} = \frac{MSS_F + MSS_G + MSS_{F*G}}{TSS}$$

Ví dụ 9.3. Chúng ta xét lại các chỉ tiêu và nhân tố như đã nêu ở ví dụ 9.2. Đó là: X = (Chi phí cho giáo dục bình quân cho một đứa trẻ đi học trong một gia đình), F = (Thành phần gia đình) và G = (Nơi cư trú – vùng). Khác với số liệu ở ví dụ 9.2, ở đây ứng với mỗi mức nhân tố chúng ta có hai quan sát (tức là số liệu của hai gia đình). Như vậy, trong mô hình phân tích phương sai hai nhân tố tác động tổng hợp ta có:

$$i = 1, 2, 3, 4 ; j = 1, 2, 3 ; m = 1, 2.$$

Sau đây là bảng số liệu cụ thể:

G \ F	1	2	3	4
1	22 ; 21	17 ; 19	28 ; 25	34 ; 33
2	21 ; 25	15 ; 18	27 ; 25	33 ; 31
3	31 ; 31	19 ; 20	30 ; 33	35 ; 33

Trước hết chúng ta lập bảng tính toán số liệu trung gian như sau (xem bảng 9.4).

Bảng 9.4

F \ G	1	2	3	4	$X_{j\bullet}$	$X_{j\bullet}^2$
1	22 ; 21	17 ; 19	28 ; 25	34 ; 33	199	39601
2	21 ; 25	15 ; 18	27 ; 25	33 ; 31	195	38025
3	31 ; 31	19 ; 20	30 ; 33	35 ; 33	232	53824
$X_{\bullet j}$	151	108	168	199	$\sum X_{kij} = 626$	$\sum X_{j\bullet}^2 = 131450$
$\sum_{k,j} X_{kij}^2$	3913	1960	4752	6609	$\sum X_{kij}^2 = 17234$	
$X_{i\bullet}^2$	22801	11664	28224	39601	$\sum X_{i\bullet}^2 = 102290$	
$\sum_{j} X_{ij\bullet}^2$	7809	3906	9482	13209	$\sum_{i,j} X_{ij\bullet}^2 = 34406$	

Theo các công thức (9.6) ta có:

$$TSS = 17234 - \frac{626^2}{24} = 905,83$$

$$MSS_F = \frac{102290}{6} - \frac{626^2}{24} = 720,17$$

$$MSS_G = \frac{131450}{8} - \frac{626^2}{24} = 103,08$$

$$RSS = 17234 - \frac{34406}{2} = 31$$

$$\Rightarrow MSS_{F*G} = 905,83 - 720,17 - 103,08 - 31 = 51,58$$

Giả sử chúng ta muốn kiểm định xem tồn tại hay không sự tác động tổng hợp của hai nhân tố F và G. Xét cặp giả thiết ở mục C, theo kết quả tính toán ở trên ta có:

$$F_{qs} = \frac{MSS_{F*G}/(p-1)(q-1)}{RSS/pq(m-1)} = \frac{51,58/6}{31/12} = 3,33$$

Với mức ý nghĩa $\alpha = 0,05$, tra bảng chúng ta được $f_{0,05}^{(6,12)} = 3$. Vì $F_{qs} > f_{0,05}^{(6,12)}$ nên bác bỏ giả thuyết H_0^{F*G}. Tức là chúng ta có thể cho rằng trên thực tế tồn tại sự tác động lẫn nhau của hai yếu tố thành phần gia đình và vùng đối với việc chi tiêu cho trẻ em đi học.

Chú ý: Bài toán phân tích phương sai trong trường hợp mô hình tổng quát hơn như số nhân tố lớn hơn hai hoặc số quan sát trong mỗi nhóm không như nhau cũng đã được giải quyết một cách đầy đủ (xem tài liệu tham khảo M.G.

Kendall, A. Stuart). Tuy nhiên đối với tất cả các bài toán phân tích phương sai việc giải nó trên máy tính là một việc hoàn toàn không có gì khó khăn.

Ví dụ, nếu sử dụng Stata thì để phân tích sự ảnh hưởng của các nhân tố F, G, L đối với biến X chúng ta chỉ cần gõ lệnh:

anova x f g l f*g f*l g*l

Đối với ví dụ 9.3 nói trên, kết quả chạy bằng Stata như sau:

anova x f g f*g

Number of obs = 24 R-squared = 0.9658

Root MSE = 1.60728 Adj R-squared = 0.9344

Source	Partial SS	df	MS	F	Prob > F
Model	874.833333	11	79.530303	30.79	0.0000
f	720.166667	3	0240.055556	92.92	0.0000
g	103.083333	2	51.5416667	19.95	0.0002
f*g	51.5833333	6	8.59722222	3.33	0.0362
Residual	31.00	12	2.58333333		
Total	905.833333	23	39.384058		

Các ký hiệu và công thức cơ bản

1. Mô hình phân tích phương sai một nhân tố

$$X_{ki} = \mu + \alpha_i + U_{ki} \qquad i = \overline{1,p};\ k = \overline{1,n_i};\ n = \sum n_i$$

Thành phần	Tổng bình phương các sai lệch	Bậc tự do	F
F	$MSS = \sum \frac{X_{i\bullet}^2}{n_i} - \frac{(\sum X_{ki})^2}{n}$	$p - 1$	$F = \frac{MSS/(p-1)}{RSS/(n-p)}$
U	$RSS = \sum X_{ki}^2 - \frac{\sum X_{i\bullet}^2}{n_i}$	$n - p$	
Σ	$TSS = \sum X_{ki}^2 - \frac{(\sum X_{ki})^2}{n}$	$n - 1$	

2. Mô hình phân tích phương sai hai nhân tố

a. Hai nhân tố tác động riêng rẽ

$$X_{ij} = \mu + \alpha_i + \beta_j + U_{ij} \qquad i = \overline{1,p}\ ;\ j = \overline{1,q}\ ;\ n = p.q$$

Thành phần	Tổng bình phương các sai lệch	Bậc tự do	F
F	$MSS_F = \frac{\sum X_{i\bullet}^2}{q} - \frac{(\sum X_{ij})^2}{pq}$	$p - 1$	$F_1 = \frac{MSS_F/(p-1)}{RSS/(p-1)(q-1)}$
G	$MSS_G = \frac{\sum X_{j\bullet}^2}{p} - \frac{(\sum X_{ij})^2}{pq}$	$q - 1$	$F_2 = \frac{MSS_G/(q-1)}{RSS/(p-1)(q-1)}$
U	$RSS = TSS - MSS_F - MSS_G$	$(p-1)\times(q-1)$	
Σ	$TSS = \sum X_{ij}^2 - \frac{(\sum X_{ij})^2}{pq}$	$pq - 1$	

b. Hai nhân tố tác động tổng hợp

$$X_{kij} = \mu + \alpha_i + \beta_j + \delta_{ij} + U_{kij}$$

$$i = \overline{1,p}\,;\ j = \overline{1,q}\,;\ k = \overline{1,m}\ ;\ n = mpq$$

Thành phần	Tổng bình phương các sai lệch	Bậc tự do	F
F	$MSS_F = \frac{\sum X_{i\bullet}^2}{mq} - \frac{(\sum X_{kij})^2}{mpq}$	p – 1	$F_1 = \frac{MSS_F/(p-1)}{RSS/pq(m-1)}$
G	$MSS_G = \frac{\sum X_{j\bullet}^2}{mp} - \frac{(\sum X_{kij})^2}{mpq}$	q – 1	$F_2 = \frac{MSS_G/(q-1)}{RSS/pq(m-1)}$
F*G	$MSS_{F*G} = TSS - MSS_F - MSS_G - RSS$	(p–1) × (q–1)	$F_3 = \frac{MSS_{F*G}/(p-1)(q-1)}{RSS/pq(m-1)}$
U	$RSS = \sum X_{kij}^2 - \frac{\sum X_{ij\bullet}^2}{m}$	pq (m –1)	
Σ	$TSS = \sum X_{kij}^2 - \frac{(\sum X_{kij})^2}{mpq}$	mpq – 1	

Câu hỏi ôn tập

1. Hãy cho biết sự khác nhau giữa hai bài toán kiểm định giả thuyết về phương sai và phân tích phương sai.

2. Hãy nêu sự khác nhau của hai tiêu chuẩn χ^2 và F khi nghiên cứu về sự phân tán của tổng thể.

3. Mô hình phân tích phương sai một nhân tố và mô hình phân tích phương sai hai nhân tố khác nhau cơ bản ở điểm nào?

4. Bảng số liệu phân tích phương sai và bảng ngẫu nhiên hai chiều có điều gì giống nhau và khác nhau?

5. Hãy giải thích về số bậc tự do của TSS, RSS và các MSS trong phân tích phương sai.

6. Có thể giải thích như thế nào khi F_G càng lớn thì sự tác động của nhân tố G càng có ý nghĩa và ngược lại.

Chương X

PHÂN TÍCH TƯƠNG QUAN VÀ HỒI QUY

§1. ĐẶT VẤN ĐỀ

Khác với các chương trước, từ chương này trở đi chúng ta sẽ nghiên cứu trực tiếp nhiều chỉ tiêu khác nhau đối với một đối tượng nghiên cứu. Các chỉ tiêu này có thể là định lượng (tức là các biến ngẫu nhiên) hay định tính (được gọi là các dấu hiệu hay nhân tố như đã đề cập ở phần phân tích phương sai). Các chỉ tiêu này được xác định cụ thể trên đối tượng quan sát (cần phân biệt đối tượng nghiên cứu và đối tượng quan sát) và được "đo lường" (tức là thu thập th tin) trực tiếp trên các đơn vị quan sát. Đó là các phần tử thuộc tập hợp đối tượng quan sát mà ta sẽ trực tiếp thu nhận thông tin từ đó. Ví dụ đối tượng quan sát là lao động trong các ngành công nghiệp chẳng hạn thì đơn vị quan sát là các cán bộ, công nhân đang làm việc trong ngành công nghiệp. Tùy theo vấn đề cần nghiên cứu, phân tích mà các chỉ tiêu, có thể là tuổi, thời gian làm việc, năng suất lao động... (chỉ tiêu định lượng)... hay giới tính, tay nghề, loại hình nghề nghiệp, trình độ văn hóa... (chỉ tiêu định tính).

Như vậy để giải quyết một vấn đề cụ thể người ta phải nghiên cứu, phân tích nhiều chỉ tiêu (đôi khi còn gọi là chỉ

tiêu thống kê) đó là các biến ngẫu nhiên (chỉ tiêu định lượng), các nhân tố, dấu hiệu (chỉ tiêu định tính). Một phương pháp hữu hiệu giúp chúng ta có thể làm rõ hơn bản chất của hiện tượng hay sự việc cần nghiên cứu để tìm ra quy luật, dự đoán được xu thế biến động của hiện tượng, sự việc đó trong tương lai, đó là phương pháp phân tích mối quan hệ phụ thuộc giữa các chỉ tiêu phản ánh nội dung của hiện tượng, sự việc cần nghiên cứu. Cụ thể chúng ta sẽ đề cập đến những vấn đề sau:

1. Mức độ phụ thuộc giữa các biến ngẫu nhiên, giữa các nhân tố, dấu hiệu (Có tồn tại sự phụ thuộc giữa chúng không? Mức độ phụ thuộc đó như thế nào nếu chúng tồn tại).

2. Mối phụ thuộc đó thuộc loại nào? Trong thống kê người ta phân biệt hai loại phụ thuộc sau đây:

+ *Định nghĩa 1:* Hai biến ngẫu nhiên X và Y được gọi là phụ thuộc hàm nếu tồn tại hàm f sao cho $Y = f(X)$.

+ *Định nghĩa 2:* Hai biến ngẫu nhiên X, Y được gọi là có phụ thuộc thống kê nếu với mỗi giá trị của X ta đều có thể xác định được quy luật phân phối xác suất có điều kiện của Y đối với X là:

$$F(y/x) = P(Y < y/X = x)$$

3. Xác định biểu thức của hàm mô tả mối phụ thuộc nói trên giữa các biến ngẫu nhiên.

Nghiên cứu vấn đề 1,2 là nội dung của bài toán phân tích tương quan.

Nghiên cứu vấn đề 3 là nội dung của bài toán phân tích hồi quy.

§2. PHÂN TÍCH TƯƠNG QUAN

2.1. Phân tích tương quan bảng số liệu định lượng

Chúng ta tiến hành phân tích tương quan chủ yếu dựa trên cơ sở phân tích hai đặc trưng cơ bản là hệ số tương quan (trường hợp hai biến ngẫu nhiên) và hệ số tương quan riêng phần (trường hợp nhiều hơn hai biến ngẫu nhiên).

1. Phân tích hệ số tương quan

Giả sử X và Y là hai biến ngẫu nhiên có Var(X) > 0 và Var(Y) > 0. Hệ số tương quan của hai biến ngẫu nhiên là X và Y, ký hiệu là ρ_{XY} và được xác định như sau (xem Chương IV, §6).

$$\rho_{XY} = \frac{E(X - EX)(Y - EY)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Như chúng ta đã biết có thể dùng ρ_{XY} để đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên. $|\rho|$ càng lớn thì sự phụ thuộc tuyến tính càng rõ. Trường hợp đặc biệt $\rho_{XY} = 0$ thì ta nói giữa X và Y không có sự tương quan với nhau. Nếu X và Y độc lập với nhau thì $\rho_{XY} = 0$, điều ngược lại chưa chắc đã đúng. Tuy nhiên đối với biến ngẫu nhiên có phân phối chuẩn thì sự độc lập và không tương quan là tương đương nhau. Trong trường hợp tổng quát cần phân biệt hai khái niệm độc lập và không tương quan giữa hai biến ngẫu nhiên.

Định nghĩa. Giả sử ta có $\{(X_i, Y_i)\}\, i = \overline{1,n}$ là một mẫu ngẫu nhiên hai chiều thu được khi quan sát vectơ ngẫu nhiên (X, Y) thì hệ số tương quan mẫu r_{XY} của X và Y được xác định như sau:

$$r_{XY} = \frac{\dfrac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{n}}{\sqrt{\dfrac{\sum_i (X_i - \overline{X})^2}{n}}\sqrt{\dfrac{\sum_i (Y_i - \overline{Y})^2}{n}}}$$

$$= \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{\overline{X}^2 - (\overline{X})^2}\sqrt{\overline{Y}^2 - (\overline{Y})^2}} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{MS_X}\sqrt{MS_Y}} \qquad (10.1)$$

Trong đó:

$$\overline{X} = \frac{\left(\sum_i X_i\right)}{n}; \quad \overline{X}^2 = \frac{\left(\sum_i X_i^2\right)}{n}$$

$$\overline{Y} = \frac{\left(\sum_i Y_i\right)}{n}; \quad \overline{Y}^2 = \frac{\left(\sum_i Y_i^2\right)}{n}$$

$$\overline{XY} = \frac{\left(\sum_i X_i Y_i\right)}{n}$$

Ví dụ 1. Chúng ta xét tập hợp số liệu có tên là XSTK10_1 về một số chỉ tiêu phát triển kinh tế Việt Nam trong thời kỳ 1980 - 1996 (đơn vị tính 1000 tỷ đồng VN năm 1989). Nội dung của tập số liệu đó như sau:

Bảng 10.1

```
. d
Contains data from C:\WINSTATA/XSTK10_1.DATA
  obs:                17
 vars:                 6                          11 Aug 1998 19:02
 size:               476 (85.9% of memory free)
-------------------------------------------------------------------------------
1. nam      float   %9.0g
2. gdp      float   %9.2g  Tong gia tri san pham trong nuoc
3. ex       float   %9.2g  Tong gia tri hang hoa xuat khau
4. im       float   %9.2g  Tong gia tri hang hoa nhap khau
5. gip      float   %9.2g  Tong gia tri h. hoa cong nghiep
6. gap      float   %9.0g  Tong gia tri h. hoa nong nghiep
-------------------------------------------------------------------------------
Sorted by:
. list
```

	nam	gdp	ex	im	gip	gap
1.	1980	16.80	.30	6.80	6.80	9.7
2.	1981	17.20	.40	6.90	6.90	10.1
3.	1982	18.70	.50	7.50	7.50	11.2
4.	1983	20.10	.60	8.50	8.50	11.5
5.	1984	21.80	.60	9.60	9.60	12.2
6.	1985	23.00	.70	10.50	10.50	12.5
7	1986	23.80	.80	11.20	11.20	13.1
8.	1987	24.70	.90	12.30	12.30	13.1
9.	1988	25.90	1.00	14.00	14.00	13.7
10.	1989	28.00	1.90	13.60	13.60	14.7

11.	1990	29.50	2.40	14.00	14.00	14.9
12.	1991	31.30	2.10	15.50	15.50	15.4
13.	1992	34.00	2.60	18.10	18.10	16.6
14.	1993	36.70	3.00	20.40	20.40	17.6
15.	1994	40.00	4.10	23.20	23.20	18.6
16.	1995	43.80	5.37	26.50	26.50	19.4
17.	1996	47.90	7.10	30.23	30.23	20.35

Kí hiệu Y = (gdp) và X = (gap), để tính hệ số tương quan (mẫu) r_{XY} ta lập bảng tính toán sau đây:

STT	y_i	x_i	y_i^2	x_i^2	x_iy_i
1	16.8	9.70	282.24	94.0900	162.9600
2	17.2	10.10	295.84	102.0100	173.7200
3	18.7	11.20	349.69	125.4400	209.4400
4	20.1	11.50	404.01	132.2500	231.1500
5	21.8	12.20	475.24	148.8400	265.9600
6	23.0	12.50	529.00	156.2500	287.5000
7	23.8	13.10	566.44	171.6100	311.7800
8	24.7	13.10	610.09	176.6100	323.5700
9	25.9	13.70	670.81	187.6900	354.8300
10	28.0	14.70	784.00	216.0900	411.6000
11	29.5	14.90	870.25	222.0100	439.5500
12	31.3	15.40	979.69	237.1600	482.0200
13	34.0	16.60	1156.00	275.5600	564.4000
14	36.7	17.60	1346.89	309.7600	645.9200
15	40.0	18.60	1600.00	345.9600	744.0000
16	43.8	19.40	1918.44	376.3600	849.7200
17	47.9	20.35	2294.41	4144.1225	974.7650
Σ	483.2	244.65	15133.04	3686.8125	7432.8850

Từ kết quả trên thu được:

$$\overline{x} = 14,3912,\ \overline{x}^2 = 216,8713$$

$$\overline{y} = 28,4235,\ \overline{y}^2 = 890,1788,\ \overline{xy} = 437,2285$$

Áp dụng công thức (10.1) ta được:

$$r_{XY} = \frac{437,2285 - 14,3912.28,4235}{\sqrt{216,8713 - 14,3912^2}\sqrt{890,1788 - 28,4235^2}}$$

$$= \frac{28,180}{3,1248.9,07} = 0,9942$$

Trường hợp nếu chúng ta cần xét p biến ngẫu nhiên X_1, X_2, ..., X_p ($p > 2$) ta ký hiệu $r_{x_i y_j} = r_{ij}$, khi đó ta có tập hợp $\{r_{ij}\}$; $i, j = \overline{1,p}$ là ma trận hệ số tương quan của p biến ngẫu nhiên nói trên. Dễ dàng thấy rằng $\{r_{ij}\}$ là ma trận đối xứng và các phần tử trên đường chéo là $r_{ii} = 1$. Với số liệu đã cho ở bảng 10.1 Stata cho ta ma trận hệ số tương quan của các biến gdp, ex, im, gip, gap như sau:

Bảng 10.2

• corr gdp ex im gip gap

(obs = 17)

	gd	ex	im	gip	gap
gdp	1.0000				
ex	0.9653	1.0000			
im	0.8926	0.9605	1.0000		
gip	0.9942	0.9720	0.9244	1:0000	
gap	0.9942	0.9375	0.8508	0.9815	1.0000

Qua bảng hệ số tương quan trên ta thấy các chỉ tiêu (biến ngẫu nhiên) nói trên có sự tương quan với nhau rất mạnh, hầu hết hệ số tương quan giữa các cặp đều lớn hơn 0,9. Điều đó chứng tỏ rất có khả năng giữa chúng tồn tại mối quan hệ phụ thuộc tuyến tính.

a. Khoảng tin cậy cho hệ số tương quan ρ_{XY}

Fisher đã chứng minh rằng ngay với những giá trị n (kích thước mẫu) không lớn lắm thống kê Z sau đây:

$$Z = \frac{1}{2}\ln\frac{1+r_{XY}}{1-r_{XY}}$$

cũng có phân phối xấp xỉ chuẩn với

$$E(Z) = \frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}} + \frac{\rho_{XY}}{2(n-1)} \text{ và } Var\,(Z) = \frac{1}{n-3}$$

Từ đây ta có khoảng tin cậy $(1-\alpha)$ của $\frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}}$ là:

$$\left(Z - \frac{r_{XY}}{2(n-1)} - u_{\alpha/2}\frac{1}{\sqrt{n-3}}; Z - \frac{r_{XY}}{2(n-1)} + u_{\alpha/2}\frac{1}{\sqrt{n-3}}\right)$$

Tra bảng các giá trị của hàm $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ ta sẽ suy ra khoảng tin cậy cho hệ số tương quan ρ_{XY}.

b. Kiểm định giả thuyết về giá trị của ρ

Xét cặp giả thuyết:

$$H_0 : (\rho_{XY} = \rho_0)$$

$$H_1 : (\rho_{XY} \neq \rho_0)$$

Sử dụng thống kê Z nêu trên ta thấy nếu H_0 đúng thì:

$$U = \left(Z - \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)} \right) \sqrt{n-3}$$

sẽ có phân phối xấp xỉ chuẩn N(0,1). Do đó miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ U = \left(Z - \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1} \right) \sqrt{n-3}; |U| > u_{\alpha/2} \right\}$$

Trường hợp đặc biệt ta giả thiết X và Y đều tuân theo quy luật phân phối chuẩn. Xét cặp giả thuyết

$$H_0: (\rho_{XY} = 0)$$

$$H_1: (\rho_{XY} \neq 0)$$

Với kích thước mẫu bất kỳ ta luôn có:

$$T = r_{XY} \sqrt{\frac{n-2}{1-r_{XY}^2}}$$

tuân theo quy luật phân phối Student với bậc tự do (n – 2).

Vậy miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ t = r_{XY} \sqrt{\frac{n-2}{1-r_{XY}^2}}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

2. Phân tích hệ số tương quan riêng phần

Xét tập hợp biến ngẫu nhiên $X_1, X_2, ..., X_p$. Hệ số tương quan riêng phần của X_1 và X_2 khi X_3 cố định kí hiệu là $r_{12,3}$. Khi đó ta có công thức xác định $r_{12,3}$ như sau:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Tổng quát ta có hệ số tương quan riêng phần của X_i và X_j khi cố định X_k là $r_{ij,\,k}$:

$$r_{ij,k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1-r_{ik}^2)(1-r_{jk}^2)}} \qquad (10.2)$$

Hệ số tương quan riêng phần giữa X_1 và X_2 khi cố định X_3, X_4 là:

$$r_{12,34} = \frac{r_{12,3} - r_{14,3}r_{24,3}}{\sqrt{(1-r_{14,3}^2)(1-r_{24,3}^2)}}$$

Tương tự ta sẽ có công thức xác định hệ số tương quan riêng phần giữa X_1 với X_2 khi cố định X_3, X_4, ... X_p là:

$$r_{12,34...p} = \frac{r_{12,34...(p-1)} - r_{1p,34...(p-1)}r_{2p,34...(p-1)}}{\sqrt{(1-r_{1p,34...(p-1)}^2)(1-r_{2p,34...(p-1)}^2)}}$$

Trên thực tế khi cần nghiên cứu một tập biến ngẫu nhiên X_1, X_2, ... X_p thì chúng ta thường quan tâm đến hệ số tương quan riêng phần của một biến nào đó (thông thường là biến mà chúng ta đang muốn quan tâm phân tích, giải thích sự biến động của nó) với tất cả các biến còn lại. Ví dụ đó là biến X_1 chẳng hạn thì để đơn giản chúng ta chỉ cần viết:

r_{12}, ... (Hệ số tương quan riêng phần của X_1và X_2 khi X_3, X_4 ..., X_p cố định).

...

r_{1p}, ... (Hệ số tương quan riêng phần của X_1 và X_p khi X_2, X_3, ... X_{p-1} cố định).

Ví dụ 2: Căn cứ vào ma trận hệ số tương quan (bảng 10.3) chúng ta sẽ tính hệ số tương quan riêng phần của X_1 = (gdp) với X_2 = (ex) và X_3 = (im). Theo công thức (10.2) ta có:

$$r_{12,3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = \frac{0,9653 - 0,8926.0,9605}{\sqrt{(1-0,8926^2)(1-0,9605^2)}} =$$

$$= \frac{0,1079577}{0,1254624} = 0,8605$$

$$r_{13,2} = \frac{r_{13} - r_{12}.r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}} = \frac{0,8926 - 0,9653.0,9805}{\sqrt{(1-0,9653^2)(1-0,9605^2)}} =$$

$$= \frac{-0,03457}{0,07267} = -0,4757$$

Với Stata chúng ta có thể dễ dàng tính được các hệ số tương quan riêng phần bằng lệnh pcorr sau đây:

pcorr x_1 x_2 x_3 ... x_p

Ở đây có nghĩa là tính hệ số tương quan riêng phần của x_1 với lần lượt các biến x_2, x_3 ... x_p khi cố định các biến còn lại. Theo như ký hiệu ở trên thì ta sẽ có tất cả các r_{1i}, ... với $i = \overline{2,p}$.

Sau đây là kết quả chạy bằng Stata cho tệp số liệu XSTK10_1.

• pcorr gdp ex im

(obs = 17)

Partial correlation of gdp with

Variable	Corr.	Sig.
ex	0.8610	0.000
im	–0.4771	0.062

• pcorr gdp ex im gip gap

(obs = 17)

Partial corelation of gdp with

Variable	Corr.	Sig.
ex	0.7958	0.001
im	–0.5830	0.029
gip	0.8024	0.001
gap	0.8514	0.000

Theo kết quả nêu trên ta thấy nếu cố định một chỉ tiêu ví dụ (im) chẳng hạn thì hệ số tương quan riêng phần của gdp với ex là 0,8610, còn khi cố định (ex) thì hệ số tương quan riêng phần của gdp với im lại là một số âm: –0,4771 (ở đây có một chút sai số so với tính tay vì làm tròn số).

Ta cũng thấy hệ số tương quan riêng phần của gdp với ex khi cố định tất cả các chỉ tiêu là 0,7958. Dễ dàng nhận thấy rằng nếu cố định các chỉ tiêu còn lại thì hệ số tương quan riêng phần của gdp với gap là cao nhất và cũng rất lớn: 0,8514. Kết hợp với phương pháp hồi quy, các bạn sẽ thấy chúng ta có thể rút ra được nhiều nhận xét, kết luận rất thú vị.

2.2. Phân tích tương quan bảng số liệu định tính

1. Bảng ngẫu nhiên hai chiều

Bảng này chủ yếu dùng cho trường hợp các chỉ tiêu định tính được "đo" bằng thang đo phân loại. Như các bạn đã biết trong chương VII chúng ta đã xét bài toán kiểm định sự độc lập của hai dấu hiệu (chỉ tiêu định tính) A và B. Tuy nhiên

nếu A và B không độc lập với nhau thì tiêu chuẩn χ^2 không cho chúng ta biết mức độ tương quan giữa A và B.

Với những bảng ngẫu nhiên hai chiều cấp (2 × 2) dạng:

A \ B	b_1	b_2	Σ
a_1	n_1	n_2	$(n_1 + n_2)$
a_2	n_3	n_4	$(n_3 + n_4)$
Σ	$(n_1 + n_3)$	$(n_2 + n_4)$	$n_1 + n_2 + n_3 + n_4 = n$

Chúng ta có thể đo mối tương quan giữa A và B bằng các hệ số Q và F được xác định như sau:

Hệ số Q:
$$Q = \frac{n_1 n_4 - n_2 n_3}{n_1 n_4 + n_2 n_3} \qquad (10.3)$$

Hệ số F:
$$F = \frac{n_1 n_4 - n_2 n_3}{\sqrt{(n_1 + n_2)(n_1 + n_3)(n_2 + n_4)(n_3 + n_4)}}$$

Ví dụ 3. Trên cơ sở bảng số liệu điều tra về thái độ của sinh viên đối với việc chấp hành nội quy thi (tệp số liệu XSTK10_2) chúng ta có bảng ngẫu nhiên hai chiều sau đây:

Bảng 10.3a

• Tab thaido gioitinh

Chap hanh noi quy thi	Sinh vien nam hay nu		Total
	0	1	
0	40	40	80
1	20	0	20
Total	60	40	100

Bảng 10.3b

• Tab thaido vung

Chap hanh noi quy thi	Nong thon hay thanh thi		Total
	nt	tt	
0	43	37	80
1	8	12	20
Total	51	49	100

Trong đó A: (thaido) có hai giá trị $a_1 = 0$ là không vi phạm, $a_2 = 1$ là có vi phạm, B = (gioitinh) có hai giá trị $b_1 = 0$ là nữ, $b_2 = 1$ là nam và C = (vung) có hai loại: nt là nông thôn, tt là thành thị. Như vậy là riêng đối với chỉ tiêu C chúng ta không "đo" bằng thang đo phân loại mà vẫn giữ nguyên các ký hiệu phân loại.

Phân tích bảng 10.3a ta có:

$$Q = \frac{40.0 - 40.20}{40.0 + 40.20} = -1$$

$$F = \frac{40.0 - 40.20}{80.60.20.40} = -0{,}4$$

Nhận xét: Ta nhận thấy giá trị của Q và F được tính trên cùng một bảng số liệu lại rất khác nhau: Giá trị Q lớn hơn hai lần giá trị của F (về giá trị tuyệt đối). Do đâu mà như vậy? Trước hết ta thấy rằng các hệ số Q và F đo các mặt khác nhau của mối liên hệ trong bảng bốn ô (bảng ngẫu nhiên 2 chiều cấp (2 × 2)). Trong ví dụ này hệ số |Q| = +1 vì tất cả các trường hợp vi phạm nội quy thi trong suốt quá trình khảo sát (điều tra) đều là của các sinh viên nam. Song sẽ hoàn toàn

không đúng nếu giải thích Q có nghĩa là hầu như tất cả sinh viên nam đều vi phạm nội quy thi.

Hệ số F cho phép phản ánh mức độ liên hệ qua lại hai chiều giữa các dấu hiệu nghiên cứu, trong lúc đó hệ số Q chỉ phản ánh mối liên hệ một chiều. Giá trị của các hệ số chỉ trùng nhau khi có mối liên hệ qua lại hai chiều hoàn toàn.

Nếu $n_1 = n_4 = 0$ thì $F = -1$. Nếu $n_2 = n_3 = 0$ thì $F = +1$. Chúng ta có thể dễ dàng kiểm tra ý nghĩa của F thông qua tiêu chuẩn χ^2 vì $\chi^2 = nF^2$.

Đối với bảng ngẫu nhiên hai chiều cấp (p × q). Giả sử chúng ta đã tính được giá trị của tiêu chuẩn χ^2 (tức là giá trị χ^2_{qs}) và đã kết luận được hai dấu hiệu A và B không độc lập với nhau. Khi đó, chúng ta có thể đo mối liên hệ (mức độ tương quan) giữa hai dấu hiệu đó bằng hệ số liên hợp Pearson (P):

$$P = \sqrt{\frac{\chi^2}{n + \chi^2}} \quad 0 \le P \le 1 \qquad (10.4)$$

Ví dụ 4. Nghiên cứu tình trạng hôn nhân trước ngày cưới của 542 cặp vợ chồng ta có bảng số liệu sau:

Tình trạng hôn nhân của vợ / Tình trạng hôn nhân của chồng	Chưa kết hôn lần nào	Ly hôn	Góa	Σ
Chưa kết hôn lần nào	180	34	36	250
Ly hôn	58	76	54	188
Góa	43	34	27	104
Σ	281	144	117	542

Ta tính được $\chi^2_{qs} = 11,96$; $\chi^2_{0,05}(4) = 9,5 \rightarrow$ Bác bỏ tính độc lập giữa tình trạng hôn nhân của vợ (A) và tình trạng hôn nhân của chồng (B) trước ngày cưới.

Vậy: $$P = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{11,96}{542+11,96}} = 0,147$$

$P = 0$ khi $\chi^2 = 0$, tức là khi có sự độc lập hoàn toàn giữa các dấu hiệu. Nhược điểm của hệ số P là giá trị cực đại của nó phụ thuộc vào quy mô của bảng ngẫu nhiên hai chiều. Do đó nẩy sinh những khó khăn nhất định cho việc giải thích và về thực chất nên coi P như một chỉ số bằng số chứ không phải như một độ đo. Để khắc phục nhược điểm trên ta có hệ số Kramer (K):

$$K = \left\{\frac{\chi^2}{n.\min(p-1, q-1)}\right\}^{1/2}$$

Hệ số K bao giờ cũng có thể đạt +1 không phụ thuộc vào dạng của bảng ngẫu nhiên hai chiều.

Theo ví dụ 10.4 ta có:

$$K = \left\{\frac{11,96}{542.\min(2,2)}\right\}^{1/2} = 0,105$$

Như vậy, ta có thể nói rằng tình trạng hôn nhân của vợ và chồng trước khi cưới có phụ thuộc vào nhau nhưng mức độ phụ thuộc đó không lớn lắm. Có lẽ đó chỉ là yếu tố mang tính chất tham khảo chứ không mang tính chất quyết định.

2. Các độ đo sự tương quan ở cấp đo theo thang thứ bậc

Trước hết, ta xét một ví dụ. Cho tệp số liệu tên gọi là XSTK10_3 với nội dung như sau:

Bảng 10.4

Contains data from: C\WINSTATA\XSTK10_3.dta

obs: 10

vars: 2 17 Aug 1998 15: 26

size: 60 (86.0% of memory free)

1. diem 1 byte %8 . 0g Diem cua giam khao nguoi Canada

2. diem 2 byte %8 . 0g Diem cua giam khao nguoi Nhat

Sorted by:

diem 1	95	90	86	84	75	70	62	60	57	50
diem 2	92	93	83	80	55	60	45	72	62	70

Trong đó A = (diem1), B = (diem2) là các điểm do hai giám khảo người Canada và Nhật chấm cho một vận động viên trong một cuộc thi quốc tế về trượt băng nghệ thuật.

Các giá trị A và B được đo theo thang đo thứ bậc. Bây giờ ta sẽ xếp hạng cho các giá trị của A và B. Trước hết, ta xếp các giá trị của A theo thứ tự (tăng hoặc giảm dần, sau đó ta gán các hạng cho các giá trị đó đúng bằng số thứ tự đã sắp xếp.

A = (diem 1)	95	90	86	84	75	70	62	60	57	50
Hạng (i)	1	2	3	4	5	6	7	8	9	10

Làm tương tự như đối với A ta có các hạng của B như sau:

B = (diem 2)	93	92	93	80	72	70	62	60	55	45
Hạng (k)	1	2	3	4	5	6	7	8	9	10

Tiếp theo ta sắp xếp các hạng của A và B theo đúng thứ tự các giá trị của A và B đã cho ở bảng 10.4.

Hạng A: i	1	2	3	4	5	6	7	8	9	10
Hạng B: k_i	2	1	3	4	9	8	10	5	7	6

Như vậy là tương ứng với mỗi cặp giá trị (a_i, b_i) của (A, B) ta có một cặp hạng tương ứng là (i, k_i).

+ Hệ số tương quan hạng r_S của Spearman cho hai dấu hiệu A và B được đo cùng thang đo thứ bậc xác định như sau:

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \qquad (10.5)$$

Trong đó $d_i = i - k_i$ là hiệu giữa hai giá trị hạng của cặp hạng (i, k_i) $(i = \overline{1,n})$, n là số cặp hạng.

Ký hiệu ρ_s là hệ số tương quan hạng Spearman của tổng thể thì để kiểm định cặp giả thuyết:

$$H_0: (\rho_s = 0),\ H_1: (\rho_s \neq 0)$$

ta sử dụng tiêu chuẩn

$$T = \sqrt{\frac{r_s^2(n-2)}{1-r_s^2}}$$

Nếu H_0 đúng thì T có phân phối Student với (n–2) bậc tự do. Do đó, ta có miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ t = \sqrt{\frac{r_s^2(n-2)}{1-r_s^2}}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

Ví dụ 5. Sử dụng số liệu đã cho ở bảng 10.4 ta tính hệ số tương quan hạng giữa hai dãy điểm số đã cho của hai giám khảo người Canađa và Nhật.

Theo cách xếp hạng trên ta được:

$$\sum_i d_i^2 = 1 + 1 + 16 + 4 + 9 + 9 + 4 + 16 = 60$$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6.60}{10(10^2-1)} = 0{,}6364$$

Để kiểm định xem cách cho điểm của hai giám khảo nói trên có độc lập với nhau hay không ta tính:

$$t_{qs} = \sqrt{\frac{r_s^2(r-2)}{1-r_s^2}} = \sqrt{\frac{0{,}6364^2.8}{1-0{,}6364^2}} = 2{,}356$$

Với $\alpha = 0{,}05$ tra bảng ta được $t_{0,025}^{(8)} = 2{,}306$, $t_{qs} > t_{0,025}^{(8)}$. Ta có thể bác bỏ giả thuyết H_0: ($\rho_s = 0$). Tuy nhiên, ta thấy giá trị quan sát t_{qs} chênh lệch so với giá trị tới hạn quá ít và nếu chọn mức ý nghĩa là 0,01 thì giá trị tới hạn $t_{\alpha/2}^{(8)} = t_{0,005}^{(8)} = 3{,}335$ thì $t_{qs} \notin W_\alpha$, tức là ta không có cơ sở bác bỏ giả thuyết H_0. Trong những trường hợp như vậy ta có thể phân tích thêm bằng hệ số tương quan hạng Kendall.

+ Hệ số tương quan hạng Kendall ký hiệu là τ và được xác định bằng công thức sau:

$$\tau = \frac{S}{\frac{n(n-1)}{2}} \qquad (10.6)$$

Việc tính giá trị của S được minh họa qua ví dụ cụ thể sau đây.

Trên cơ sở số liệu đã xếp hạng về điểm số của hai giám

khảo người Canađa và Nhật trong ví dụ 10.5 ta lập bảng tính S như sau:

Hạng của A	Hạng của B	S_i^+	S_i^-	$S_i^+ - S_i^-$
1	2	8	1	7
2	1	8	0	8
3	3	7	0	7
4	4	6	0	6
5	9	1	4	–3
6	8	1	3	–2
7	10	0	3	–3
8	5	2	0	2
9	7	0	1	–1
10	6	-	-	-
				$S = \sum(S_i^+ - S_i^-) = 21$

Cách tính S_i^+ và S_i^- như sau. Trước hết, lấy giá trị đầu tiên của cột hạng của B (tức là k_1) và đếm số hạng có giá trị lớn hơn k_1 đó chính là S_1^+, số hạng còn lại có giá trị nhỏ hơn k_1 là S_1^-. Tiếp tục như vậy cho đến giá trị k_{n-1} thì kết thúc. Tổng quát ta có thể mô tả cách tính S_i^+ và S_i^- như sau:

S_i^+ = số hạng k_j (j > i) thỏa mãn: $k_j > k_i$

S_i^- = số hạng k_j (j > i) thỏa mãn $k_j < k_i$

Cuối cùng là $S = \sum_i (S_i^+ - S_i^-)$. (S còn được gọi là số điểm Kendall – Kendall's score)

Ký hiệu ρ_k là hệ số tương quan hạng Kendall của tổng thể thì để kiểm định cặp giả thuyết:

$$H_0 : (\rho_K = 0) \; ; \; H_1 : (\rho_K \neq 0)$$

ta sử dụng miền bác bỏ sau đây:

$$W_\alpha = \left\{ U = \frac{\tau}{\sqrt{\frac{4n+10}{9n(n-1)}}} ; |U| > u_{\alpha/2} \right\}$$

Với số liệu đã tính ở trên ta có:

$$\tau = \frac{S}{n(n-1)/2} = \frac{21}{10(10-1)/2} = 0{,}467$$

và $U_{qs} = \dfrac{0{,}467}{\sqrt{\frac{40+10}{90.9}}} = 1{,}88$

Với $\alpha = 0{,}05$ ta có $u_{0{,}025} = 1{,}96 \rightarrow U_{qs} \notin W_\alpha$. Vậy không có cơ sở bác bỏ giả thuyết H_0: $(\rho_k = 0)$. Có nghĩa là có thể coi điểm của hai giám khảo người Canada và Nhật là độc lập với nhau với mức ý nghĩa 5%.

Sau đây là kết quả chạy bằng Stata đối với ví dụ 5.

```
• spearman  diem1   diem2

Number of obs =        10

Spearman's rho =    0.6364

Test of Ho: diem1 and diem2 independent

Pr > |t| =          0.0479
```

• Ktau diem1 diem2

Number of obs = 10

Kendall's tau-a = 0.4667

Kendall's tau-b = 0.4667

Kendall's score = 21

SE of score = 11.180

Test of Ho: diem1 and diem2 independent

Pr > |z| = 0.0736 (continuity corrected)

§3. PHÂN TÍCH HỒI QUY

Trên đây chúng ta đã trình bày cách phân tích mối tương quan giữa p biến ngẫu nhiên: X_1, X_2, ..., X_p. Tuy nhiên trên thực tế, nhiều khi chúng ta muốn đi sâu nghiên cứu rõ hơn sự tác động của một, hay nhiều biến trong số đó đến sự biến động của một biến nào đó, ví dụ như biến ngẫu nhiên X_1 chẳng hạn. Có nghĩa là chúng ta muốn giải thích được nhiều hơn và cụ thể hơn sự biến động của X_1 thông qua sự tác động của những biến đó. Đó chính là mục đích phương pháp phân tích hồi quy. Để tiện cho việc trình bày, chúng ta vẫn xét một tập hợp p biến ngẫu nhiên, nhưng trong đó X_1 được thay thế bằng Y. Tức là ta có: Y, X_2, X_3, ..., X_p.

3.1. Hàm hồi quy

Giả sử ta muốn tìm một hàm f nào đó của $X_2, X_3, ..., X_p$ sao cho nó xấp xỉ Y tốt nhất theo nghĩa cực tiểu sai số bình phương trung bình. Có nghĩa là:

$$E(Y - f(X_2, X_3, ..., X_p))^2 = E(Y - f(X))^2$$
$$= \min_{f(x)} E(Y - f(X))^2$$

Trong đó ký hiệu $X = (X_2, X_3, ..., X_p)$.

Người ta đã chứng minh được rằng $E(Y - f(X))^2$ sẽ đạt cực tiểu khi hàm f(X) là kỳ vọng có điều kiện của Y đối với $X_2, X_3, ..., X_p$. Tức là khi:

$$f(X) = E(Y/X_2, X_3, ..., X_p)$$

Định nghĩa: Hàm hồi quy (hay hàm hồi quy kỳ vọng của Y đối với một vectơ $X = (X_2, X_3, ..., X_p)$ là kỳ vọng có điều kiện của Y đối với X.

$$f_Y(X) = E(Y/ X_2, X_3, ..., X_p) = E(Y/X)$$

Hàm hồi quy có thể có nhiều dạng khác nhau. Ví dụ với p = 2 và ta có hai biến ngẫu nhiên là Y, X thì hàm hồi quy có dạng :

$$f_Y(X) = aX + b$$

được gọi là hàm *hồi quy tuyến tính đơn*. Trường hợp tổng quát (p > 2) ta có :

$$f_Y(X_2, X_3, ..., X_p) = \beta_1 + \beta_2X_2 + \beta_3X_3 + ...+ \beta_pX_p$$

và nó được gọi là *hàm hồi quy tuyến tính bội*. Ngoài ra, hàm hồi quy còn có các dạng khác như hàm đa thức bậc 2, bậc 3, ... hay dạng hàm logarít... Các hằng số $\beta_1, \beta_2, ..., \beta_p$ được gọi là *tham số* của hàm hồi quy.

Nội dung của bài toán phân tích hồi quy gồm một số điểm cơ bản như sau:

Dựa trên số liệu điều tra thống kê (mẫu ngẫu nhiên) thu được khi quan sát ($Y, X_2, X_3, ..., X_p$) hãy:

1. Ước lượng các tham số của hàm hồi quy

2. Kiểm định giả thuyết về giá trị của các tham số đó.

3. Đánh giá các sai số ước lượng và kiểm tra tính phù hợp (đúng đắn) của hàm hồi quy.

4. Dự báo (hay dự đoán) các giá trị của Y theo $X_2, X_3, ..., X_p$.

3.2. Mô hình hồi quy tuyến tính đơn

Trong phần này chúng ta xét trường hợp $p = 2$, tức là chúng ta xét hàm hồi quy có dạng sau đây:

$$f_Y(X) = aX + b$$

$f_Y(X)$ còn được gọi là hàm hồi quy lý thuyết. Giả sử ta có một mẫu ngẫu nhiên kích thước n thu được khi quan sát (Y, X) là:

$$\{(Y_i, X_i)\} = \{(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)\}$$

Khi đó chúng ta có thể viết: $f_{Yi}(X_i) = E(Y_i/X_i) = aX_i + b$

Hay:

$$Y_i = aX_i + b + U_i \quad i = \overline{1,n} \tag{10.7}$$

Trong đó U_i là các sai số ngẫu nhiên và giả thiết rằng chúng độc lập với nhau, cùng tuân theo quy luật phân phối chuẩn $N(0, \sigma^2)$.

(10.7) được gọi là mô hình hồi quy tuyến tính đơn.

1. Ước lượng các hệ số a, b

Ký hiệu $\hat{a}$, $\hat{b}$ là **các ước lượng của a và b thì** $\hat{Y} = \hat{a}X + \hat{b}$ được gọi là hàm hồi quy ước lượng (của hàm hồi quy lý thuyết $f_Y(X) = aX + b$.

Xét mô hình (10.7) và ta đi tìm $\hat{a}$, $\hat{b}$ bằng phương pháp bình phương bé nhất tức là tìm $\hat{a}$, $\hat{b}$ sao cho cực tiểu hóa được tổng các bình phương sai số (e_i^2):

$$S(\hat{a}, \hat{b}) = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \hat{a}X_i - \hat{b})^2 = \sum_i e_i^2 =$$

$$= \min_{a,b} \sum (Y_i - aX_i - b)^2$$

Dễ dàng nhận thấy rằng $\hat{a}$, $\hat{b}$ là nghiệm của hệ phương trình sau:

$$\begin{cases} \dfrac{\partial S(a,b)}{\partial a} = \dfrac{\partial \sum (Y_i - aX_i - b)^2}{\partial a} = 0 \\ \dfrac{\partial S(a,b)}{\partial b} = \dfrac{\partial \sum (Y_i - aX_i - b)^2}{\partial b} = 0 \end{cases}$$

hay

$$\begin{cases} nb + a\sum X_i = \sum Y_i \\ b\sum X_i + a\sum X_i^2 = \sum X_i Y_i \end{cases} \qquad (10.8)$$

Giải hệ phương trình (10.8) ta được:

$$\hat{a} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2} = r_{XY} \frac{\sqrt{MS_Y}}{\sqrt{MS_X}}$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X} \qquad (10.9)$$

Trong đó:

$$MS_X = \frac{\sum (X_i - \bar{X})^2}{n} = \overline{X}^2 - (\bar{X})^2$$

$$MS_Y = \frac{\sum (Y_i - \bar{Y})^2}{n} = \overline{Y}^2 - (\bar{Y})^2$$

$$\overline{XY} = (\sum_i (X_i Y_i)/n$$

$$\overline{X}^2 = (\sum_i X_i^2)/n; \quad \overline{Y}^2 = (\sum_i Y_i^2)/n$$

Người ta cũng chứng minh được rằng $\hat{a}$, $\hat{b}$ là các ước lượng không chệch tốt nhất của a, b và:

$$Var(\hat{a}) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma^2/n}{\overline{X}^2 - (\bar{X})^2} = \frac{\sigma^2/n}{MS_X}$$

$$Var(\hat{b}) = \frac{\sigma^2 \overline{X}^2}{\sum (X_i - \bar{X})^2} = \frac{(\sigma^2/n)\overline{X}^2}{\overline{X}^2 - (\bar{X})^2} = \frac{(\sigma^2/n)\overline{X}^2}{MS_X}$$

Nếu σ^2 chưa biết ta dùng ước lượng không chệch của nó là:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

Với các giả thuyết đã nêu ra đối với mô hình (10.7) ta cũng có:

+ $\hat{a} \sim N(a, Var(\hat{a}))$ hay:

$$U = \frac{\hat{a} - a}{\sqrt{Var(\hat{a})}} \sim N(0,1)$$

+ $\hat{b} \sim N(b, Var(\hat{b}))$ hay:

$$U = \frac{\hat{b} - b}{\sqrt{Var(\hat{b})}} \sim N(0,1)$$

Tường hợp σ^2 chưa biết, thay thế nó bằng $\hat{\sigma}^2$, ta suy ra rằng:

$$T = \frac{\hat{a} - a}{Se(\hat{a})} \sim T(n-2)$$

(Phân phối Student với (n – 2) bậc tự do)

$$T = \frac{\hat{b} - b}{Se(\hat{b})} \sim T(n-2)$$

Trong đó

$$Se(\hat{a}) = \sqrt{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{\hat{\sigma}^2 / n}{\overline{X}^2 - (\bar{X})^2}} \qquad (10.10)$$

$$Se(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2 \overline{X}^2}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{(\hat{\sigma}^2 / n)\overline{X}^2}{\overline{X}^2 - (\bar{X})^2}}$$

Từ đây ta có khoảng tin cậy (1 – α) cho các tham số a, b là:

$$\left(\hat{a} - Se(\hat{a})t_{\alpha/2}^{(n-2)} < a < \hat{a} + Se(\hat{a}).t_{\alpha/2}^{(n-2)}\right) \qquad (10.11)$$

và $$\left(\hat{b} - Se(\hat{b})t_{\alpha/2}^{(n-2)} < b < \hat{b} + Se(\hat{b}).t_{\alpha/2}^{(n-2)}\right)$$

Ví dụ 6. Trở lại với tệp số liệu XSTK10_1 (Xem bảng 10.1).

Ta gọi chỉ tiêu gdp là Y và gap là X, dựa vào kết quả đã tính toán được ở ví dụ 1 ta có:

$$\hat{a} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X}^2 - (\overline{X})^2} = \frac{437{,}23 - 14{,}391.28{,}424}{216{,}87 - 14{,}391^2} = 2{,}885$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X} = 28{,}424 - 2{,}885.14{,}391 = -13{,}095$$

vậy: $\hat{Y} = \hat{a}X + \hat{b} = 2{,}885X - 13{,}095$

Để tính được $\hat{\sigma}^2$ mà không cần phải tính các giá trị $e_i = Y_i - \hat{Y}_i$ ta xét biểu thức sau đây:

$$TSS = \sum_i (Y_i - \overline{Y})^2 = \sum_i \left[(Y_i - \hat{Y}_i) + \hat{Y}_i - \overline{Y})\right]^2$$

$$= \sum_i (Y_i - \overline{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 = ESS + RSS \qquad (10.12)$$

Trong đó $ESS = \sum_i (\hat{Y}_i - \overline{Y})^2$ (Explained sum of squares)

$$RSS = \sum_i (Y_i - \hat{Y}_i)^2 \text{ (Residual sum of Squares)}$$

Chú ý rằng $ESS = \sum_i (\hat{Y}_i - \overline{Y})^2 = \sum_i (\hat{a}X_i + \hat{b} - \overline{Y})^2$

$$= \sum_i (\hat{a}X_i + \overline{Y} - \hat{a}\overline{X} - \overline{Y})^2 = (\hat{a})^2 \sum_i (X_i - \overline{X})^2$$

$$= (\hat{a})^2 \left[\sum_i X_i^2 - n(\overline{X})^2\right]$$

Tiếp tục ví dụ 6, theo bảng số liệu đã tính toán ở ví dụ 1 ta có:

$$TSS = \sum_i (Y_i - \overline{Y})^2 = \sum_i Y_i^2 - n(\overline{Y})^2$$

$$= 15133{,}04 - 17.(28{,}4235)^2 = 1398{,}8$$

$$ESS = (\hat{a})^2 \left[\sum_i X_i^2 - n(\overline{X})^2\right]$$

$$= 2{,}885^2 (3686{,}8125 - 17.144{,}391^2) = 1380{,}4$$

Vì RSS = TSS – ESS nên ta có:

$$RSS = TSS - ESS = 1398,8 - 1380,4 = 16,4$$

Vậy $\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{16,4}{15} = 1,09$

Theo công thức (10.10) ta cũng có:

$$Se(\hat{a}) = \sqrt{\frac{\frac{\hat{\sigma}^2}{n}}{X^2 - (\overline{X})^2}} = \sqrt{\frac{\frac{1,09}{17}}{216,8713 - 14,391}} = 0,081$$

$$Se(\hat{b}) = \sqrt{\frac{\left(\frac{\hat{\sigma}^2}{n}\right)\overline{X}^2}{X^2 - (\overline{X})^2}} = \sqrt{\frac{\left(\frac{1,09}{17}\right).216,8713}{9,77}} = 1,91$$

Từ đây ta có thể tính được khoảng tin cậy cho a và b theo công thức (10.11) như sau:

$$(2,885 - 0,081.2,131 < a < 2,885 + 0,081.2,131)$$

$$\rightarrow \quad (2,713 < a < 3,058)$$

$$(-13,905 - 1,191.2,131 < b < -13,095 + 1,191.2,131)$$

$$\rightarrow \quad (-15,633 < b < -10,556)$$

2. Kiểm định giả thuyết đối với a, b

+ Với cặp giả thuyết

$$H_0 : (a = a_0) \; ; \; H_1 : (a \neq a_0)$$

Ta thấy nếu H_0 đúng thì

$$T = \frac{\hat{a} - a_0}{Se(\hat{a})} \sim T(n-2)$$

Do đó miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ t = \frac{\hat{a} - a_0}{Se(\hat{a})} ; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

+ Với cặp giả thuyết: $H_0 : (b = b_0)$; $H_1 : (b \neq b_0)$

Tương tự ta có miền bác bỏ là:

$$W_\alpha = \left\{ t = \frac{\hat{b} - b_0}{Se(b)} ; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

Trở lại ví dụ 10.6, giả sử ta muốn kiểm định cặp giả thuyết

$$H_0 : (a = a_0 = 3) ; H_1 : (a \neq 3)$$

Trước hết ta tính

$$t_{qs} = \frac{\hat{a} - a_0}{Se(\hat{a})} = \frac{2,885 - 3}{0,081} = -1,419$$

Ta thấy $|t_{qs}| < t_{\alpha/2}^{(n-2)} = t_{0,025}^{(15)} = 2,131$ do đó $t_{qs} \notin W_\alpha$. Chưa có cơ sở bác bỏ giả thuyết H_0, do đó có thể xem như hệ số a = 3.

3. Kiểm định sự phù hợp của hàm hồi quy

Phần này chúng ta sẽ làm tương tự như đối với bài toán phân tích phương sai (xem chương IX).

Theo công thức (10.12) ta có:

$$TSS = ESS + RSS$$

Như vậy nếu chúng ta lấy TSS để đo mức độ biến động của các giá trị của biến phụ thuộc Y xung quanh giá trị trung bình $(\overline{Y})$ của nó thì ta thấy TSS được chia thành 2 phần là ESS và TSS. Trong đó:

$$ESS = \sum (\hat{Y}_i - \overline{Y})^2$$

có thể được coi như là ước lượng của TSS và RSS là sai số do ước lượng mô hình hồi quy mà có. Có thể thấy ngay mô hình hồi quy sẽ càng tốt nếu kết quả ước lượng cho ta RSS càng nhỏ, tức là ESS càng gần với TSS.

Chia hai vế cho TSS ta được

$$1 = 100\% = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

Giá trị $R^2 = \frac{ESS}{TSS}$ được gọi là *hệ số xác định* (hay xác định bội trong trường hợp $p > 2$) **của mô hình hồi quy.**

Ý nghĩa của R^2: R^2 cho ta biết trong 100% của toàn bộ sự sai lệch (biến động) của Y so với giá trị trung bình của nó thì có bao nhiêu phần trăm là do biến X (còn gọi là biến giải thích) gây nên. Nói một cách khác mô hình đã giải thích được (R^2.100) phần trăm sự biến động của Y xung quanh giá trị trung bình của nó, số phần trăm còn lại do sai số ngẫu nhiên và do các yếu tố khác (nếu có) mà ta chưa đưa vào mô hình để xem xét.

Cũng với ví dụ 10.6 ta có:

$$R^2 = \frac{ESS}{TSS} = \frac{1382,4}{1398,8} = 0,988$$

Như vậy nếu R^2 càng lớn thì mô hình càng tốt: Mô hình càng giải thích được nhiều về sự biến động của Y (còn gọi là biến được giải thích). Việc kiểm định sự phù hợp, hay đúng đắn của hàm hồi quy chính là đi kiểm định cặp giả thuyết sau đây:

$$H_0 : (R^2 = 0) \; ; \; H_1 : (R^2 \neq 0) \qquad (10.13)$$

Cặp giả thuyết trên tương đương với cặp giả thuyết:

$$H_0 : (a = 0) \; ; \; H_1 : (a \neq 0)$$

Để kiểm định cặp giả thuyết (10.13) người ta sử dụng tiêu chuẩn sau đây:

$$F = \frac{\left(\frac{R^2}{1}\right)}{\frac{(1-R^2)}{(n-2)}}$$

Nếu H_0 đúng, F có phân phối $F(1, n-2)$.

Vậy miền bác bỏ để kiểm định giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{R^2}{\frac{(1-R^2)}{(n-2)}} ; F > f_\alpha(1, n-2) \right\}$$

Nhận xét: Đối với mô hình hồi quy tuyến tính đơn ($p = 2$) ta luôn có $R^2 = r_{xy}^2$

Trở lại ví dụ 6, ta đã tính được $r_{XY} = 0{,}9942$ do đó $r_{XY}^2 = 0{,}9942^2 = 0{,}9884 = R^2$.

4. Dự báo

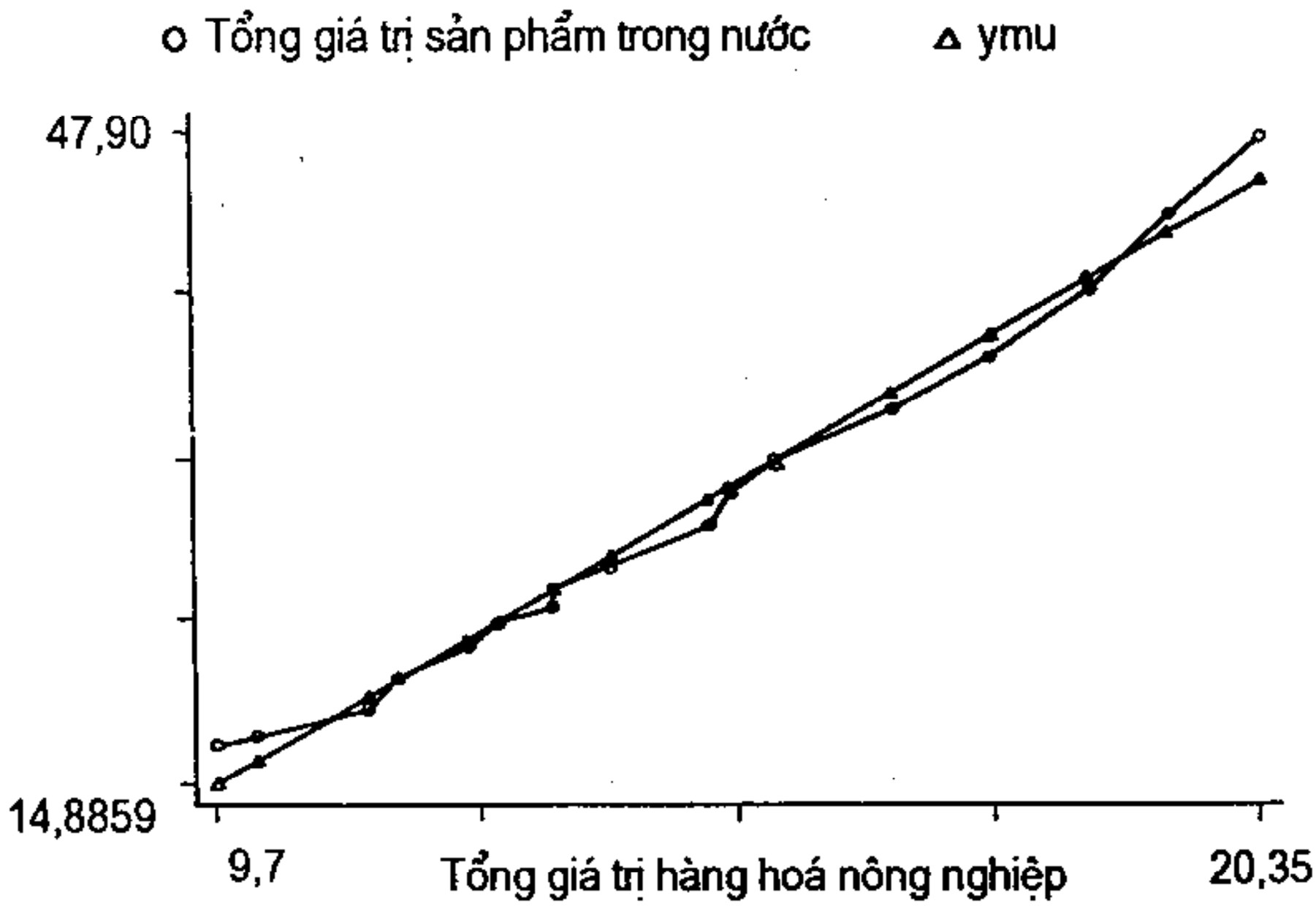
Chúng ta có $\hat{Y} = \hat{a}X + \hat{b}$ là hàm hồi quy ước lượng của hàm hồi quy lý thuyết:

$$f_Y(X) = aX + b$$

Tất nhiên ta cũng thấy Y cũng có quan hệ chặt chẽ với $\hat{Y}$. Với cùng các giá trị $\{X_i\}$ ta tính các giá trị hồi quy ước lượng:

$$\hat{Y}_i = \hat{a}X_i + \hat{b}$$

Sau đó ta vẽ đồ thị Y và các điểm $\hat{Y}_i$ trên cùng một hệ tọa độ, lúc đó ta sẽ thấy rõ hơn quan hệ giữa các giá trị thực tế $\{Y_i\}$ và các giá trị hồi quy ước lượng $\{\hat{Y}_i\}$. Với ví dụ 6 ta có hình ảnh sau đây:



Có thể xem $\hat{Y}$ như là đường xu thế của Y. $\hat{Y}$ là một công cụ đắc lực trong công tác dự báo ngắn hạn. Ta có hai loại dự báo sau đây (trên cơ sở sử dụng $\hat{Y}$):

- Dự báo trung bình có điều kiện của Y khi X nhận giá trị X_0.
- Dự báo giá trị cá biệt của Y khi $X = X_0$

+ Dự báo giá trị trung bình:

Giả sử $X = X_0$, ta có $\hat{Y}_0 = \hat{a}X_0 + \hat{b}$ là ước lượng điểm của $f_Y(X_0) = E(Y/X=x_0) = aX_0 + b$.

$\hat{Y}_0$ là ước lượng hiệu quả của $f_Y(X_0)$. Phương sai của $\hat{Y}_0$ là:

$$Var(\hat{Y}_0 = \sigma^2\left[\frac{1}{n} + \frac{X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$$

Thay σ^2 bằng $\hat{\sigma}^2$ ta có:

$$T = \frac{\hat{Y}_0 - (aX_0 + b)}{Se(\hat{Y}_0)} \sim T(n-2)$$

Trong đó $Se(\hat{Y}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]}$

Do đó khoảng tin cậy $(1 - \alpha)$ của $m_Y(x_0)$ là:

$$\left(\hat{Y}_0 - t_{\alpha/2}^{(n-2)} Se(\hat{Y}_0) < f_Y(X_0) < \hat{Y}_0 + t_{\alpha/2}^{(n-2)} Se(\hat{Y}_0)\right)$$

Ví dụ: Dự báo giá trị trung bình của tổng sản phẩm trong nước khi tổng giá trị hàng hóa nông nghiệp đạt mức 25 triệu tấn ($X_0 = 25$).

$$Se(\hat{Y}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]}$$

$$= \sqrt{1{,}09\left[\frac{1}{17} + \frac{(25 - 14{,}391)^2}{166{,}097}\right]} = 0{,}896$$

$$(1 - \alpha) = 0{,}95,\ t_{\alpha/2}^{(n-2)} = t_{0{,}025}^{(15)} = 2{,}131$$

$\hat{Y}_0 = \hat{a}X_0 + \hat{b} = 2{,}885.25 - 13{,}095 = 59{,}03$. Ta cũng có khoảng tin cậy cho f_y (25) là:

$$59,03 - 2,131 \,.\, 0,896 < f_Y(25) < 59,03 + 2,131 \,.\, 0,896$$

$$57,12 < f_Y(25) < 60,94$$

+ Dự báo giá trị cá biệt

Ký hiệu giá trị của Y ứng với $X = X_0$ là Y_0, ta cũng lấy $\hat{Y}_0 = \hat{a}X_0 + \hat{b}$ làm ước lượng điểm cho giá trị Y_0.

$$\text{Vì } Var(Y_0 - \hat{Y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]$$

$$\text{nên ta có: } T = \frac{Y_0 - \hat{Y}_0}{Se(Y_0 - \hat{Y}_0)} \sim T(n-2)$$

$$\text{Trong đó: } Se(Y_0 - \hat{Y}_0) = \sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right]}$$

Do đó khoảng tin cậy của Y_0 là:

$$\left(\hat{Y}_0 - t_{\alpha/2}^{(n-2)} Se(Y_0 - \hat{Y}_0) < Y_0 < \hat{Y}_0 + t_{\alpha/2}^{(n-2)} Se(Y_0 - \hat{Y}_0)\right)$$

Ví dụ: Ta đi tìm khoảng tin cậy cho giá trị tổng sản phẩm trong nước khi tổng giá trị hàng hóa nông nghiệp đạt mức 25 triệu tấn ($X = X_0 = 25$).

$$Se(Y_0 - \hat{Y}_0) = \sqrt{1,09\left[1 + \frac{1}{17} + \frac{(25 - 14,391)^2}{166,097}\right]} = 1,376$$

$$59,03 - 2,131 \,.\, 1,376 < Y_0 < 59,03 + 2,131 \,.\, 1,376$$

$$\rightarrow \qquad 56,098 < Y_0 < 61,962$$

Sau đây là kết quả phân tích hồi quy đối với bài toán đã nêu ở ví dụ 10.6 được chạy bởi Stata. Các bạn hãy thử đọc kết

quả được in ra và so sánh với những kết quả mà chúng ta đã tính bằng tay ở trên.

• regress gdp gap

Source	SS	df	MS	Number of obs	=	17
Model	1382.47717	1	1382.47717	F(1,15)	=	1271.17
Residual	16.313456	15	1.08756373	Prob > F	=	0.0000
				R-squared	=	0.9883
Total	1398.79063	16	87.4244141	Adj R-squared	=	0.9876
				Root MSE	=	1.0429

gdp	Coef.	Std.Err.	t	P > |t|	[95% Conf. Interval]	
gap	2.885761	.0809392	35.653	0.000	2.713243	3.058278
cons	−13.105	1.19195	−10.995	0.000	−15.64655	−10.56537

3.3. Mô hình hồi quy tuyến tính bội

Ta xét trường hợp tổng quát: $p > 2$, tức là ta có p biến ngẫu nhiên: $Y, X_2, X_3 \ldots X_p$.

Giả sử hàm hồi quy (lý thuyết) của Y đối với các biến $X_2, X_3, \ldots, X_p$ có dạng sau đây:

$$f_Y(X_2, X_3, \ldots, X_p) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

$$= \beta_1 + \sum_{j=2}^{p} \beta_j X_j$$

Với ma trận quan sát là:

$$\begin{pmatrix} Y_1 & X_{12} & X_{13} & \cdots & X_{1p} \\ Y_2 & X_{22} & X_{23} & \cdots & X_{2p} \\ & & \cdots & & \\ Y_n & X_{n2} & X_{n3} & \cdots & X_{np} \end{pmatrix}$$

Ta có mô hình hồi quy như sau:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_p X_{ip} + U_i \quad i = \overline{1,n} \quad (10.14)$$

Trong đó ta cũng giả thiết các sai số ngẫu nhiên U_i độc lập với nhau và cùng tuân theo quy luật $N(0, \sigma^2)$.

Ký hiệu:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{pmatrix}; X = \begin{pmatrix} 1 & X_{12} & ... & X_{1p} \\ 1 & X_{22} & ... & X_{2p} \\ & & ... & \\ 1 & X_{n2} & ... & X_{np} \end{pmatrix}; X_j = \begin{pmatrix} Y_{1j} \\ Y_{2j} \\ ... \\ Y_{nj} \end{pmatrix} \; j = \overline{2,p}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{pmatrix} \text{ và } U = \begin{pmatrix} U_1 \\ U_2 \\ ... \\ U_n \end{pmatrix}; I = \begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix}$$

Ta có thể viết mô hình (10.14) dưới dạng ma trận như sau:

$$Y = X\beta + U$$

1. Ước lượng vectơ tham số β

Ta đi tìm ước lượng $\hat{\beta}$ theo phương pháp bình phương bé nhất, tức là tìm $\hat{\beta}$ sao cho:

$$\sum_i (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_p X_{pi})^2 =$$

$$= \sum e_i^2 = \min_\beta \sum (Y_i - \beta_1 - \beta_2 X_{2i} - ... - \beta_p X_{pi})^2$$

$$= \min_\beta S(\beta_1, \beta_2, ..., \beta_p)$$

Hoặc viết dưới dạng vectơ:

$$\sum_i (Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2 = \|e\|^2 = \min$$

Dễ dàng thấy rằng $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ là nghiệm của hệ phương trình tuyến tính sau đây:

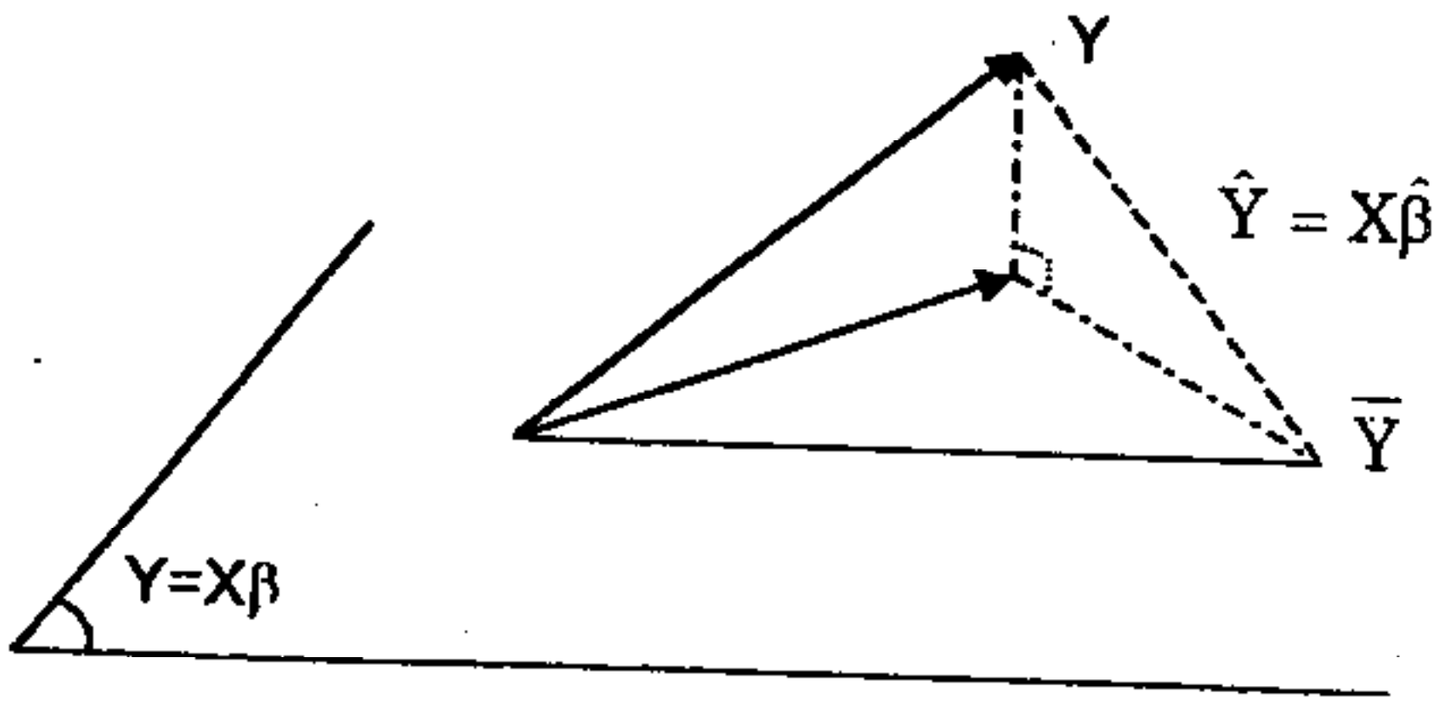
$$\begin{cases} \dfrac{\partial S(\beta_1, \beta_2, ... \beta_p)}{\partial \beta_j} = 0 \\ j = \overline{1,p} \end{cases}$$

Tuy nhiên việc giải hệ phương trình tuyến tính trên một cách tổng quát sẽ rất cồng kềnh và phức tạp. Có thể sử dụng hình học giải tích và một số phép tính cơ bản của ma trận để tìm ra biểu thức tính vectơ $\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, ... \hat{\beta}_p)$ một cách nhanh chóng và đơn giản hơn xét về mặt mô tả ký hiệu.

Nếu coi β như một vectơ biến thì $F = X\beta$ chính là phương trình của một siêu phẳng trong không gian R^n với các cơ sở là các vectơ I, X_2, X_3, ...X_p. Tức là ta cũng phải giả thiết rằng I, X_2, ... X_p là độc lập tuyến tính hay ma trận X có hạng là p.

Vì $\|Y - X\hat{\beta}\|^2 = \|Y - \hat{Y}\|^2 = \min$ nên ta suy ra rằng $\hat{Y}$ là hình chiếu của Y lên siêu phẳng F. Tức là ta có: $(Y - \hat{Y})$ trực giao với F. Vì I, X_2 ... X_p là cơ sở của F do đó $(Y - \hat{Y})$ cũng trực giao với tất cả những vectơ đó.

Ta có thể mô tả bằng hình vẽ sau đây:



Hình 10.1

Vì $(Y - \hat{Y})$ vuông góc với $I, X_2, ..., X_p$ nên tích vô hướng của $(Y - \hat{Y})$ với những vectơ đó phải bằng không. Tức là ta có:

$$(X_j,(Y - \hat{Y})) = \sum X_{ij}(Y_i - \hat{Y}_i) = 0 \ (j = \overline{2,p})$$

và $(I,(Y - \hat{Y})) = \sum(Y_i - \hat{Y}_i) = 0$

Sử dụng ký hiệu ma trận ta có thể viết lại như sau:

$$(X_j,(Y - \hat{Y})) = X'_j(Y - \hat{Y}) = 0 \ \ (j = \overline{2,p})$$

$$(I,(Y - \hat{Y})) = I'(Y - \hat{Y}) = 0$$

Hay: $X'(Y - \hat{Y}) = 0$. Thay $\hat{Y} = X\hat{\beta}$ vào ta được

$$X'(Y - \hat{Y}) = X'(Y - X\hat{\beta}) = 0 \rightarrow X'Y = X'X\hat{\beta}$$

Vì hạng của X bằng p nên (X'X) cũng có hạng bằng p, do đó sẽ tồn tại ma trận nghịch đảo $(X'X)^{-1}$. Ta có:

$$(X'X)^{-1}X'Y = (X'X)^{-1}X'X\hat{\beta}$$

Vậy: $\hat{\beta} = (X'X)^{-1}X'Y$ (10.15)

$\hat{\beta}$ có ma trận hiệp phương sai là:

$$\text{cov}(\hat{\beta}) = \{\text{cov}(\hat{\beta}_k,\hat{\beta}_j)\} = E[\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

Thay $\hat{\beta} = (X'X)^{-1}X'(X\beta + U)$ vào biểu thức trên ta được:

$$\text{cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \{\sigma_{kj}\} \ k, \ \ j = \overline{1,p}$$

Vì σ^2 chưa biết, chúng ta thay nó bằng ước lượng không chệch:

$$\hat{\sigma}^2 = \frac{\left\|Y - \hat{Y}\right\|^2}{n - p} = \frac{\|e\|^2}{n - p} = \frac{\sum e_i^2}{n - p} \qquad (10.16)$$

Các phần tử nằm trên đường chéo của ma trận hiệp

phương sai $\{\sigma kj\} = \sigma^2 (X'X)^{-1}$ chính là các phương sai của các ước lượng $\hat{\beta}_j$. Tức là: $Var(\hat{\beta}_j) = \sigma_{jj}, \quad j = \overline{1,p}$

Do đó ta có: $Se(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)} = \sqrt{\sigma_{jj}}$

2. Khoảng tin cậy và kiểm định giả thuyết về các tham số hồi quy

a. Khoảng tin cậy cho tham số hồi quy

Thông thường thì σ^2 chưa biết, do đó chúng ta thay bằng $\hat{\sigma}^2$. Khi đó các thống kê:

$$T = \frac{\hat{\beta}_j - \beta_j}{Se(\hat{\beta}_j)}, \qquad j = \overline{1,p}$$

sẽ có phân bố Student với (n – p) bậc tự do: T(n – p). Vậy khoảng tin cậy (1 – α) của các β_j sẽ là:

$$\left(\hat{\beta}_j - t_{\alpha/2}^{(n-p)} Se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{\alpha/2}^{(n-p)} Se(\hat{\beta}_j)\right)$$

b. Kiểm định giả thuyết về tham số hồi quy

+ Nếu muốn biết X_j có độc lập với Y không chúng ta kiểm định cặp giả thuyết sau:

$$H_0 : (\beta_j = 0) \; ; \; H_1 : (\beta_j \neq 0)$$

với miền bác bỏ được xác định như sau:

$$W_\alpha = \left\{ t = \frac{\hat{\beta}_j}{Se(\hat{\beta}_j)} ; |t| > t_{\alpha/2}^{(n-p)} \right\}$$

+ Muốn kiểm định về mức độ ảnh hưởng của X_j đối với sự biến động của Y chúng ta xét cặp giả thuyết:

$$H_0 : (\beta_j = \beta^*); \; H_1 : \beta_j \neq \beta^*)$$

Miền bác bỏ trong trường hợp này là:

$$W_\alpha = \left\{ t = \frac{\hat{\beta}_j - \beta^*}{Se(\hat{\beta}_j)}; |t| > t_{\alpha/2}^{(n-p)} \right\}$$

+ Để so sánh mức độ ảnh hưởng của hai biến X_j và X_k đối với sự biến động của Y chúng ta có thể tiến hành kiểm định cặp giả thuyết sau:

$$H_0 : (\beta_j \pm \beta_k = \beta^*);\ H_1 : (\beta_j \pm \beta_k \neq \beta^*)$$

với miền bác bỏ là:

$$W_\alpha = \left\{ t = \frac{\hat{\beta}_j \pm \hat{\beta}_k - \beta^*}{Se(\hat{\beta}_j \pm \hat{\beta}_k)}; |t| > t_{\alpha/2}^{(n-p)} \right\}$$

Trong đó $Se(\hat{\beta}_j \pm \hat{\beta}_k)$ xác định như sau:

$$Se(\hat{\beta}_j \pm \hat{\beta}_k) = \sqrt{Var(\hat{\beta}_j) + Var(\hat{\beta}_k) \pm 2\,cov(\hat{\beta}_j, \hat{\beta}_k)}$$

3. Hệ số xác định bội

Từ hình vẽ 10.1 dễ dàng thấy rằng:

$$\left\| Y - \overline{Y} \right\|^2 = \left\| Y - \hat{Y} \right\|^2 + \left\| \hat{Y} - \overline{Y} \right\|^2$$

Hay: $\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \overline{Y})^2$

Giống như trường hợp chúng ta xét mô hình hồi quy đơn, ở đây chúng ta cũng có:

$$TSS = RSS + ESS$$

Chia 2 vế cho TSS ta được:

$$1 = 100\% = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

$$\text{Đặt: } R^2 = \frac{ESS}{TSS} = \frac{\left\|\hat{Y} - \overline{Y}\right\|^2}{\left\|Y - \overline{Y}\right\|^2} = 1 - \frac{RSS}{TSS}$$

R^2 chính là tỷ lệ (số phần trăm) sự biến động của Y xung quanh giá trị trung bình $\overline{Y}$ được giải thích bởi mô hình hồi quy. Mặt khác, từ ý nghĩa hình học ta thấy R^2 cũng chính là cosin của góc được tạo bởi hai vectơ $(\hat{Y} - \overline{Y})$ và $(Y - \overline{Y})$. Góc đó càng nhỏ thì độ chính xác của ước lượng càng cao, tức là R^2 càng lớn thì mô hình càng tốt.

+ Xét cặp giả thuyết:

$$H_0 : (R^2 = 0); H_1 : (R^2 \neq 0)$$

Cặp giả thuyết này tương đương với cặp giả thuyết sau:

$H_0 : (\beta_2 = \beta_3 = ... = \beta_p = 0)$

H_1: (Tồn tại ít nhất một $\beta_j \neq 0$)

Chúng ta sử dụng tiêu chuẩn: $F = \dfrac{\dfrac{R^2}{(p-1)}}{\dfrac{(1-R^2)}{(n-p)}}$

F sẽ tuân theo quy luật $F(p-1, n-p)$ nếu H_0 đúng.

Vậy miền bác bỏ để kiểm định cặp giả thuyết trên là:

$$W_\alpha = \left\{ F = \frac{\dfrac{R^2}{(p-1)}}{\dfrac{(1-R^2)}{(n-p)}} ; F > f_\alpha(p-1, n-p) \right\}$$

Chú ý:

1. Về mặt thực hành, việc áp dụng mô hình hồi quy

tuyến tính cùng với phương pháp bình phương bé nhất chỉ có ý nghĩa (đáng tin cậy) khi các giả thiết về mô hình được thỏa mãn. Để kiểm tra xem các giả thiết này có được thỏa mãn không người ta có thể sử dụng các tiêu chuẩn khác nhau. Ví dụ, để kiểm định giả thuyết về sự phân phối chuẩn của các U_i, chúng ta có thể sử dụng các tiêu chuẩn χ^2, Kolmogorov, Sktest... như đã trình bày ở chương 8. Về vấn đề này có thể tham khảo các sách về kinh tế lượng (ví dụ như cuốn: *Basic econometrics* của Damodar N. Gujarati hoặc giáo trình *Kinh tế lượng*, NXB KHKT - 1998).

2. Trên thực tế số liệu thống kê cần phân tích thường rất lớn do đó việc tính toán bằng tay là hầu như không thể làm được. Để có thể sử dụng thống kê như một công cụ phân tích có hiệu quả thì việc biết sử dụng các phần mềm thống kê là gần như bắt buộc. Các phần mềm thống kê hay được sử dụng và được phổ biến rộng rãi hiện nay là: SPSS, STATA, ...

Cuối cùng, để minh họa cho phần này chúng ta hãy trở lại với tệp số liệu XSTK10_1 (xem §2), ta có kết quả phân tích hồi quy mô hình sau đây:

$$gdp_i = \beta_1 ex_i + \beta_2 im_i + \beta_3 gip_i + \beta_4 gap_i + \beta_5 + Ui \quad (i = \overline{1,17})$$

bằng Stata như sau:

```
regress gdp gap
```

Source	SS	df	MS	Number of obs	=	17
Model	1397.57741	4	349.394352	F(4,12)	=	3455.88
Residual	1.21321847	12	.101101539	Prob > F	=	0.0000
				R-squared	=	0.9991
Total	1398.79063	16	87.4244141	Adj R-squared	=	0.9988
				Root MSE	=	.31796

gdp	Coef.	Std.Err.	t	P > |t|	[95% Conf.	Interval]
gap	1.232316	.2191744	5.623	0.000	.7547759	1.709856
gip	.6063006	.1301608	4.658	0.001	.3227045	.8898966
ex	1.181809	.2596261	4.552	0.001	.6161321	1.747485
im	-.43687	.175753	-2.486	0.029	-.3198064	-.0539405
cons	.84561	1.687897	0.501	0.625	-2.832	4.623224

Muốn kiểm định giả thuyết về các hệ số của hàm hồi quy, ví dụ muốn kiểm định giả thuyết hệ số của biến ex bằng 1 ta gõ lệnh test ex = 1, hoặc kiểm định giả thuyết về hệ số tự do (hệ số chặn) bằng 0 ta gõ lệnh test cons = 0. Sau đây là kết quả của các kiểm định trên:

. test ex=1

(1) ex = 1.0

F(1, 12) = 0.49

Prob > F = 0.4971

. test _cons=0

(1) _cons = 0.0

F(1, 12) = 0,25

Prob > F = 0.6254

Để kết thúc phần này ta xét thêm một ví dụ nữa để tham khảo. Nếu có điều kiện độc giả hãy thử chạy chương trình Stata và so sánh với kết quả in ra dưới đây:

```
.d
Contains data from C:\WINSTATA\XSTK10_4.dta
obs   :  16
vars  :   5                                  25 Aug 1998 10:08
size  :  384 (85.9% of memory free)
```

1. nam	float %9.0g	
2. gdp	float %9.2g	Tong s.p trong nuoc (1000 ty)
3. dien	float %9.2g	San luong dien (ty Kwh)
4. ximang	float %9.2g	San luong xi mang (trieu tan)
5.dien thoai	float %9.2g	So may dien thoai (1000 chiec)

Sorted by:

list	nam	gdp	dien	ximang	dthoai
1	1980	16.30	3.60	.60	90.60
2	1981	17.20	3.80	.60	97.40
3	1982	18.70	4.10	.70	98.50
4	1983	20.10	4.30	1.00	103.80
5	1984	21.80	5.00	1.30	108.70
6	1985	23.00	5.20	1.50	103.10
7	1986	23.80	5.70	1.50	113.50
8	1987	24.70	6.20	1.70	116.10
9	1988	25.90	7.00	2.00	109.10
10	1989	28.00	7.90	2.10	110.70
11	1990	29.50	8.80	2.50	114.40
12	1991	31.30	9.30	3.10	121.10
13	1992	34.00	9.80	3.90	132.10
14	1993	36.70	10.90	4.80	268.30
15	1994	40.00	12.50	5.40	420.00
16	1995	43.80	14.70	5.90	600.00

. correlate gdp dien ximang dthoai

(obs = 16)

	gdp	dien	ximang	dthoai
gdp	1.0000			
dien	0.9946	1.0000		
ximang	0.9888	0.9821	1.0000	
dthoai	0.8176	0.8338	0.8488	1.0000

. pcorr gdp dien ximang dthoai

(obs = 16)

Partial correlation of gdp with

Variable	Corr.	Sig.
dien	0.8719	0.000
ximang	0.7207	0.004
dthoai	–0.5075	0.064

. regress gdp dien ximang dthoai

Source	SS	df	MS
Model	990.84684	3	330.28228
Residual	4.90251119	12	.408542599
Total	995.749351	15	66.38329

Number of obs	=	16
F(3,12)	=	808.44
Prob > F	=	0.0000
R-squared	=	0.9951
Adj R- quared	=	0.9938
Root MSE	=	.63917

gdp	Coef.	Std.Err.	t	P > |t|	[95% Conf.	Interval]
dien	1.623106	.2631666	6.168	0.000	1.049715	2.196497
ximang	1.908267	–5298553	3.601	0.004	.753812	3.062723
dthoai	–.004454	.0021831	–2.041	0.064	–.009211	.0003019
_cons	11.30476	.8021872	14.092	0.000	9.556943	13.05257

. test dthoai = 0

(1) dthoai = 0.0

F(1, 12) = 4.16

Prob > F = 0.0639

. test dien = 2

(1) dien = 2.0

F(1, 12) = 2.05

Prob > F = 0.1776

3.4. Một số dạng hàm hồi quy phi tuyến có thể đưa về dạng hàm hồi quy tuyến tính

1. Hồi quy lũy thừa

$$f_Y(X_2, X_3, ..., X_p) = \alpha X_2^{\beta_2} X_3^{\beta_3} ... X_p^{\beta_p}$$

Lấy logarit hai vế ta được:

$$\ln f_Y(X_2, ..., X_p) = \ln\alpha + \beta_2 \ln X_2 + ... + \beta_p \ln X_p$$

Đặt: $\ln\alpha = \beta_1$, $\ln X_j = V_j$ và $\ln Y = Z$ ta có mô hình hồi quy tuyến tính sau đây:

$$Zi = \beta_1 + \beta_2 V_{i2} + \beta_3 V_{i3} + ... + \beta_p V_{ip} + U_i \qquad i = \overline{1,n}$$

Ví dụ: Xét hàm sản xuất Cobb – Douglas;

$$Q = aK^{\alpha} L^{\beta} e^{\gamma t} X^{\delta} e^{u}$$

Ta có thể ước lượng các hệ số của hàm này bằng cách sử dụng mô hình sau đây:

$$\ln Q_i = \ln a + \alpha \ln K_i + \beta \ln L_i + \gamma t_i + \delta \ln X_i + U_i \qquad i = \overline{1,n}$$

2. Hồi quy mũ

$$f_Y(X_2,...,X_p) = e^{\beta_1+\beta_2X_2+...+\beta_pX_p}$$

Lấy logarit hai vế ta có hàm hồi quy tuyến tính:

$$\ln f_Y(X_2, ..., X_p) = \beta_1 + \beta_2X_2 + ... + \beta_pX_p$$

Đặt $\ln Y = Z$ ta có mô hình sau đây:

$$Zi = \beta_1 + \beta_2 X_{i2} + ... + \beta_p X_{ip} + U_i$$

Khi đó:

$$\hat{Z} = \hat{\beta}_1 + \hat{\beta}_2X_2 + ... + \hat{\beta}_pX_p$$

là ước lượng của $\ln f_Y(X_2, ..., X_p) = \ln E(Y/X_2, ..., X_p)$

3. Hồi quy parabol

$$f_Y(X) = aX^2 + bX + c$$

Đặt $X = V_1$, $X^2 = V_2$ ta có mô hình tuyến tính:

$$Y_i = aV_{i2} + bV_{i1} + C + U_i \qquad i = \overline{1,n}$$

4. Hồi quy hyperbol bội

$$f_y(X) = \beta_1 + \frac{\beta_2}{X_2} + \frac{\beta_3}{X_3} + ... + \frac{\beta_p}{X_p}$$

Đặt $\frac{1}{X_j} = V_j$ $(j = \overline{2,p})$ ta có mô hình tuyến tính:

$$Y_i = \beta_1 + \beta_2V_{i2} + \beta_3V_{i3} + ... + \beta_pV_{ip} + U_i \qquad (i = \overline{1,n})$$

5. Hồi quy logarit

Trên thực tế chúng ta gặp nhiều trường hợp hàm hồi quy không phải là tuyến tính nhưng nếu thực hiện phép logarit hóa các biến xem xét thì người ta lại thấy chúng thỏa mãn mô hình hồi quy tuyến tính. Tức là chúng ta có mô hình sau đây:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{i2} + ... + \beta_p \ln X_{ip} + U_i \qquad (i = \overline{1,n})$$

Chúng ta có thể áp dụng phương pháp bình phương bé nhất như đã làm ở phần §3.3 để tìm các ước lượng $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_p$ bằng cách đặt $\ln Y = Z$ và $\ln X_j = V_j$.

Các ký hiệu và công thức cơ bản

1. Phân tích tương quan bảng số liệu định lượng:

+ Hệ số tương quan mẫu:

$$r_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{MS_X}\sqrt{MS_Y}}$$

+ Hệ số tương quan riêng phần:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$r_{12,34...p} = \frac{r_{12,34...(p-1)} - r_{1p,34...(p-1)}r_{2p,34...(p-1)}}{\sqrt{(1 - r_{1p,34...(p-1)}^2)(1 - r_{2p,34...(p-1)}^2)}}$$

2. Phân tích tương quan bảng số liệu định tính

+ Bảng ngẫu nhiên hai chiều (2 × 2)

A \ B	b_1	b_2
a_1	n_1	n_2
a_2	n_3	n_4

- Hệ số $Q = \dfrac{n_1 n_4 - n_2 n_3}{n_1 n_4 + n_2 n_3}$

- Hệ số $F = \dfrac{n_1 n_4 - n_2 n_3}{\sqrt{(n_1 + n_2)(n_1 + n_3)(n_2 + n_4)(n_3 + n_4)}}$

+ Bảng ngẫu nhiên hai chiều (p × q)

- Hệ số liên hợp Pearson: $P = \sqrt{\dfrac{\chi^2}{n + \chi^2}}$

- Hệ số Cramer: $K = \left\{\dfrac{\chi^2}{n.\min(p-1, q-1)}\right\}^{1/2}$

+ Nếu A, B được xếp hạng là $\{(i, k_i)\}$ với $i = \overline{1, n}$ ta có:

- Hệ số tương quan hạng Spearman:

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (d_i = i - k_i)$$

- Hệ số tương quan hạng Kendall:

$$\tau = \frac{S}{\dfrac{n(n-1)}{2}}$$

$S = \sum(S_i^+ - S_i^-)$ = số điểm Kendall

S_i^+ = số hạng k_j (j > i) thỏa mãn $k_j > k_i$

S_i^- = số hạng k_j (j > i) thỏa mãn $k_j < k_i$

3. Phân tích hồi quy

+ Hàm hồi quy

$$f_Y(X_2, X_3, \ldots, X_p) = E(Y/X_2, X_3, \ldots, X_p)$$

+ Hàm hồi quy tuyến tính:

$$f_Y(X_2, X_3, ..., X_p) = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p$$

+ Mô hình hồi quy tuyến tính đơn:

$$Y_i = aX_i + b + U_i \qquad i = \overline{1,n}$$

- Ước lượng tham số hồi quy:

$$\hat{a} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X}^2 - (\overline{X})^2} = r_{XY}\frac{\sqrt{MS_Y}}{\sqrt{MS_X}}$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}$$

- Khoảng tin cậy của tham số hồi quy:

$$\hat{a} - t_{\alpha/2}^{(n-2)} Se(\hat{a}) < a < \hat{a} + t_{\alpha/2}^{(n-2)} Se(\hat{a})$$

- Kiểm định giả thuyết về tham số hồi quy

$$H_0 : (a = a_0);\ H_1 : (a \neq a_0)$$

$$W_\alpha = \left\{ t = \frac{\hat{a} - a_0}{Se(\hat{a})}; |t| > t_{\alpha/2}^{(n-2)} \right\}$$

Chú ý: Nếu thay a bằng b ta sẽ có công thức khoảng tin cậy và kiểm định giả thuyết về tham số b.

- Kiểm định sự phù hợp của hàm hồi quy:

$$H_0 : (R^2 = 0);\ H_1 : (R^2 \neq 0)$$

$$W_\alpha = \left\{ F = \frac{R^2}{\frac{(1-R^2)}{(n-2)}}; F > f_\alpha(1, n-2) \right\}$$

- Dự báo giá trị trung bình $E(Y/X=X_0)$

$$\hat{y}_0 = \hat{a}X_0 + \hat{b}$$

$$Se\hat{Y}_0 = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]}; \hat{\sigma}^2 = \frac{RSS}{n-2}$$

$$\left(\hat{Y}_0 - t^{(n-2)}_{(\alpha/2)} Se(\hat{Y}_0) < E(Y/X_0) < \hat{Y}_0 + t^{(n-2)}_{\alpha/2} Se(\hat{Y}_0)\right)$$

Câu hỏi ôn tập

1. Cho biết sự giống nhau và khác nhau giữa hàm hồi quy và mô hình hồi quy.

2. Khi nào các ước lượng bình phương bé nhất của các tham số hồi quy là các ước lượng không chệch tốt nhất? Hãy giải thích rõ các tính chất đó có nghĩa là gì?

3. Phân tích hồi quy và phân tích tương quan giống nhau và khác nhau ở những điểm nào?

4. Nội dung của phương pháp bình phương bé nhất trong phân tích hồi quy là gì? Phân biệt e_i và U_i như thế nào?

5. Khi nào thì 2 cặp giả thiết sau đây là tương đương nhau:

a) $H_0 : (\beta_j = 0)$; $H_1 : (\beta_j \neq 0)$

b) $H_0 : (R^2 = 0)$; $H_1 : (R^2 \neq 0)$

6. Cho mô hình hồi quy:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad i = \overline{1,n}$$

Trong đó Y là nhu cầu hàng hoá A, X_2 là thu nhập của người tiêu dùng, X_3 là giá của hàng hoá A.

a) Nếu thu nhập tăng một đơn vị thì nhu cầu trung bình tăng tối đa bao nhiêu?

b) Nếu giá của A tăng một đơn vị thì nhu cầu trung bình giảm tối đa bao nhiêu?

Phần phụ lục: CÁC BẢNG SỐ

Phụ lục 1: Phân phối nhị thức

Phụ lục 2: Phân phối Poisson

Phụ lục 3: Giá trị hàm e^{-x}

Phụ lục 4: Giá trị hàm

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Phụ lục 5: Giá trị hàm

$$\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{z^2}{2}} dz$$

Phụ lục 6: Giá trị tới hạn U_α

Phụ lục 7: Giá trị tới hạn $\chi_\alpha^{2(n)}$

Phụ lục 8: Giá trị tới hạn $T_\alpha^{(n)}$

Phụ lục 9: Giá trị tới hạn $F_\alpha^{(n_1,n_2)}$

Phụ lục 10: Bảng số ngẫu nhiên

Phụ lục 11: Giá trị kiểm định tổng hạng Wilcoxon

Phụ lục 12: Giá trị kiểm định tổng hạng theo dấu Wilcoxon

Phụ lục 13: Giá trị phân phối Cochran

Phụ lục 14: Giá trị kiểm định Kolmogorov

Phụ lục 15: Giá trị kiểm định Lilliefors

Phụ lục 1: Phân phối nhị thức
P(X = x) = b(x; n, p)

n	x	p = 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0 0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4239	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0312
	1	0,2036	0,3280	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1562
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0004	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1562
6	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0312
	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	00305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3310	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1338	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2153	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2760	0,1977	0,1373	0,0896	0,0548	0,0312
	2	0,0515	0,1488	0,2376	0,2936	0,3135	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0231	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,004	0,0012	0,0033	0,0079	0,0164	0,0312
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0007	0,0039

Phụ lục 1 *(tiếp theo)*

n	x	p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0277	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	,0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2308	0,2119	0,16.41
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2315	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
	2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0533	0,0269
	3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
	4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
	5	0,0003	0,0025	0,0132	0,0388	0,0803	0,1231	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0066	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002

Phụ lục 1 *(tiếp theo)*

n	x	p = 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,3312	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
	8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0006	0,0001
14	0	0,4877	0,2268	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
	2	0,1229	0,2570	0,2912	0,2501	0,1502	0,1134	0,0634	0,0317	0,0141	0,0056
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
	4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0200	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1266	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1956
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16	0	0,4401	0,1853	0,0743	0,0283	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
	1	0,3706	0,3294	0,2097	0,1126	0,0535	0,0228	0,0087	0,0030	0,0009	0,0002
	2	0,1463	0,2745	0,2775	0,2111	0,1336	0,0732	0,0353	0,0150	0,0056	0,0018

Phụ lục 1 *(tiếp theo)*

n	x	p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
16	3	0,0359	0,1423	0,2285	0,2463	0,2079	0,1465	0,0008	0,0468	0,0215	0,0083
	4	0,0061	0,0514	0,1311	0,2001	0,2252	0,2040	0,1553	0,1014	0,0572	0,0278
	5	0,0008	0,0137	0,0555	0,1201	0,1802	0,2099	0,2008	0,1623	0,1123	0,0667
	6	0,0001	0,0028	0,0108	0,0550	0,1101	0,1649	0,1982	0,1983	0,1684	0,1222
	7	0,0000	0,0004	0,0045	0,0197	0,0524	0,1010	0,1524	0,1889	0,1969	0,1746
	8	0,0000	0,0001	0,0009	0,0053	0,0197	0,0487	0,0923	0,1417	0,1812	0,1964
	9	0,0000	0,0000	0,0001	0,0012	0,0058	0,0185	0,0442	0,0840	0,1318	0,1746
	10	0,0000	0,0000	0,0000	0,0002	0,0014	0,0056	0,0167	0,0392	0,0755	0,1222
	11	0,0000	0,0000	0,0000	0,0000	0,0002	0,0013	0,0049	0,0142	0,0337	0,0667
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0040	0,0115	0,0278
	13	0,0000	0,0000	0,0000	0,0000	0,0800	0,0000	0,0002	0,0008	0,0029	0,0085
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
17	0	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
	1	0,3741	0,3150	0,1893	0,0957	0,0426	0,0169	0,0060	0,0019	0,0005	0,0001
	2	0,1575	0,2800	0,2673	0,1914	0,1136	0,0581	0,0260	0,0102	0,0035	0,0010
	3	0,0415	0,1556	0,2359	0,2393	0,1893	0,1245	0,0701	0,0341	0,0144	0,0052
	4	0,0076	0,0605	0,1457	0,2093	0,2209	0,1868	0,1320	0,0796	0,0411	0,0182
	5	0,0010	0,0175	0,0668	0,1361	0,1914	0,2081	0,1849	0,1319	0,0875	0,0472
	6	0,0001	0,0039	0,0236	0,0680	0,1276	0,1784	0,1991	0,1839	0,1432	0,0944
	7	0,0000	0,0007	0,0065	0,0267	0,0668	0,1201	0,1685	0,1927	0,1841	0,1484
	8	0,0000	0,0001	0,0014	0,0084	0,0279	0,0644	0,1143	0,1606	0,1883	0,1855
	9	0,0000	0,0000	0,0003	0,0021	0,0093	0,0276	0,0611	0,1070	0,1540	0,1855
	10	0,0000	0,0000	0,0000	0,0004	0,0025	0,0095	0,0263	0,0571	0,1008	0,1404
	11	0,0000	0,0000	0,0000	0,0001	0,0005	0,0026	0,0090	0,0242	0,0525	0,0944
	12	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0081	0,0215	0,0472
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0021	0,0068	0,0182
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000,	0,0001	0,0004	0,0016	0,0052
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
18	0	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,3763	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
	2	0,1683	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
	3	0,0473	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0033
	4	0,0093	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
	5	0,0014	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
	6	0,0002	0,0052	0,0316	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708
	7	0,0000	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
	8	0,0000	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
	9	0,0000	0,0000	0,0004	0,0033	0,0139	0,0306	0,0794	0,1284	0,1694	0,1855
	10	0,0000	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669
	11	0,0000	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
	12	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0045	0,0134	0,0327
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117

Phụ lục 1 *(tiếp theo)*

n	x	p = 0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
19	0	0,3774	0,1351	0,0456	0,0044	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
	1	0,3774	0,2852	0,1529	0,0685	0,0268	0,0093	0,0029	0,0008	0,0002	0,0000
	2	0,1787	0,2852	0,2428	0,1540	0,0803	0,0358	0,0138	0,0046	0,0013	0,0003
	3	0,0533	0,1796	0,2428	0,2182	0,1517	0,0869	0,0422	0,0175	0,0062	0,0018
	4	0,0112	0,0798	0,1714	0,2182	0,2023	0,1491	0,0909	0,0467	0,0203	0,0074
	5	0,0018	0,0266	0,0907	0,1636	0,2023	0,1916	0,1468	0,0933	0,0497	0,0222
	6	0,0002	0,0069	0,0374	0,0955	0,1374	0,1916	0,1844	0,1451	0,0949	0,0518
	7	0,0000	0,0014	0,0122	0,0443	0,0974	0,1525	0,1844	0,1797	0,1443	0,0961
	8	0,0000.	0,0002	0,0032	0,0166	0,0487	0,0981	0,1489	0,1797	0,1771	0,1442
	9	0,0000	0,0000	0,0007	0,0051	0,0198	0,0514	0,0980	0,1464	0,1771	0,1762
	10	0,0000	0,0000	0,0001	0,0013	0,0066	0,0220	0,0528	0,0976	0,1449	0,1762
	11	0,0000	0,0000	0,0000	0,0003	0,0018	0,0077	0,0233	0,0532	0,0970	0,1442
	12	0,0000	0,0000	0,0000	0,0000	0,0004	0,0022	0,0083	0,0237	0,0529	0,0961
	13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0085	0,0233	0,0518
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0082	0,0222
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0022	0,0074
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,3774	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
	2	0,1887	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
	3	0,0596	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
	4	0,0133	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
	5	0,0022	0,0319	0,1028	0,1746	0,2023	0,17*9	0,1272	0,0746	0,0365	0,0148
	6	0,0003	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
	7	0,0000	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
	8	0,0000	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201
	9	0,0000	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
	10	0,0000	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
	11	0,0000	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
	12	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
	13	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0130	0,0370
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Phụ lục 2: Phân phối poisson p (X; λ)

x	λ									
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1433	0,1647	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0002	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

x	λ									
	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1351
1	0,3662	0,3614	0,3543	0,3452	0,3347	0,3230	0,3106	0,2975	0,2842	0,2707
2	0,2014	0,2169	0,2303	0,2417	0,2510	0,2584	0,2640	0,2678	0,2700	0,2702
3	0,0738	0,0867	0,0998	0,1128	0,1255	0,1378	0,1496	0,1607	0,1710	0,1804
4	0,0203	0,0260	0,0324	0,0395	0,0471	0,0551	0,0636	0,0723	0,0812	0,0902
5	0,0045	0,0062	0,0034	0,0111	0,0141	0,0176	0,0216	0,0260	0,0309	0,0361
6	0,0008	0,0002	0,0018	0,0026	0,0035	0,0047	0,0061	0,0078	0,0098	0,0120
7	0,0001	0,0002	0,0003	0,0005	0,0008	0,0011	0,0015	0,0020	0,0027	0,0034
8	0,0000	0,0000	0,0000	0,0001	0,0001,	0,0002	0,0003	0,0005	0,0006	0,0009
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002

x	λ									
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	0,2572	0,2438	0,2306	0,2177	0,2052	0,1931	0,1815	0,1703	0,1596	0,1494
2	0,2700	0,2681	0,2652	0,2613	0,2565	0,2510	0,2450	0,2384	0,2314	0,2240
3	0,1890	0,1966	0,2033	0,2090	0,2138	0,2176	0,2205	0,2225	0,2237	0,2240
4	0,0992	0,1002	0,1169	0,1254	0,1336	0,1414	0,1408	0,1557	0,1622	0,1680
5	0,0417	0,0476	0,0538	0,0602	0,0668	0,0735	0,0804	0,0872	0,0940	0,1008
6	0,0146	0,0174	0,0206	0,0241	0,0278	0,0319	0,0362	0,0407	0,0455	0,0540
7	0,0044	0,0056	0,0068	0,0083	0,0099	0,0118	0,0139	0,0163	0,0188	0,0216
8	0,0011	0,0015	0,0019	0,0025	0,0031	0,0038	0,0047	0,0057	0,0068	0,0081
9	0,0003	0,0004	0,0005	0,0007	0,0009	0,0011	0,0014	0,0018	0,0022	0,0027
10	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0004	0,0005	0,0006	0,0003
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0002
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

x	λ									
	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,0450	0,0408	0,0369	0,0344	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183
1	0,1397	0,1304	0,1217	0,1135	0,1057	0,0984	0,0915	0,0850	0,0789	0,0733
2	0,2165	0,2087	0,2008	0,1929	0,1050	0,1771	0,1692	0,1615	0,1539	0,1465
3	0,2237	0,2226	0,2209	0,2186	0,2158	0,2125	0,2087	0,2046	0,2001	0,1954
4	0,1734	0,1781	0,1823	0,1850	0,1888	0,1912	0,1931	0,1944	0,1951	0,1954

Phụ lục 2 *(tiếp theo)*

x	λ 3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
5	0,1075	0,1140	0,1203	0,1264	0,1322	0,1377	0,1429	0,1477	0,1522	0,1563
6	0,0555	0,0608	0,0662	0,0716	0,0771	0,0826	0,0881	0,0936	0,0989	0,1042
7	0,0246	0,0278	0,0312	0,0348	0,0385	0,0425	0,0466	0,0508	0,0551	0,0595
8	0,0095	0,0111	0,0129	0,0148	0,0169	0,0191	0,0215	0,0241	0,0269	0,0298
9	0,0033	0,0040	0,0047	0,0056	0,0066	0,0076	0,0089	0,0102	0,0116	0,0132
10	0,0010	0,0013	0,0016	0,0019	0,0023	0,0028	0,0033	0,0039	0,0045	0,0053
11	0,0003	0,0004	0,0005	0,0006	0,0007	0,0009	0,0011	0,0013	0,0016	0,0019
12	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006
13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

x	λ 4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,0166	0,0150	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067
1	0,0679	0,0630	0,0583	0,0540	0,0500	0,0462	0,0427	0,0395	0,0365	0,0337
2	0,1393	0,1323	0,1254	0,1188	0,1125	0,1063	0,1005	0,0948	0,0894	0,0842
3	0,1904	0,1852	0,1790	0,1743	0,1687	0,1631	0,1574	0,1517	0,1460	0,1404
4	0,1951	0,1944	0,1933	0,1917	0,1898	0,1875	0,1849	0,1820	0,1789	0,1755
5	0,1600	0,1633	0,1662	0,1687	0,1708	0,1725	0,1738	0,1747	0,1753	0,1755
6	0,1093	0,1143	0,1191	0,1237	0,1281	0,1323	0,1362	0,1398	0,1432	0,1462
7	0,0640	0,0686	0,0732	0,0778	0,0824	0,0869	0,0914	0,0959	0,1002	0,1044
8	0,0328	0,0360	0,0393	0,0428	0,0463	0,0500	0,0537	0,0575	0,0614	0,0653
9	0,0150	0,0168	0,0188	0,0209	0,0232	0,0255	0,0280	0,0307	0,0334	0,0363
10	0,0061	0,0071	0,0081	0,0092	0,0104	0,0118	0,0132	0,0147	0,0164	0,0181
11	0,0023	0,0027	0,0032	0,0037	0,0043	0,0049	0,0056	0,0064	0,0073	0,0082
12	0,0008	0,0009	0,0011	0,0014	0,0016	0,0019	0,0022	0,0026	0,0030	0,0034
13	0,0002	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013
14	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005
15	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002

x	λ 5,1	5,2	5,3	3,4	5,5	5,6	5,7	5,8	5,9	6,0
0	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037	0,0033	0,0030	0,0027	0,0025
1	0,0311	0,0287	0,0265	0,0244	0,0225	0,0207	0,0191	0,0176	0,0162	0,0149
2	0,0793	0,0746	0,0701	0,0659	0,0618	0,0580	0,0544	0,0509	0,0477	0,0446
3	0,1348	0,1293	0,1239	0,1105	0,1133	0,1082	0,1033	0,0985	0,0938	0,0892
4	0,1719	0,1681	0,1641	0,1600	0,1558	0,1515	0,1472	0,1428	0,1333	0,1339
5	0,1753	0,1748	0,1740	0,1728	0,1714	0,1697	0,1678	0,1656	0,1632	0,1606
6	0,1490	0,1515	0,1537	0,1555	0,1571	0,1584	0,1594	0,1601	0,1605	0,1606
7	0,1086	0,1125	0,1163	0,1200	0,1234	0,1267	0,1298	0,1326	0,1353	0,1377
8	0,0692	0,0731	0,0771	0,0810	0,0849	0,0087	0,0925	0,0962	0,0998	0,1033
9	0,0392	0,0423	0,0454	0,0486	0,0519	0,0552	0,0586	0,0620	0,0654	0,0688

Phụ lục 2 *(tiếp theo)*

x	λ 5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
10	0,0200	0,0220	0,0241	0,0262	0,0285	0,0309	0,0334	0,0359	0.0386	0,0413
11	0,0093	0,0104	0,0116	0,0129	0,0143	0,0157	0,0173	0,0190	0,0207	0,0225
12	0,0039	0,0045	0,0051	0,0058	0.0065	0.0073	0,0082	0,0092	0,0102	0.0113
13	0,0015	0,0018	0.0021	0,0024	0,0025	0.0032	0,0036	0,0041	0,0046	0.0052
14	0.0006	0.0007	0,0008	0,0009	0,0011	0.0013	0.0015	0.0017	0,0019	0,0022
15	0,0002	0,0002	0,0003	0,0003	0,0004	0.0005	0,0006	0.0007	0,0008	0,0009
16	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001

x	λ 6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8	6,9	7,0
0	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0137	0,0126	0,0116	0,0106	0,0098	0,0090	0.0082	0.0076	0,0070	0,0064
2	0,0417	0,0390	0,0364	0,0340	0,0318	0,0296	0,0276	0,0258	0,0240	0.0223
3	0,0848	0,0806	0,0765	0.0726	0,0688	0 0652	0,0617	0,0584	0,0552	0,0521
4	0,1294	0,1249	0,1205	0.1162	0.1118	0.1076	0,1034	0,0992	0,0952	0,0912
5	0,1579	0,1549	0,1519	0,1487	0.1454	0,1420	0,1385	0.1349	0,1314	0,1277
6	0,1605	0,1601	0,1595	0,1586	0,1575	0,1562	0,1546	0.1529	0,1511	0,1490
7	0,1399	0,1418	0,1435	0,1450	0,1462	0,1472	0.1480	0.1486	0,1489	0,1490
8	0.1066	0,1099	0.1130	0,1160	0,1188	0,1215	0,1240	0.1263	0,1284	0,1304
9	0,0723	0,0757	0,0791	0,0825	0,0858	0,0891	0,0923	0,0954	0,0985	0,1014
10	0,0441	0,0469	0.0498	0,0528	0,0558	0,0388	0.0618	0,0649	0,0679	0,0710
11	0,0245	0,0265	0,0285	0,0307	0,0330	0,0353	0,0377	0,0401	0,0426	0,0452
12	0,0124	0.0137	0.0150	0,0164	0,0179	0,0194	0,0210	0,0227	0,0245	0,0264
13	0,0058	0,0065	0.0073	0.0081	0,0089	0,0098	0.0108	0,0119	0,0130	0,0142
14	0,0025	0.0029	0,0033	0.0037	0,0041	0,0046	0,0052	0.0058	0,0064	0,0071
15	0,0010	0,0012	0,0014	0.0016	0,0018	0,0020	0,0023	0,0026	0,0029	0,0033
16	0,0004	0,0005	0,0003	0.0006	0,0007	0,0008	0,0010	0,0011	0,0013	0,0014
17	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006
18	0,0000	0,0001	%0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002
19	0,0000	0,0000	0.0000	0,0000	0,0000	0.0000	0,0000	0,0001	0,0001	0,0001

x	λ 7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
0	0,0008	0.0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0.0004	0,0003
1	0,0059	0,0054	0,0049	0,0045	0,0041	0,0038	0,0035	0.0032	0,0029	0,0027
2	0,0208	0,0194	0,0180	0,0167	0,0156	0,0145	0,0134	0,0125	0.0116	0,0107
3	0,0492	0,0464	0,0438	0,0413	0.0339	0,0366	0.0343	0,0324	0,0305	0,0286
4	0,0874	0.0836	0,0799	0,0764	0.0729	0,0696	0,0663	0,0632	0.0602	0,0573
5	0,1241	0,1204	0.1167	0,1130	0,1094	0.1057	0,1021	0,0986	0,0951	0,0916
6	0,1468	0,1445	0.1420	0,1394	0.1367	0,1339	0,1311	0,1282	0.1252	0.1221
7	0,1489	0,1486	0,1481	0,1474	0,1465	0.1454	0,1442	0,1428	0,1413	0,1396
8	0,1321	0,1337	0,1351	0,1363	0,1373	0,1382	0,1388	0,1392	0,1395	0,1396
9	0,1042	0,1070	0.1096	0,1121	0,1144	0,1167	0,1187	0,1207	0,1224	0,1241
10	0,0740	0,0770	0,0800	0,0829	0,0058	0,0887	0.0914	0,0941	0,0967	0,0993
11	0,0478	0.0504	0,0531	0,0558	0,0585	0,0613	0.0640	0,0667	0,0695	0,0722

Phụ lục 2 *(tiếp theo)*

x	λ 7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
12	0,0283	0,0303	0,0323	0,0344	0,0366	0,0388	0,0411	0,0434	0,0457	0,0481
13	0,0154	0,0168	0,0181	0,0196	0,0211	0,0227	0,0243	0,0260	0,0278	0,0296
14	0,0078	0,0086	0,0095	0,0104	0,0113	0,0723	0,0134	0,0145	0,0157	0,0169
15	0,0037	0,0041	0,0046	0,0051	0,0057	0,0062	0,0069	0,0075	0,0083	0,0090
16	0,0016	0 0019	0,0021	0,0024	0,0026	0,0030	0,0033	0,0037	0,0041	0,0045
17	0,0007	0,0008	0,0009	0,0010	0,0012	0,0013	0,0015	0,0017	0,0039	0,0021
18	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0006	0,0009
19	0,0001	0,0001	0,0001	0;0002	0,0002	0,0002	0,0003	0,0003	0,0003	0,0004
20	0,0000	0,0000	0,0001	0,0001	0,0001	0,0000	0,0001	0,0001	0,0001	0,0002
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0001

x	λ 8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	9,0
0	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
1	0,0025	0,0023	0,0021	0,0019	0,0017	0,0016	0,0014	0,0013	0,0012	0,0011
2	0,0100	0,0092	0,0086	0,0079	0,0074	0,0068	0,0063	0,0058	0,0054	0,0050
3	0,0269	0,0252	0,0237	0,0222	0,0208	0,0195	0,0183	0,0171	0,0160	0,0150
4	0,0544	0,0517	0,0493	0,0466	0,0443	0,0420	0,0398	0,0377	0,0357	0,0337
5	0,0882	0,0849	0,0816	0,0784	0,0752	0,0722	0,0692	0,0663	0,0635	0,0607
6	0,1191	0,1160	0,1128	0,1097	0,1066	0,1034	0,1003	0,0972	0,0941	0,0911
7	0,1378	0,1358	0,1338	0,1317	0,1294	0,1271	0,1247	0,1222	0,1197	0,1171
8	0,1395	0,1392	0,1338	0,1382	0,3375	0,1366	0,1356	0,1344	0,1332	0,1318
9	0,1256	0,1269	0,1280	0,1290	0,1299	0,1306	0,1311	0,1315	0,1311	0,1318
10	0,1017	0,1040	0,1063	0,1084	0,1104	0,1123	0,1140	0,1157	0,1172	0,1186
11	0,0749	0,0776	0,0802	0,0828	0,0853	0,0870	0,0902	0,0925	0,0948	0,0970
12	0,0505	0,0530	0,0555	0,0579	0,0604	0,0629	0,0654	0,0679	0,0703	0,0728
13	0,0335	0,0334	0,0354	0,0374	0,0395	0,0416	0,0438	0,0459	0,0481	0,0504
14	0,0182	0,0196	0,0210	0,0225	0,0240	0,0256	0,0272	0,0289	0,0306	0,0324
15	0,0098	0,0107	0,0116	0,0126	0,0136	0,0147	0,0158	0,0169	0,0182	0,0194
16	0,0050	0,0055	0,0060	0,0066	0,0072	0,0079	0,0086	0,0093	0,0101	0,0109
17	0,0024	0,0026	0,0029	0,0033	0,0036	0,0040	0,0044	0,0048	0,0053	0,0058
18	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019	0,0021	0,0024	0,0026	0,0029
19	0,0005	0,0005	0,0006	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014
20	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0005	0,0006
21	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003
22	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

x	λ 9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005	0,0005
2	0,0046	0,0043	0,0040	0,0037	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
3	0,0140	0,0131	0,0123	0,0115	0,0107	0,0100	0,0093	0,0087	0,0081	0,0076
4	0,0319	0,0302	0,0285	0,0269	0,0254	0,0240	0,0226	0,0213	0,0201	0,0189

Phụ lục 2 *(tiếp theo)*

x	λ									
	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
5	0,0581	0,0555	0,0530	0,0504	0,0483	0,0460	0,0439	0,0410	0,0393	0,0378
6	0,0881	0,0851	0,0322	0,0793	0,0764	0,0736	0,0709	0,0682	0,0656	0,0631
7	0,1145	0,1118	0,1091	0,1064	0,1037	0,1010	0,0932	0,0955	0,0928	0,0901
8	0,1302	0,1236	0,1269	0,1251	0,1232	0,1212	0,1191	0,1170	0,1148	0,1326
9	0,1317	0,1315	0,1311	0,1306	0,1300	0,1293	0,1284	0,1274	0,1263	0,1251
10	0,1198	0,1210	0,1219	0,1228	0,1235	0,1241	0,1245	0,1249	0,1250	0,1251
11	0,0991	0,1012	0,1031	0,1049	0,1067	0,1083	0,1093	0,1112	0,1125	0,1137
12	0,0752	0,0776	0,0799	0,0822	0,0844	0,0866	0,0888	0,0908	0,0928	0,0948
13	0,0526	0,0549	0,0572	0,0594	0,0417	0,0440	0,0662	0,0685	0,0707	0,0729
14	0,0342	0,0361	0,0380	0,0399	0,0419	0,0439	0,0459	0,0479	0,0500	0,0521
15	0,0208	0,0221	0,0235	0,0250	0,0265	0,0281	0,0297	0,0313	0,0330	0,0347
16	0,0118	0,0127	0,0137	0,0147	0,0157	0,0160	0,0180	0,0192	0,0204	0,0217
17	0,0063	0,0069	0,0075	0,0081	0,0088	0,0095	0,0103	0,0111	0,0119	0,0128
18	0,0032	0,0035	0,0039	0,0042	0,0046	0,0051	0,0055	0,0060	0,0065	0,0071
19	0,0015	0,0017	0,0019	0,0021	0,0023	0,0026	0,0028	0,0031	0,0034	0,0037
20	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019
21	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
22	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004
23	0,0000	0,0001	0,0001	0,0001	0,0003	0,0001	0,0001	0,0001	0,0002	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

x	λ									
	11	12	13	14	15	16	17	18	19	20
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0010	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0037	0,0018	0,0008	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
4	0,0102	0,0053	0,0027	0,0013	0,0006	0,0003	0,0001	0,0001	0,0000	0,0000
5	0,0224	0,0127	0,0070	0,0037	0,0019	0,0010	0,0005	0,0002	0,0001	0,0001
6	0,0411	0,0255	0,0152	0,0087	0,0048	0,0026	0,0014	0,0007	0,0004	0,0002
7	0,0646	0,0437	0,0281	0,0174	0,0104	0,0060	0,0034	0,0018	0,0010	0,0005
8	0,0888	0,0655	0,0457	0,0304	0,0194	0,0120	0,0072	0,0042	0,0024	0,0013
9	0,1085	0,0874	0,0661	0,0473	0,0324	0,0213	0,0135	0,0003	0,0050	0,0029
10	0,1194	0,1048	0,0859	0,0663	0,0486	0,0341	0,0230	0,0150	0,0095	0,0058
11	0,1194	0,1144	0,1015	0,0844	0,0663	0,0496	0,0355	0,0245	0,0164	0,0106
12	0,1094	0,1144	0,1099	0,0884	0,0029	0,0661	0,0504	0,0368	0,0259	0,0176
13	0,0926	0,1056	0,1099	0,1060	0,0956	0,0814	0,0658	0,0509	0,0378	0,0271
14	0,0728	0,0905	0,1021	0,1060	0,1024	0,0930	0,0800	0,0655	0,0514	0,0307
15	0,0534	0,0724	0,0885	0,0992	0,1024	0,0985	0,0906	0,0786	0,0650	0,0516
16	0,0367	0,0543	0,0719	0,0866	0,0960	0,0992	0,0963	0,0884	0,0772	0,0646
17	0,0237	0,0383	0,0550	0,0713	0,0847	0,0934	0,0963	0,0936	0,0863	0,0760
18	0,0145	0,0256	0,0397	0,0554	0,0706	0,0030	0,0909	0,0936	0,0911	0,0844
19	0,0084	0,0161	0,0272	0,0409	0,0557	0,0699	0,0814	0,0887	0,0911	0,0888
20	0,0046	0,0097	0,0177	0,0286	0,0418	0,0559	0,0692	0,0798	0,0366	0,0888
21	0,0024	0,0055	0,0109	0,0191	0,0299	0,0426	0,0560	0,0684	0,0783	0,0846
22	0,0012	0,0030	0,0065	0,0121	0,0204	0,0310	0,0433	0,0560	0,0676	0,0769
23	0,0006	0,0016	0,0037	0,0074	0,0133	0,0216	0,0320	0,0438	0,0559	0,0669

Phụ lục 2 *(tiếp theo)*

x	λ									
	11	12	13	14	15	16	17	18	19	20
24	0,0003	0,0008	0,0020	0,0083	0,0033	0,0144	0,0226	0,0328	0,0442	0,0557
25	0,0001	0,0004	0,0010	0,0024	0,0050	0.0092	0,0154	0,0237	0,0336	0.0446
26	0,0000	0.0002	0.0005	0,0013	0,0029	0,0057	0.0101	0.0164	0,0246	0,0343
27	0,0000	0,0001	0.0002	0,0007	0,0016	0,0034	0,0063	0,0109	0.0173	0.0254
28	0.0000	0,0000	0,0001	0.0003	0,0009	0.0019	0,0038	0,0070	0,0117	0,0181
29	0,0000	0,0000	0,0001	0,0002	0,0004	0,0011	0,0023	0.0044	0,0077	0,0125
30	0,0000	0,0000	0,0000	0.0001	0,0002	0,0006	0,0013	0,0026	0,0049	0,0083
31	0,0000	0,0000	0.0000	0,0000	0,0001	0.0003	0,0007	0,0015	0,0030	0,0054
32	0.0000	0.0000	0,0000	0,0000	0,0001	0,0001	0.0004	0,0009	0,0018	0,0034
33	0,0000	0,0000	0.0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0010	0,0020
34	0,0000	0.0000	0,0000	0,0000	0.0000	0,0000	0.0001	0,0002	0,0006	0,0012
35	0,0000	0,0000	0.0000	0,0000	0,0000	0,0000	0.0000	0,0001	0,0003	0,0007
36	0,0000	4,0000	0.0000	0,0000	0,0000	0.0000	0,0000	0,0001	0,0002	0,0004
37	0.0000	0,0000	0.0000	0,0000	0,0000	0.0000	0.0000	0,0000	0,0001	0,0002
31	0,0000	0.0000	0.0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
39	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

Phụ lục 3: Phân phối luỹ thừa

x	e^{-x}	x	e^{-x}	x	e^{-x}
0,0	1,000	1,5	0,223	3,0	0,050
0,1	0,905	1,6	0,202	3,1	0,045
0,2	0.819	1,7	0,183	3,2	0,041
0,3	0,741	1,8	0,165	3,3	0,037
0,4	0,670	1,9	0,150	3,4	0,033
0,5	0,607	2,0	0,135	3,5	0,030
0,6	0,549	2,1	0,122	3,6	0,027
0,7	0,497	2,2	0,111	3,7	0,025
0,8	0,449	2,3	0,100	3,8	0,022
0,9	0,407	2,4	0,091	3,9	0,020
1,0	0,368	2,5	0,082	4,0	0,018
1,1	0,333	2,6	0,074	4,5	0,011
1,2	0,301	2,7	0,067	5,0	0,007
1.3	0,273	2,8	0,061	6,0	0,002
1,4	0,247	2,9	0,055	7,0	0,001

Phụ lục 4: Giá trị hàm $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0303	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Phụ lục 5: Giá trị hàm $\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-z^2/2} dz$

u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0,00	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.10	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.20	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.30	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.40	.1554	.1591	1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.50	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.60	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.70	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.80	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.90	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.00	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.10	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.20	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.30	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.40	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.50	.4332	.4345	.4357	.4370	.4382	.4394	:4406	.4418	.4429	.4441
1.60	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.70	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.80	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.90	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.00	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.10	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.20	.4861	.4864	.4868	.4871	.4875	.4878	.4881	4884	.4887	.4890
2.30	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.40	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.50	.4938	.4940	.4941	.4943	4945	.4946	.4948	.4949	.4951	.4952
2.60	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.70	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.80	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.90	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.00	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

u	Area
3.50	.49976737
4.00	.49996833
4.50	.49999660
5.00	.49999971

Phụ lục 6: Giá trị tới hạn chuẩn

u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.10	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.20	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.30	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.40	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.50	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.60	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.70	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.80	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.90	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.00	.1587	.1562	.1539	.1515	.1492	.1469	1446	.1423	.1401	.1379
1.10	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.20	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.30	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.40	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.50	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.60	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.70	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.80	.0359	.0331	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.90	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.00	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.10	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.20	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.30	.0107	.0104	0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.40	.0082	.0080	.0078	.0073	.0073	.0071	.0069	.0068	.0066	.0064
2.50	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.60	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.70	.0035	.0031	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.80	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.90	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.00	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

u	Area
3.500	.00023263
4.000	.00003167
4.500	.00000340
5.000	.00000029

Phụ lục 7: Giá trị tới hạn χ^2

df	α = .999	α = .995	α = .99	α = .975	α = .95	α = .9
1	.000002	.000039	.000157	.000982	.003932	.01579
2	.002001	.01003	.02010	.05064	.1026	.2107
3	.02430	.07172	.1148	.2158	.3518	.5844
4	.09080	.2070	.2971	.4844	.7107	1.064
5	.2102	.4117	.5543	.8312	1.145	1.610
6	.3811	.6757	.8721	1.237	1.635	2.204
7	.5985	.9893	1.239	1.690	2.167	2.833
8	.8571	1.344	1.646	2.180	2.733	3.490
9	1.152	1.735	2.088	2.700	3.325	4.168
10	1.479	2.156	2.558	3.247	3.940	4.865
11	1.834	2.603	3.053	3.816	4.575	5.578
12	2.214	3.074	3.571	4.404	5.226	6.304
13	2.617	3.565	4.107	5.009	5.892	7.042
14	3.041	4.075	4.660	5.629	6.571	7.790
15	3.483	4.601	5 229	6.262	7.261	8.547
16	3.942	5.142	5.812	6.908	7.962	9.312
17	4.416	5.697	6.408	7.564	8.672	10.09
18	4.905	6.265	7.015	8.231	9 390	10.86
19	5.407	6.844	7,633	8.907	10.12	11.65
20	5.921	7 .434	8.260	9.591	10.85	12.44
21	6.447	8.034	8.897	10.28	11.59	13.24
22	6.983	8.643	9.542	10.98	12.34	14.04
23	7.529	9.260	10.20	11.69	13.09	14.85
24	8.085	9.886	10.86	12.40	1385	15.66
25	8.649	10.52	11.32	13.12	14.61	16.47
26	9.222	11.16	12.20	13.84	15.38	17.29
27	9.803	11.81	12.88	14.57	16.15	18.11
28	10.39	12.46	13.56	15 31	16.93	18.94
29	10.99	13.12	14.26	16.06	17.71	19.77
30	11.59	13.79	14.95	16.79	18.49	20.60
40	17.92	20.71	22.16	24.43	26.51	29.05
50	24.67	27.99	29.71	32.36	34.76	37.69
60	31.74	35.53	37.48	40.48	43.19	46.46
70	39.04	43.28	45.44	48.76	51.74	55.33
80	46.52	51.17	53.54	57.13	60.39	64.28
90	54.16	59.20	61.75	65.65	69.13	73.29
100	61.92	67.33	70.06	74.22	77.93	82.36
120	77.76	83.85	86.92	91.57	95.70	100.62
240	177.95	187.32	191.99	198.98	205.14	212.39

Phụ lục 7 *(tiếp theo)*

α = .1	α = .05	α = .025	α = .01	α = 0.005	α = .001	df
2.706	1.841	5.024	6.635	7.879	10.83	1
4.605	5.991	7.378	9.210	10.60	13.82	2
6.251	7.815	9.348	11.34	12.84	16.27	3
7.779	9.488	11.14	13.28	14.86	18.47	4
9.236	11.07	12.83	15.09	16.75	20.52	5
10.64	12.59	14.45	16.81	18.53	22.46	6
12.02	14.07	16.01	18.48	20.28	24.32	7
13.36	15.51	17.51	20.09	21 95	26.12	8
14.68	16.92	19.02	21.67	23.59	27.88	9
16.99	18.31	20.48	23.21	25.19	29.59	10
17.28	19.68	21.92	24.72	26.76	31.27	11
18.55	21.03	23.34	26.22	38.10	32.91	12
19.81	22.36	24 74	27.69	29.82	34.53	13
21.06	23.68	26.12	29.14	11.32	36.12	14
22.31	25.00	37.49	30.58	32.80	37.70	15
23.54	26.30	28.8.5	32.00	34.27	39.25	16
24.77	27 59	30.19	3341	35.72	40.79	17
25.99	28.87	31.53	34.81	37.16	42.31	18
27.20	30.14	32.85	36.19	38.58	43.82	19
28.41	31.41	34.17	37.57	40.00	45.31	20
29.62	32.67	35.48	38.93	41.40	46.80	21
30.81	33.92	36.78	40.29	42.80	48.27	22
32.01	35.17	38.08	41.64	44.18	49.73	23
33.20	36,42	39.36	42.98	45.56	51.18	24
34.38	37.65	40.65	44.31	46.93	52.62	25
35.56	38.89	41.92	45.64	48.29	54.05	26
36.74	40.11	43.19	46.96	49.65	55.48	27
37.92	41.34	44.46	48.28	50.99	56.89	28
39.09	42.56	45.72	49.59	52.34	58.30	29
40.26	43.77	46.98	50.89	53.67	59.70	30
51.81	55.76	59.34	63.69	66.77	73.40	40
63.17	67.50	71.42	76.15	79.49	86.66	60
74.40	79.08	83.30	88.38	91.95	99.61	60
85.53	90.53	95.02	100.43	104.21	112.32	70
96.58	101.88	106.63	112.33	116.32	124.84	80
107.57	113.15	118.14	124.12	128.30	137.21	90
118.50	124.34	129.56	135.81	140.17	149.45	100
140.23	146.57	152.21	158.95	163.65	173.62	120
268.47	277.14	284.80	293.89	300.18	313.44	240

Phụ lục 8: Giá trị tới hạn student

df	α = .1	α = .05	α = .025	α = .01	α = .005	α = 001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.041	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
240	1.285	1.651	1.970	2.342	2.596	3.125
inf.	1.282	1.645	1.960	2.326	2.576	3.090

Phụ lục 9: Giá trị tới hạn fisher

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32
	.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963 3	968.6
	.01	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.025	38.51	39.00	39.17	3925	39.30	39.33	39.36	39.37	39.39	39.40
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4
	.001	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.4
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
	.01	34.12	30.82	29.46	28.71	28.24	27.91	2767	27.49	27.35	27.23
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43 69
	.001	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129 2
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77
	10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41

Phụ lục 9: *(tiếp theo)*

df.											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
9.41	9.49	9.58	9.63	9.67	9.7	9.76	9.80	9.83	9.85	.25	1
60.71	61.22	61.74	62.00	62.26	62.5	62.79	63.06	63.19	63.33	.10	
243.9	245.9	248.0	249 1	250.1	251.	252.2	253.3	253.8	254.3	.05	
976.7	984.9	993.1	997.2	1001	1006	1010	1014	1016	1018	.025	
6106	6157	6209	6235	6261	6287	6313	6339	6353	6366	.01	
3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.47	3.48	.25	2
9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	9.49	.10	
19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.49	19.50	.05	
39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.49	39.50	.025	
99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	99.50	.01	
199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	199.5	.005	
999.4	999.4	9994	999.5	999.5	999.5	999.5	999.5	999.5	999.5	.001	
2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47	.25	3
5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	.5.14	3.13	.10	
8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.54	8.53	.05	
14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.92	13.90	.025	
27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.17	26.13	.01	
43.39	4308	42.78	42.62	42.47	42.31	42.15	41.99	41.91	41.83	.005	
128.3	1274	126.4	125.9	123.4	125.0	124.5	124.0	123.7	123.5	.001	
2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	.25	4
3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.77	3.76	.10	
5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.64	5.63	.05	
8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.28	8.26	.025	
14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.51	13.46	.01	
20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.40	19.32	.005	
47.41	46.76	46.10	45.77	45.43	45.09	44.75	44.40	44.23	44.05	.001	
1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87	1.87	.25	5
3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11	3.10	.10	
4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.38	4.36	.05	
6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.04	6.02	.025	
9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.07	9.02	.01	
13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.21	12.14	.005	
26.42	25.91	25.39	25.13	24.87	24.60	24.33	24.06	23.92	23.79	.001	
1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74	1.74	.25	6
2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.73	2.72	.10	
4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.69	3.67	.05	
5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.88	4.85	.025	
7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.92	6.88	.01	
10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.94	8.88	.005	
17.99	17.56	17.12	16.90	16.67	16.44	16.21	15.98	15.86	15.75	.001	

Phụ lục 9: *(tiếp theo)*

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89
10	.25	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55
	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
11	.25	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52
	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	.005	12.23	8.91	7.60	6.83	6.42	6.10	5.86	5.68	5.54	5.42
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92
12	.25	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65	1.65	.25	7
2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.48	2.47	.10	
3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.25	3.23	.05	
4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.17	4.14	.025	
6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.69	5.65	.01	
8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7.13	7.08	.005	
13.71	13.32	12.93	12.73	12.53	12.33	12.12	11.91	11.80	11.70	.001	
1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58	1.58	.25	8
2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.30	2.29	.10	
3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.95	2.93	.05	
4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.70	3.67	.025	
5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.90	4.86	.01	
7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	6.01	5.95	.005	
11.19	10.84	10.48	10.30	10.11	9.92	9.73	9.53	9.43	9.33	.001	
1.58	1.57	1.56	1.56	1.55	1.54	1.64	1.53	1 53	1.53	.25	9
2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.17	2.16	.10	
3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.73	2.71	.05	
3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.36	3.33	.025	
5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.35	4.31	.01	
6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.24	5.19	.005	
9.57	9.24	8.90	8.72	8.55	8.37	8.19	8.00	7.91	7.81	.001	
1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.49	1 48	.25	10
2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.07	2.06	.10	
2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.56	2.54	.05	
3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.11	3.08	.025	
4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.95	3.91	.01	
5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.69	4.64	.005	
8.45	8.13	7.80	7.64	7.47	7.30	7.12	6.94	6.85	6.76	.001	
1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45	1.45	.25	11
2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.99	1.97	.10	
2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.43	2.40	.05	
3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.91	2.88	.025	
4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.65	3.60	.01	
5.24	5.05	4.86	4.76	4.65	4.55	4.45	4.34	4.28	4.23	.005	
7.63	7.32	7.01	6.85	6.68	6.52	6.35	6.18	6.09	6.00	.001	
1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.43	1.42	.25	12
2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.92	1.90	.10	
2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.32	2.30	.05	
3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.76	2.72	.025	
4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.41	3.36	.01	
4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.96	3.90	.005	
7.00	6.71	6.40	6.25	6.09	5.93	5.76	5.59	5.51	5.42	.001	

Phụ lục 9: (*tiếp theo*)

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
13	.25	1.45	1.55	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1 48
	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80
14	.25	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	.025	6 30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40
15	.25	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45
	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	.05	4.54	3 68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
16	.25	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	.05	4.49	3.63	3:24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81
17	.25	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43
	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
18	.25	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	.025	5.98	4.56	395	3.61	3.38	3.22	3.10	3.01	2.93	2.87
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40	1.40	.25	13
2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.86	1.85	.10	
2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.23	2.21	.05	
3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.63	2.60	.025	
3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.21	3.17	.01	
4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.70	3.65	.005	
6.52	6.23	5.93	5.78	5.63	5.47	5.30	5.14	5.05	4.97	.001	
1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38	1.38	.25	14
2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.81	1.80	.10	
2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.15	2.13	.05	
3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.52	2.49	.025	
3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.05	3.00	.01	
4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.49	3.44	.005	
6.13	5.85	5.56	5.41	5.25	5.10	4.94	4.77	4.69	4.60	.001	
1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36	1.36	.25	15
2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.77	1.76	.10	
2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.09	2.07	.05	
2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.43	2.40	.025	
3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.91	2.87	.01	
4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.32	3.26	.005	
5.81	5.54	5.25	5.10	4.95	4.80	4.64	4.47	4.39	4.31	.001	
1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.31	1.35	1.34	.25	16
1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.73	1.72	.10	
2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.03	2.01	.05	
2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.35	2.32	.025	
3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.80	2.75	.01	
4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.17	3.11	.005	
5.55	5.27	4.99	4.85	4.70	4.54	4.39	4.23	4.14	4.06	.001	
1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.33	.25	17
1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.70	1.69	.10	
2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.99	1.96	.05	
2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.28	2.25	.025	
3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.70	2.65	.01	
3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	3.04	2.98	.005	
5.32	5.05	4.78	4.13	4.48	4.33	4.18	4.02	3.93	3.85	.001	
1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.32	.25	18
1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.67	1.66	.10	
2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.94	1.92	.05	
2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.22	2.19	.025	
3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.61	2.57	.01	
3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.93	2.87	.005	
5.13	4.87	4.59	4.45	4.30	4.15	4.00	3.84	3.75	3.67	.001	

Phụ lục 9: *(tiếp theo)*

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
19	.25	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41
	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22
20	.25	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08
21	.25	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39
	.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	.025	5.83	4.42	3.82	3.48	3,25	3.09	2.97	2.87	2.80	2.73
	.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95
22	.25	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39
	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
23	.25	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38
	.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
	.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73
24	.25	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38
	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30	.25	19
1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.65	1.63	.10	
2 31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.90	1.88	.05	
2.72	2.62	2.51	2.45	2. 39	2.33	2.27	2.20	2.17	2.13	.025	
3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.54	2.49	.01	
3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.83	2.78	.005	
4.97	4.70	4.43	4.29	4.14	3.99	3.84	3.68	3.60	3.31	.001	
1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30	1.29	.25	20
1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.63	1.61	.10	
2.28	2.20	2.12	2.08	2.04	1 99	1.95	1.90	1.87	1.04	.03	
2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.12	2.09	.025	
3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.47	2.42	.01	
3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.75	2.69	.005	
4.82	4.56	4.29	4.15	4.00	3.86	3.70	3.54	3.46	3.38	.001	
1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.29	1.28	.25	21
1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.60	1.59	.10	
2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.84	1.81	.05	
2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.08	2.04	.025	
3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.41	2.36	.01	
3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.67	2.61	.005	
4.70	4.44	4.17	4.03	3.88	3.74	3.58	3.42	3.34	3.26	.001	
1.37	1.36	1.34	1.33	1 32	1.31	1.30	1.29	1.28	1.28	.25	22
1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.59	1.57	.10	
2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.81	1.78	.05	
2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.04	2.00	.025	
3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.35	2.31	.01	
3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.60	2.55	.005	
4.58	4.33	4.06	3.92	3.78	3.63	3.48	3.32	3.23	3.15	.001	
1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.28	1.27	.25	23
1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.57	1.35	.10	
2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.79	1.76	.05	
2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	2.01	1.97	.025	
3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.31	2.26	.01	
3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.54	2.48	.005	
4.48	4.23	3.96	3.82	3.68	3.53	3.38	3.22	3 14	3.05	.001	
1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.26	.25	24
1.83	1.78	1.73	1.70	1.07	1.64	1.61	1.57	1.35	1.53	.10	
2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.76	1.76	1.73	.05	
2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.97	.1.94	.025	
3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.26	2.21	.01	
3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.49	2.43	.005	
4.39	4.14	3.87	3.74	3 59	3.45	3.29	3.14	3.05	2.97	.001	

Phụ lục 9: *(tiếp theo)*

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
25	.25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37
	.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
	.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
26	.25	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
27	.25	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36
	.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
	.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	.005	9 34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41
28	.25	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
29	.25	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35
	.10	2.89	2.50	2.28	2.15	206	1.99	1.93	1.89	1.86	1.83
	.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
	.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24

Phụ lục 9: *(tiếp theo)*

df_1											
12	15	20	24	30	40	60	120	240	inf.	α	df_2
1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.26	1.25	.25	25
1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.54	1.52	.10	
2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.74	1.71	.05	
2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.94	1.91	.025	
2.99	2.85	2.70	2 62	2.54	2.45	2.36	2.27	2.22	2.17	.01	
3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.44	2.38	.005	
4.31	4.06	3.79	3.66	3.52	3.37	3.22	3.06	2.98	2.89	.001	
1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.26	1.25	.25	26
1.81	1.76	1.71	1 68	1.65	1.61	1.58	1.54	1.52	1.50	10	
2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.72	1.69	.05	
2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.92	1.88	.025	
2.96	2.81	2.66	2.58	2 50	2.42	2.33	2.23	2.18	2.13	.01	
3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.39	2.33	.005	
4.24	3.99	3.72	3.59	3.44	3.30	3.15	2.99	2.90	2.82	.001	
1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.25	1.24	.25	27
1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.51	1.49	.10	
2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.70	1.67	.05	
2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.89	1.85	.025	
2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.15	2.10	.01	
3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.35	2.29	.005	
4.17	3.92	3.66	3.52	3.38	3.23	3.08	2.92	2.84	2.75	001	
1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24	1.24	.25	28
1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.50	1.48	.10	
2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.68	1.65	.05	
2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.87	1.83	.025	
2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.12	2.06	.01	
3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.31	2.25	.005	
4.11	3.86	3.60	3.41	3.32	3.18	3.02	2.86	2.78	2.69	.001	
1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.24	1.23	.25	29
1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.49	1.47	.10	
2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.67	1.64	.05	
2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.85	1.81	.025	
2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.09	2.03	.01	
3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.27	2.21	.005	
4.05	3.80	3.54	3.46	3.27	3.12	2.97	2.81	2.73	2.64	.001	
1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23	1.23	.25	30
1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.48	1.46	.10	
2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.65	1.62	.05	
2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.83	1.79	.025	
2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.06	2.01	.01	
3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.24	2.18	.005	
4.00	3.75	3.49	3.36	3.22	3.07	2.92	2.76	2.68	2.59	.001	

Phụ lục 9: *(tiếp theo)*

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2-89	2.80
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33.	1.32	1.31	1.30
	.10	2.79	2:39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
	01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
90	.25	1.34	1.41	1.39	1.37	1.35	1.33	1.32	1.31	1.30	1.29
	.10	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	.025	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19
	.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
	.005	8.28	5.62	4,57	3.99	3.62	3.35	3.15	3.00	2.87	2.77
	.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	.3.48	3.34
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
	.025	5.15	3.80	3.23.	2.89	2.67	2.52	2.39	2.30	2.22	2.16
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71
	.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24
240	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.30	1.29	1.27	1.27
	.10	2.73	2.32	2.10	1.97	1.87	1.80	1.74	1.70	1.65	1.63
	.05	3.88	3.03	2.64	2.41	2.25	2.14	2.04	1.98	1.92	1.87
	.025	5.09	3.75	3.17	2.84	2.62	2.46	2.34	2.25	2.17	2.10
	.01	6.74	4.69	3.86	3.40	3.09	2.88	2.71	2.59	2.48	2.40
	.005	8.03	5.42	4.38	3.82	3.45	3.19	2.99	2.84	2.71	2.61
	.001	11.10	7.11	5.60	4.78	4.25	3.89	3.62	3.41	3.24	3.09
inf	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52
	.001	10.83	6,91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96

Phụ lục 9: *(tiếp theo)*

df_1											df_2
12	15	20	24	30	40	60	120	240	inf.	α	
1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.20	1.19	.25	40
1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.40	1.38	.10	
2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.54	1.51	.05	
2.29	2.18	2.07	2.01	1.94	I.88	1.80	1.72	1.68	1.64	.025	
2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.86	1.80	.01	
2.95	2,78	2.60	2.50	2.40	2.30	2.18	2.06	2.00	1.93	.005	
3.64	3.40	3.14	3.01	2.87	2.73	2.57	2.41	2.32	2.23	.001	
1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.16	1.15	.25	60
1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.32	1.29	.10	
1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.43	1.39	.05	
2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.53	1.48	.025	
2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.67	1.60	.01	
2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.76	1.69	.005	
3.32	3.08	2.83	2.69	2.55	2.41	2.25	2.08	1.99	1.89	.001	
1.27	1.25	1.23	1.22	1.20	1.19	1.17	1.15	1.13	1.12	.25	90
1.62	1.56	1.50	1.47	1.43	1.39	1.35	1.29	1.26	1.23	.10	
1.86	1.78	1.69	1.64	1.59	1.53	1.46	1.39	1.35	1.30	.05	
2.09	1.98	1.86	1.80	1.73	1.66	1.58	1 48	1.43	1.37	.025	
2.39	2.24	2.09	2.00	1.92	1.82	1.72	1 60	1.53	1.46	.01	
2.61	2.44	2.25	2.15	2.05	1.94	1.82	1.68	1.61	1.52	.005	
3.11	2.88	2.63	2.50	2.36	2.21	2.05	1.87	1.77	1.66	.001	
1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.12	1.10	.25	120
1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.23	1.19	.10	
1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.31	1.25	.05	
2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.38	1.31	.025	
2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.46	1.38	.01	
2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.52	1.43	.005	
3.02	2.78	2.53	2.40	2.26	2.11	1.95	1.77	1.66	1.54	.001	
1.25	1.23	1.21	1.19	1.18	1.16	1.14	1.11	1.09	1.07	.25	240
1.57	1.52	1.45	1.42	1.38	1.33	1.28	1.22	1.18	1.13	.10	
1.79	1.71	1.61	1.56	1.51	1.44	1.37	1.29	1.24	1.17	.05	
2.00	1.89	1.77	1.70	1.63	1.55	1.46	1.35	1.29	1.21	.025	
2.26	2.11	1.96	1.87	1.78	1.68	1.57	1.43	1.35	1.25	.01	
2.45	2.28	2.09	1.99	1.89	1.77	1.64	1.49	1.40	1.28	.005	
2.88	2.65	2.40	2.26	2.12	1.97	1.80	1.61	1.49	1.35	.001	
1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.06	1.00	.25	inf
1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1 12	1.00	.10	
1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.15	1.00	.05	
1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.19	1.00	.025	
2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.22	1.00	.01	
2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.25	1.00	.005	
2.74	2.51	2.27	2.13	1.99	1.84	1.66	1.45	1.31	1.00	.001	

Phụ lục 10: Bảng số ngẫu nhiên 4 chữ số

1559	9068	9290	8303	8508	8954	1051	6677	6415	0342
5550	6245	7313	0117	7652	5069	6354	7668	1096	5780
4735	6214	8037	1385	1882	O828	2957	0530	9210	0177
5333	1313	3063	1134	8676	6241	9960	5304	1582	6198
8495	2956	1121	8484	2920	7934	0670	5263	0968	0069
1947	3353	1197	7363	9003	9313	3434	4261	0066	2714
4785	6325	1868	5020	9100	0823	7379	7391	1250	5501
9972	9163	5833	0100	5758	3696	6496	6297	5653	7782
0472	4629	2007	4464	3312	8728	1193	2497	4219	5339
4727	6994	1175	5622	2341	8562	5192	1471	7206	2027
3658	3226	5981	9025	1080	1437	6721	7331	0792	5383
6906	9758	0244	0259	4609	1269	5957	7556	1975	7898
3793	6916	0132	8873	8987	4975	4814	2098	6683	0901
3376	5966	1614	4025	0721	1537	6695	6090	8083	5450
6126	0224	7169	3596	1593	5097	7286	2686	1796	1150
0466	7566	1320	8777	8470	5448	9575	4669	1402	3905
9908	9832	8185	8835	0384	3699	1272	1181	8627	1968
7594	3636	1224	6808	1184	3404	6752	4391	2016	6167
5715	9301	5847	3524	0077	6674	8061	5438	6508	9673
7932	4739	4567	6797	4540	6488	3639	9777	1621	7244
6311	2025	5250	6099	6718	7539	9681	3204	9637	1091
0476	1624	3470	1600	0675	3261	7749	4195	2660	2150
5317	3903	6098	9438	3482	5505	5167	9993	8191	8488
7474	8876	1918	9828	2061	6664	0391	9170	2776	4025
7460	6800	1967	2758	0737	6880	1500	5763	2061	9373
1002	1494	9972	3877	6104	4006	0477	0669	8557	0513
5449	6891	9047	6297	1075	7762	8091	7153	8881	3367
9453	0809	7151	9982	0411	1120	6129	5090	2053	7570
0471	2725	7588	6573	0546	0110	6132	1224	3124	6563
5469	2668	1996	2249	3857	6637	8010	1701	3141	6147
2782	9603	1877	4159	9809	2570	4544	0544	2660	6737
3129	7217	5020	3788	0853	9465	2186	3945	1696	2286
7092	9885	3714	8557	7804	9524	6226	7774	6674	2775
9566	0501	8352	1062	0634	2401	0379	1697	7153	6208
5863	7000	1714	9276	7218	6922	1032	4838	1954	1680
5881	9151	2321	3147	6755	2510	5759	6947	7102	0097
6416	9939	9569	0439	1705	4680	9881	7071	9596	8758
9568	3012	6316	9065	0710	2158	1639	9149	4848	8634
0452	9538	5730	1893	1186	9245	6558	9562	8534	9321
8762	5920	8989	4777	2169	7873	7082	9495	1594	8600
0194	0270	7601	0342	3897	4133	7650	9228	5558	3597
3306	5478	2797	1605	4996	0023	9780	9429	3937	7573
7198	3079	2171	6972	0926	6599	9328	0597	5948	5753

Phụ lục 10: *(tiếp theo)* Bảng số ngẫu nhiên 4 chữ số

8350	4846	1309	0612	4584	4968	4642	4430	9481	9048
7449	4279	4224	1018	2496	2091	9750	6086	1955	9860
6126	5399	0852	5491	6557	4946	9918	1541	7894	1843
1851	7940	9908	3660	1536	8011	4314	7269	7047	0382
7698	4218	2726	5130	3132	1722	8592	9662	4795	7718
0810	0118	4979	0458	1059	5739	7919	4557	0245	4861
6647	7149	1409	6809	3313	0082	9024	7477	7320	5822
3867	7111	5549	9439	3427	9793	3071	6651	4267	8099
1172	7278	7527	2492	6211	9457	5120	4903	1023	5745
6701	1668	5067	0413	7961	7825	9261	8572	0634	1140
8244	0620	8736	2649	1429	6253	4181	8120	6500	8127
8009	4031	7884	2215	2382	1931	1252	8088	2490	9122
1947	8315	9755	7187	4074	4743	6669	6060	2319	0635
9562	4821	8050	0106	2782	4665	9436	4973	4879	8900
0729	9026	9631	8096	8906	5713	3212	8854	3435	4205
6904	2569	3251	0079	8838	8738	8503	6333	0952	1641

Phụ lục 11: Bảng giá trị kiểm định tổng hạng wilcoxon

a. $\alpha = .025$ một phía $\alpha = .05$ hai phía

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	96	66	114	79	131

b. $\alpha = .05$ một phía $\alpha = .01$ hai phía

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u	T_l	T_u
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

Phụ lục 12: Bảng giá trị kiểm định tổng hạng theo dấu wilcoxon

Một phía	Hai phía	n = 5	n = 6	n = 7	n = 8	n = 9
p = .1	p = .2	2	3	5	8	10
p = .05	p = .1	0	2	3	5	8
p = .025	p = .05		0	2	3	5
p = .01	p = .02			0	1	3
p = .005	p = .01				0	1
p = .0025	p = .005					0
p = .001	p = .002					

Một phía	Hai phía	n = 15	n = 16	n = 17	n = 18	n = 19
p = .1	p = .2	36	42	48	55	62
p = .05	p = .1	30	35	41	47	53
p = .025	p = .05	25	29	34	40	46
p = .01	p = .02	19	23	27	32	37
p = .005	p = .01	15	19	23	27	32
p = .0025	p = .005	12	15	19	23	27
p = .001	p = .002	8	11	14	18	21

Một phía	Hai phía	n = 25	n = 26	n = 27	n = 28	n = 29
p = .1	p = .2	113	124	134	145	157
p = .05	p = .1	100	110	119	130	140
p = .025	p = .05	89	98	107	116	126
p = .01	p = .02	76	84	92	101	110
p = .005	p = .01	68	75	83	91	100
p = .0025	p = .005	60	67	74	82	90
p = .001	p = .002	51	58	64	71	79

Một phía	Hai phía	n = 35	n = 36	n = 37	n = 38	n = 39
p = .1	p = .2	235	250	265	231	297
p = .05	p = .1	213	227	241	256	271
p = .025	p = .05	195	208	221	235	249
p = .01	p = .02	173	185	198	211	224
p = .005	p = .01	159	171	182	194	207
p = .0025	p = .005	146	157	168	180	192
p = .001	p = .002	131	141	151	162	173

Một phía	Hai phía	n = 45	n = 46	n = 47	n = 48	n = 49
p = .1	p = .2	402	422	441	462	482
p = .05	p = .1	371	389	407	426	446
p = .025	p = .05	343	361	378	396	415
p = .01	p = .02	312	328	345	362	379
p = .005	p = .01	291	307	322	339	355
p = .0025	p = .005	272	287	302	318	334
p = .001	p = .002	249	263	277	292	307

Phụ lục 12: *(Tiếp theo)*

Một phía	Hai phía	n = 10	n = 11	n = 12	n = 13	n = 14
p = .1	p = .2	14	17	21	26	31
p = .05	p = .1	10	13	17	21	25
p = .025	p = .05	8	10	13	17	21
p = .01	p = .02	5	7	9	12	15
p = .005	p = .01	3	5	7	9	12
p = .0025	p = .005	1	3	5	7	9
p = .001	p = .002	0	1	2	4	6

Một phía	Hai phía	n = 20	n = 21	n = 22	n = 23	n = 24
p = .1	p = .2	69	77	86	94	104
p = .05	p = .1	60	67	75	83	91
p = .025	p = .05	52	58	65	73	81
p = .01	p = .02	43	49	55	62	69
p = .005	p = .01	37	42	48	54	61
p = .0025	p = .005	32	37	42	48	54
p = .001	p = .002	26	30	35	40	45

Một phía	Hai phía	n = 30	n = 31	n = 32	n = 33	n = 34
p = .1	p = .2	169	181	194	207	221
p = .05	p = .1	151	163	175	187	200
p = .025	p = .05	137	147	159	170	182
p = .01	p = .02	120	130	140	151	162
p = .005	p = .01	109	118	128	138	148
p = .0025	p = .005	98	107	116	126	136
p = .001	p = .002	86	94	103	112	121

Một phía	Hai phía	n = 40	n = 41	n = 42	n = 43	n = 44
p = .1	p = .2	313	330	348	365	384
p = .05	p = .1	286	302	319	336	353
p = .025	p = .05	264	279	294	310	327
p = .01	p = .02	238	252	266	281	296
p = .005	p = .01	220	233	247	261	276
p = .0025	p = .005	204	217	230	244	258
p = .001	p = .002	185	197	209	222	235

Một phía	Hai phía	n = 50	n = 51	n = 52	n = 53	n = 54
p = .1	p = .2	503	525	547	569	592
p = .05	p = .1	466	486	507	529	514
p = .025	p = .05	434	453	473	494	514
p = .01	p = .02	397	416	434	454	473
p = .005	p = .01	373	390	408	427	445
p = .0025	p = .005	350	367	384	402	420
p = .001	p = .002	323	339	355	372	389

Phụ lục 13: Bảng giá trị tới hạn của phân phối cochran

k - số bậc tự do, l - số lượng mẫu; mức ý nghĩa $\alpha = 0,05$

l \ k	1	2	3	4	5	6	7	8	9	10	16	36	144	+∞
2	0.985	0.9750	0.9392	0.9057	0.8772	0.8534	0.8332	0.8159	0.8010	0.7880	0.7341	0.6602	0.5813	0.5000
3	9669	8709	7977	7457	7071	6771	6530	6333	6167	6025	5466	4748	4031	3333
4	9065	7679	6841	6287	5895	5598	5365	5175	5017	4884	4366	3720	3093	2500
5	0.8412	0.6338	0.5981	0.5440	0.5063	0.4783	0.4564	0.4387	0.4241	0.4118	0.3645	0.3066	0.2013	0.2000
6	7808	6161	5321	4803	4447	4184	3980	3817	3682	3568	3135	2612	2119	1667
7	7271	5612	4800	4307	3974	3726	3535	3384	3259	3154	2756	2278	1833	1429
8	0.6798	0.5157	0.4377	0.3910	0.3595	0.3362	0.3185	0.3043	0.2926	0.2829	0.2462	0.2022	0.1616	0.1250
9	6385	4775	4027	3584	3286	3067	2901	2768	2659	2568	2226	1820	1446	1111
10	6020	4450	3733	3311	3029	2823	2666	2541	2439	2353	2032	1655	1308	1000
12	0.5410	0.3924	0.3624	0.2880	0.2624	0.2439	0.2299	0.2187	0.2098	0.2020	0.1737	0.1403	0.1100	0.0833
15	4709	3346	2758	2419	2195	2034	1911	1815	1736	1671	1429	1144	0889	0667
20	3894	2705	2205	1921	1735	1602	1501	1422	1357	1303	1108	0879	0675	0500
24	0.3434	0.2354	0.1907	0.1656	0.1493	0.1374	0.1286	0.1216	0.1160	0.1113	0.0942	0.0743	0.0567	0.0417
30	2929	1980	1593	1377	1237	1137	1061	1002	0958	0921	0771	0604	0457	0333
40	2370	1576	1259	1082	0968	0887	0827	0780	0745	0713	0595	0462	0347	0250
60	0.1737	0.1131	0.0895	0.0765	0.0682	0.0623	0.0583	0.0552	0.0520	0.0497	0.0411	0.0316	0.0234	0167
120	0998	0632	0495	0419	0371	0337	0312	0292	0279	0266	0218	0165	0120	0083
+∞	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000

Phụ lục 13: Bảng giá trị tới hạn của phân phối cochran

k - số bậc tự do; l - số lượng mẫu; mức ý nghĩa $\alpha = 0,01$

l \ k	1	2	3	4	5	6	7	8	9	10	16	36	144	+∞
2	0,9999	0,9950	0,9794	0,9586	0,9373	0,9172	0,8988	0,8823	0,8674	0,8539	0,7949	0,7067	0,6062	0,5000
3	9933	9423	8831	8335	7938	7606	7335	7107	6912	6743	6059	5153	4230	3333
4	9676	8643	7814	7212	6761	6410	6129	5897	5702	5536	4884	4057	3251	2500
5	9279	0,7885	6957	0,6329	0,5875	0,5531	0,5259	0,5037	0,4854	0,4697	0,4094	0,3351	0,2644	0,2000
6	8828	7218	6258	5635	5195	4866	4608	4401	4229	4084	3529	2858	2229	1667
7	8376	6644	5685	5080	4659	4347	4105	3911	3751	3616	3105	2494	1929	1429
8	0,7945	0,6152	0,5209	0,4627	0,4226	0,3932	0,3704	0,3522	0,3373	0,3248	0,2779	0,2214	0,1700	0,1250
9	7544	5727	4810	4251	3870	3592	3378	3207	3007	2950	2514	1992	1521	1111
10	7175	5358	4469	3934	3572	3308	3106	2945	2810	2704	2297	1811	1376	1000
12	0,6528	0,4751	0,3919	0,3428	0,3099	0,2861	0,2680	0,2535	0,2413	0,2320	0,1961	0,1535	0,1157	0,0833
15	5747	4069	3317	2882	2593	2386	2228	2104	2002	1918	1612	1251	0934	0667
20	4799	3297	2654	2288	2048	1877	1748	1646	1567	1501	1248	0960	0709	0500
24	0,4247	0,2871	0,2295	0,1970	0,1759	0,1608	0,1495	0,1406	0,1338	0,1283	0,1060	0,0810	0,0595	0,0417
30	3632	2412	1913	1635	1454	1327	1232	1157	1100	1054	0867	0658	0480	0333
40	2940	1915	1508	1281	1135	1033	0957	0898	0853	0816	0668	0503	0363	0250
60	2151	0,1371	0,1069	0,0902	0,0796	0,0722	0,0668	0,0625	0,0594	0,0567	0,0461	0,0344	0,0245	0,0167
120	1225	0759	0585	0489	0429	0387	0357	0334	0316	0302	0242	0178	0125	0083
+∞	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000	0000

Phụ lục 14: Bảng giá trị tới hạn của phân phối kolmogorov

α	0.5	0.4	0.3	0.2	0,1	0,05	0,02	0.01	0.001
λ_α	0.828	0,895	0.974	1,073	1.224	1.358	1,520	1,627	1,950

Phụ lục 15: Bảng kiểm định Lilliefors

n	*a*				
	0,2	0,15	0,1	0,05	0,01
4	0,300	0,319	0,352	0,381	0,417
5	0,285	0,299	0,315	0,337	0,405
6	0,265	0,277	0,294	0,319	0,364
7	0,247	0,258	0,276	0,300	0,348
8	0,233	0,244	0,261	0,285	0,331
9	0,223	0,233	0,249	0,271	0,311
10	0,215	0,224	0,239	0,258	0,294
11	0,206	0,217	0,230	0,249	0,284
12	1,199	0,212	0,223	0,242	0,275
13	0,190	0,202	0,214	0,234	0,268
14	0,183	0,194	0,207	0,227	0,261
15	0,177	0,187	0,201	0,220	0,257
16	0,173	0,182	0,195	0,213	0,250
17	0,169	0,177	0,189	0,206	0,245
18	0,166	0,173	0,184	0,200	0,239
19	0,163	0,169	0,179	0,195	0,235
20	0,160	0,166	0,174.	0,190	0,231
25	0,142	0,147	0,158	0,173	0,200
30	0,131	0,136	0,144	0,161	0,187
>30	0,736	0,768	0,805	0,886	1,031
	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$

TÀI LIỆU THAM KHẢO

1. Nguyễn Tấn Lập - Giáo trình thống kê toán. ĐH Kinh tế Kế hoạch. Hà Nội. 1972.

2. Lê Văn Phong. Giáo trình lý thuyết xác suất. ĐH Kinh tế Kế hoạch. Hà Nội. 1972.

3. Hoàng Khoan. Giáo trình thống kê toán. ĐH Kinh tế Kế hoạch. Hà Nội. 1979.

4. Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như. Thống kê toán học. NXB Đại học và THCN. Hà Nội. 1984.

5. Tô Cẩm Tú. Phân tích số liệu nhiều chiều. NXB Nông nghiệp. Hà Nội. 1992.

6. Nguyễn Đình Cử, Nguyễn Cao Văn. Giáo trình lí thuyết xác suất và thống kê toán. NXB Thống kê. Hà Nội. 1991.

7. Harald Cramer. Phương pháp toán học trong thống kê. NXB Khoa học và Kỹ thuật. Hà Nội. 1970.

8. Гмурман ВЕ Теория вероятностей и математическая статистика. Москва. 1972.

9. Гнеденко Б. В. Теория вероятностей. Москва 1969.

10. РАО С.Р. Линейные статистические методы и их применения. Москва. 1968.

11. Уйлкс Математическая статистика. Москва 1982.

12. Многомерный статитистический анализ. Москва. 1982.

13. Bencrézi J.P. Analyse des données. Dunod. Paris. 1973.

14. Baillageon G et Rainville J. Introduction à la Statistique appliquée. Québec. 1976.

15. Giard V. Statistique appliquée à la gestion. Economica Paris. 1992.

16. Amir D Aczel. Complete Business Statistics. Irwin Boston. 1993.

17. Alen Webster. Applied Statistics for business and economics. Boston. 1992.

18. Anderson T.W. An introduction to multivariate statistical analysis. New york. 1958.

19. Dillon W and Goldstein M. Multivariate analysis Methods and Applications. New yourk. 1984.

20. De groot M. Probability and Statistics. New york. 1989.

21. Kendall M and Stuart A. The advanced theory of Statistics. California. 1993.

22. OTT L.R. An introduction to statistical methods and data analysis. California. 1993.

23. Scheaffer R. Mendenhall W and OTT L. Elementary survey sampling. California. 1996.

24. Gujarati D. Basic econometrics. McGraw Hill. New york. 1995.

25. Wonnacott T. and Wonnacott R. Introductory Statistics for business and economics. John Willey and sons New york. 1990.

MỤC LỤC

	Trang
LỜI NÓI ĐẦU	3
PHẦN I. LÝ THUYẾT XÁC SUẤT	7
***Chương I.* Biến cố ngẫu nhiên và xác suất**	9
§1. Phép thử và các loại biến cố	9
§2. Xác suất của biến cố	11
§3. Định nghĩa cổ điển về xác suất	12
§4. Định nghĩa thống kê về xác suất	21
§5. Một số định nghĩa khác về xác suất	24
§6. Nguyên lý xác suất lớn và xác suất nhỏ	26
§7. Định lý cộng xác suất	27
§8. Định lý nhân xác suất	35
§9. Các hệ quả của định lý cộng và định lý nhân xác suất	48
9.1. Định lý	48
9.2. Hệ quả	49
9.3. Định lý	50
9.4. Công thức Bernoulli	55
9.5. Công thức xác suất đầy đủ	57
9.6. Công thức Bayes	60
Các ký hiệu và công thức cơ bản	65
Câu hỏi ôn tập	67

Chương II. **Biến ngẫu nhiên và quy luật phân phối xác suất** 71

§1. Định nghĩa và phân loại biến ngẫu nhiên 71

§2. Quy luật phân phối xác suất của biến ngẫu nhiên 73

2.1. Định nghĩa 74

2.2. Bảng phân phối xác suất 74

2.3. Hàm phân bố xác suất 77

2.4. Hàm mật độ xác suất 83

§3. Các tham số đặc trưng của biên ngẫu nhiên 92

3.1. Kỳ vọng toán 93

3.2. Trung vị 104

3.3. Mốt 105

3.4. Phương sai 106

3.5. Độ lệch chuẩn 114

3.6. Hệ số biến thiên 115

3.7. Giá trị tới hạn 115

3.8. Một vài tham số đặc trưng dạng phân phối xác suất 116

Các ký hiệu và công thức cơ bản 117

Câu hỏi ôn tập 119

Chương III. **Một số quy luật phân phối xác suất thông dụng** 123

§1. Quy luật không - một - A(p) 123

§2. Quy luật nhị thức - B(n, p) 125

§3. Quy luật Poisson - P(λ) 134

§4. Quy luật siêu bội - M(N, n) 140

§5. Quy luật phân phối đều - U(a, b) 143

§6. Quy luật phân phối lũy thừa - E(λ) 146

§7. Quy luật phân phối chuẩn - $N(\mu,\sigma^2)$ 151

§8. Quy luật khi bình phương $\chi^2(n)$ 169

§9. Quy luật Student - T(n) 171

§10. Quy luật Fisher - Snedecor - $F(n_1, n_2)$ 173

Các ký hiệu và công thức cơ bản 175

Câu hỏi ôn tập 178

Chương IV. **Biến ngẫu nhiên hai chiều. Hàm các biến ngẫu nhiên** **183**

§1. Khái niệm về biến ngẫu nhiên nhiều chiều 183

§2. Bảng phân phối xác suất của biến ngẫu nhiên hai chiều 184

§3. Hàm phân bố xác suất của biến ngẫu nhiên hai chiều 187

§4. Hàm mật độ xác suất của biến ngẫu nhiên hai chiều 192

§5. Quy luật phân phối xác suất có điều kiện của các thành phần của hệ hai biến ngẫu nhiên 198

§6. Các tham số đặc trưng của hệ hai biến ngẫu nhiên 202

§7. Kỳ vọng toán có điều kiện - Hàm hồi quy 210

§8. Phân phối chuẩn hai chiều 213

§9. Quy luật phân phối xác suất của hàm các biến ngẫu nhiên 214

Các ký hiệu và công thức cơ bản 223

Câu hỏi ôn tập 228

Chương V. Các định lý giới hạn **233**

§1. Bất đẳng thức Trêbưsép 234

§2. Định lý Trêbưsép 236

§3. Định lý Bernoulli 240

§4. Định lý giới hạn trung tâm 242

Các ký hiệu và công thức cơ bản 249

Câu hỏi ôn tập 250

PHẦN II. THỐNG KÊ TOÁN **253**

Chương VI. Cơ sở lý thuyết mẫu **255**

§1. Khái niệm về phương pháp mẫu 255

§2. Tổng thể nghiên cứu 257

2.1. Định nghĩa 257

2.2. Các phương pháp mô tả tổng thể 257

2.3. Các tham số đặc trưng của tổng thể 259

§3. Mẫu ngẫu nhiên 266

3.1. Định nghĩa mẫu ngẫu nhiên 266

3.2. Các phương pháp chọn mẫu 270

3.3. Thang đo các giá trị mẫu 273

3.4. Các phương pháp mô tả số liệu mẫu 275

§4. Thống kê 288

4.1. Định nghĩa 288

4.2. Một số thống kê đặc trưng của mẫu ngẫu nhiên 289

4.3. Đồ thị hình hộp (Box - Plot) 308

§5. Mẫu ngẫu nhiên hai chiều 312
5.1. Khái niệm 312
5.2. Phương pháp mô tả ngẫu nhiên hai chiều 313
5.3. Một số thống kê đặc trưng của mẫu ngẫu nhiên hai chiều 314
§6. Quy luật phân phối xác suất của một số thống kê đặc trưng mẫu 317
6.1. Trường hợp biến ngẫu nhiên gốc phân phối theo quy luật chuẩn 318
6.2. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật chuẩn 320
6.3. Trường hợp biến ngẫu nhiên gốc X phân phối theo quy luật không - một 323
6.4. Trường hợp có hai biến ngẫu nhiên gốc cùng phân phối theo quy luật không - một 325
§7. Suy diễn thống kê 326
7.1. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối chuẩn 327
7.2. Suy diễn về mẫu ngẫu nhiên rút ra từ tổng thể phân phối không - một 331
Các ký hiệu và công thức cơ bản 333
Câu hỏi ôn tập 338
***Chương VII.* Ước lượng các tham số của biến ngẫu nhiên 341**
§1. Phương pháp ước lượng điểm 342
1.1. Phương pháp hàm ước lượng (phương pháp mômen) 342
1.2. Phương pháp ước lượng hợp lý tối đa 351

§2. Phương pháp ước lượng bằng khoảng tin cậy 354
2.1. Khái niệm 354
2.2. Ước lượng kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật chuẩn 356
2.3. Ước lượng hiệu hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn 371
2.4. Ước lượng xác suất p của biến ngẫu nhiên phân phối theo quy luật không - một 379
2.5. Ước lượng hiệu hai tham số p của hai biến ngẫu nhiên phân phối không - một 385
2.6. Ước lượng phương sai của biến ngẫu nhiên phân phối theo quy luật chuẩn 387
2.7. Ước lượng tỷ số của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn 393
2.8. Ước lượng trung vị của tổng thể nghiên cứu 395
Các ký hiệu và công thức cơ bản 399
Câu hỏi ôn tập 402

***Chương VIII.* Kiểm định giả thuyết thống kê 405**

§1. Khái niệm chung 405
§2. Kiểm định tham số 413
2.1. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên phân phối theo quy luật chuẩn khi đã biết phương sai 413
2.2. Kiểm định giả thuyết về kỳ vọng toán của biến ngẫu nhiên phối chuẩn khi chưa biết phương sai 427

2.3. Kiểm định giả thuyết về hai kỳ vọng toán của hai biến ngẫu nhiên phân phối chuẩn 433

2.4. Kiểm định giả thuyết về tham số p của biến ngẫu nhiên phân phối không - một 453

2.5. Kiểm định giả thuyết về hai tham số p của hai biến ngẫu nhiên phân phối chuẩn A(p) 460

2.6. Kiểm định giả thuyết về phương sai của biến ngẫu nhiên phân phối chuẩn 465

2.7. Kiểm định giả thuyết về sự bằng nhau của hai phương sai của hai biến ngẫu nhiên phân phối chuẩn 467

2.8. Kiểm định K phương sai của K biến ngẫu nhiên phân phối chuẩn 470

§3. Kiểm định phi tham số 475

3.1. Kiểm định khi bình phương 476

3.2. Một số kiểm định khác về quy luật phân phối xác suất 494

3.3. Kiểm định theo dấu 504

3.4. Kiểm định tổng hạng Wilcoxon về hai kỳ vọng toán của hai biến ngẫu nhiên 508

3.5. Kiểm định tổng hạng theo dấu của Wilcoxon 515

3.6. Kiểm định Kruskal - Wallis về k kỳ vọng toán 519

3.7. Kiểm định đoạn mạch 524

Các ký hiệu và công thức cơ bản 529

Câu hỏi ôn tập 536

Chương IX. Phân tích phương sai **539**

§1. Đặt vấn đề 539

§2. Mô hình phân tích phương sai một nhân tố 542

§3. Mô hình phân tích phương sai hai nhân tố 548

3.1. Mô hình hai nhân tố tác động riêng rẽ 549

3.2. Mô hình phân tích phương sai hai nhân tố tác động tổng hợp 556

Các ký hiệu và công thức cơ bản 562

Câu hỏi ôn tập 564

Chương X. Phân tích tương quan và hồi quy **565**

§1. Đặt vấn đề 565

§2. Phân tích tương quan 567

2.1. Phân tích tương quan bảng số liệu định lượng 567

2.2. Phân tích tương quan bảng số liệu định tính 576

§3. Phân tích hồi quy 586

3.1. Hàm hồi quy 587

3.2. Mô hình hồi quy tuyến tính đơn 588

3.3. Mô hình hồi quy tuyến tính bội 600

3.4. Một số dạng hàm hồi quy phi tuyến có thể đưa về dạng hàm hồi quy tuyến tính 611

Các ký hiệu và công thức cơ bản 613

Câu hỏi ôn tập 616

Phần phụ lục **617**

Tài liệu tham khảo **654**

BAN BIÊN TẬP - NXB THỐNG KÊ

98 Thụy Khuê - Tây Hồ - Hà Nội

ĐT: 04.8471397, Fax: 04.8457814

Chịu trách nhiệm xuất bản:

CÁT VĂN THÀNH

Biên tập: ĐƯ VINH - NGUYỄN VÂN ANH

Trình bày: LÊ ANH TUẤN - MAI ANH

Sửa bản in: BAN BIÊN TẬP

Sách do Ban Biên tập - NXB Thống kê chế bản và triển khai in.

GIÁO TRÌNH LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ TOÁN

In 5020 cuốn, khổ 14,5 × 20,5cm tại Cty In và Văn hóa phẩm
Số xuất bản: 17-133/XB-QLXB, do Cục Xuất bản, Bộ Văn hóa - Thông tin cấp ngày 13 tháng 02 năm 2004.
In xong, nộp lưu chiểu: tháng 1 năm 2005.

NXB
THỐNG KÊ

GIÁO TRÌNH LÝ THUYẾT XÁC SUẤT & THỐNG