Chương I : Biến cố ngẫu nhiên và xác suất

1.1. Phép thử và các loại biến cố

1.1.1. Phép thử

a) Các thí dụ

+) Muốn biết sản phẩm trong hộp là sản phẩm tốt hay xấu thì ta lấy ra từ hộp một sản phẩm và quan sát xem nó là sản phẩm tốt hay xấu.

v.v.

b) Khái niệm phép thử

Việc thực hiện một nhóm các điều kiện cơ bản để quan sát một hiện tượng nào đó xảy ra hay không xảy ra được gọi là *thực hiện một phép thử*.

Chú ý : Ứng với mỗi phép thử bao giờ cũng gắn với một hành động và một mục đích quan sát.

1.1.2. Biến cố

Khái niệm : Hiện tượng có thể xảy ra hay không xảy ra trong kết quả của một phép thử được gọi là *biến cố*

Thí dụ : Một hộp đựng 10 sản phẩm trong đó có 7 sản phẩm tốt, 3 sản phẩm xấu. Lấy ra một sản phẩm (tức là ta thực hiện một phép thử), gọi A = (Lấy được sản phẩm tốt) thì A là một biến cố.

1.1.3. Phân loại biến cố

+) Biến cố chắc chắn (ký hiệu bằng chữ U): Là biến cố nhất định xảy ra khi thực hiện một phép thử.

+) Biến cố không thể có (ký hiệu bằng chữ V): Là biến cố nhất định không xảy ra khi thực hiện một phép thử.

+) Biến cố ngẫu nhiên (ký hiệu bằng các chữ cái như A, B, C,...): Là biến cố có thể xảy ra khi thực hiện một phép thử.

Thí dụ 1: Tung một đồng xu có 2 mặt Sấp(S) và Ngửa(N). Gọi A = (Đồng xu xuất hiện mặt sấp), ta có A là biến cố ngẫu nhiên.

Thí dụ 2: Gieo một con xúc xắc (giải thích con xúc xắc)

Gọi U = (Con xúc xắc xuất hiện mặt có số chấm $\leq$ 6), ta có U là biến cố chắc chắn.

V = (Con xúc xắc xuất hiện mặt 7 chấm), ta có V là biến cố không thể có.

A_1 = (Con xúc xắc xuất hiện mặt 1 chấm), ta có A_1 là biến cố ngẫu nhiên.

C = (Con xúc xắc xuất hiện mặt có số chấm chẵn), ta có C là biến cố ngẫu nhiên.

Chú ý : Việc đưa biến cố U, V vào chỉ để hoàn thiện về mặt lý thuyết , thực tế ta chỉ quan tâm tới biến cố ngẫu nhiên, từ đây khi nói biến cố ta hiểu đó là biến cố ngẫu nhiên.

1.2. Xác suất của biến cố, định nghĩa cổ điển về xác suất

1.2.1. Khái niệm xác suất của biến cố

Cho A là một biến cố, *xác suất* của biến cố A, ký hiệu P(A) (Probability of event A) là một con số đặc trưng cho khả năng khách quan xuất hiện biến cố A khi thực hiện một phép thử

1.2.2. Định nghĩa cổ điển về xác suất của một biến cố

a) Kết cục duy nhất đồng khả năng có thể xảy ra

Thí dụ 1: Tung một đồng xu cân đối và đồng chất, giả sử khả năng đồng xu xuất hiện mặt sấp hay mặt ngửa là như nhau. Khi đó ta có hai kết cục duy nhất đồng khả năng có thể xảy ra, đó là: {S; N}

Thí dụ 2: Gieo một con xúc xắc cân đối và đồng chất. Gọi A_i = (Con xúc xắc xuất hiện mặt i chấm); $1 \le i \le 6$. Khi đó ta có 6 kết cục duy nhất đồng khả năng có thể xảy ra, đó là {A_1; A_2;.....;A_6}

Thí dụ 3: Một hộp đựng 10 sản phẩm cùng loại, trong đó có 7 chính phẩm và 3 phế phẩm, lấy 1 sản phẩm từ hộp. Khi đó ta có 10 kết cục duy nhất đồng khả năng có thể xảy ra

b) Kết cục thuận lợi cho một biến cố

Thí dụ 1: Trở lại thí dụ 2 gọi C = (Con xúc xắc xuất hiện mặt có số chấm chẵn), khi đó C xảy khi A_2 xảy ra hoặc A_4 xảy ra, hoặc A_6 xảy ra. Do vậy các kết cục {A_2; A_4; A_6} gọi là các kết cục thuận lợi cho biến cố C xảy ra, và ta nói có 3 kết cục thuận lợi cho C

Thí dụ 2: Một hộp đựng 10 sản phẩm cùng loại, trong đó có 7 chính phẩm và 3 phế phẩm, lấy 1 sản phẩm từ hộp, gọi A = (Lấy được chính phẩm) khi đó ta có 7 kết cục thuận lợi cho A

Vậy những kết cục xảy ra làm cho biến cố A xảy ra khi thực hiện một phép thử được gọi là các kết cục thuận lợi cho biến cố A

c) Định nghĩa cổ điển về xác suất

Định nghĩa: Xét một phép thử, gọi *n* là số kết cục duy nhất đồng khả năng có thể xảy ra, gọi *m* là số kết cục thuận lợi cho biến cố A xảy ra, khi đó

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

(P(A) là xác suất xảy ra biến cố A)

Thí dụ 1: Gieo một con xúc xắc cân đối và đồng chất, tính xác suất để con xúc xắc xuất hiện măt có số chấm chẵn

Lời giải: Gọi C = (Con xúc xắc xuất hiện mặt có số chấm chẵn), ta có n = 6, m_C = 3 do đó:

$$P(C) = \frac{3}{6} = 0,5$$

Thí dụ 2: Một hộp đựng 10 quả cầu giống hệt nhau về mặt hình thức, trong đó có 8 quả màu đỏ, 2 quả màu xanh. Lấy ngẫu nhiên 1 quả cầu từ hộp, tính xác suất lấy được quả cầu màu đỏ

Lời giải: Gọi A = (Lấy được quả cầu màu đỏ), ta có n = 10, m_A = 8 do đó

$$P(A) = \frac{8}{10} = 0,8$$

d) Các tính chất của xác suất

+) Nếu A là biến cố ngẫu nhiên thì 0 < P(A) < 1

+) Nếu B là biến cố bất kỳ thì 0 ≤ P(B) ≤ 1

+) Nếu U là biến cố chắc chắn thì P(U) = 1

+) Nếu V là biến cố không thể có thì P(V) = 0

Chú ý : P(A) = 1 nhưng chưa chắc A là biến cố chắc chắn
P(B) = 0 nhưng chưa chắc B là biến cố không thể có

Thí dụ :

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

1.3.1. Phương pháp suy luận trực tiếp

Thí dụ 1: (xem thí dụ trong giáo trình)

Tính xác suất bằng cách vẽ hình (biểu đồ Ven, hình cây). Tính xác suất bằng cách liệt kê tất cả các giá trị có thể có khi thực hiện một phép thử, và đếm các kết cục thuận lợi cho một biến cố, sau đó áp dụng công thức tính xác suất bằng định nghĩa cổ điển

Thí dụ 2: Tung 3 đồng xu giống nhau và mỗi đồng xu cân đối và đồng chất, tính xác suất để có 2 đồng xu xuất hiện mặt ngửa

Lời giải : Gọi A = (Có 2 đồng xu xuất hiện mặt ngửa)

Những khả năng có thể xảy ra khi tung đồng thời 3đồng xu là

{NNN, NNS, NSN, NSS, SNN, SSN, SNS, SSS}

ta thấy n = 8, m_A = 3 do vậy $P(A) = \frac{3}{8}$

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

(Nhắc lại ý nghĩa và phương pháp tính các công thức $n!$, $C_n^k, A_n^k, \overline{A_n^k}$)

Thí dụ 1: Một hộp đựng 10 quả cầu có kích thước giống nhau trong đó có 6 quả màu xanh, 4 quả màu đỏ. Lấy ngẫu nhiên từ hộp 3 quả cầu, tính xác suất để

a) Lấy được cả 3 quả màu xanh

b) Lấy được đúng 2 quả màu đỏ

Lời giải :

Ta có số kết cục duy nhất đồng khả năng có thể xảy ra là $n = C_{10}^3$

a) Gọi A = (Lấy được 3 quả màu xanh), ta có $m_A = C_6^3$

do vậy $$P(A) = \frac{C_6^3}{C_{10}^3} = \frac{20}{120} = \frac{1}{6}$$

b) Gọi B = (Lấy được đúng 2 quả màu đỏ), ta có $m_B = C_6^1.C_4^2$

do vậy $$P(B) = \frac{C_6^1.C_4^2}{C_{10}^3} = \frac{36}{120} = 0,3$$

Thí dụ 2: Một công ty cần tuyển 5 người. Có 20 người nộp đơn trong đó có 8 nam và 12 nữ. Giả sử khả năng trúng tuyển của 20 người là như nhau, tính xác suất để

a) Có 2 nam trúng tuyển

b) Có ít nhất 3 nữ trúng tuyển

Lời giải: Số khả năng có thể xảy ra là $n = C_{20}^{5} = 15504$.

a) Gọi A = (có 2 nam trúng tuyển); có $m_A = C_8^2.C_{12}^3 = 6160$

do vậy ta có $$P(A) = \frac{C_8^2.C_{12}^3}{C_{20}^5} = \frac{6160}{15504} = 0,3973$$

b) Gọi B = (có ít nhất 3 nữ trúng tuyển); có $m_B = C_{12}^3.C_8^2 + C_{12}^4.C_8^1 + C_{12}^5 = 10912$

do vậy ta có $$P(B) = \frac{10912}{15504} = 0,70382$$

1.3.3. Ưu điểm và hạn chế của phương pháp cổ điển

*) Ưu điểm :

+) Không cần thực hiện phép thử, phép thử chỉ tiến hành một cách giả định

+) Cho phép tìm được một cách chính xác giá trị của xác suất

*) Hạn chế :

+) Số kết cục duy nhất đồng khả năng phải hữu hạn nhưng trong thực tế có nhiều phép thử mà số kết cục có thể là vô hạn

+) Tính đối xứng hay tính đồng khả năng thực sự hiếm gặp trong thực tế

1.4. Định nghĩa xác suất bằng tần suất

1.4.1. Tần suất xuất hiện biến cố

Ta biết rằng với mỗi phép thử thì ta có hoặc biến cố A (mà ta quan tâm) xuất hiện hoặc không xuất hiện. Giả sử ta thực hiện *n* phép thử độc lập, trong *n* phép thử đó biến cố A xuất hiện *k* lần khi đó *tần suất xuất hiện biến cố A* ký hiệu là $f(A)$ được xác định:

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

Thí dụ : Kiểm tra ngẫu nhiên 100 sản phẩm do một máy sản xuất người ta phát hiện ra 3 phế phẩm. Gọi A là biến cố (lấy được một phế phẩm) trong 100 sản phẩm khi đó $f(A) = \frac{3}{100} = 0,03$

1.4.2. Định nghĩa xác suất bằng tần suất

Khi số phép thử *n* tăng lên khá lớn (tùy thuộc tình huống thực tế) thì ta định nghĩa xác suất để biến cố A xảy ra là $P(A) = f(A)$

1.4.3. Ưu điểm và hạn chế của phương pháp tần suất

*) Ưu điểm : Không đòi hỏi các điều kiện áp dụng như đối với định nghĩa cổ điển

*) Hạn chế : Phải thực hiện phép thử với số lần khá lớn dẫn đến tốn kém mất nhiều thời gian.

1.5. Nguyên lý xác suất lớn nguyên lý xác suất nhỏ

*) Nguyên lý xác suất lớn : Biến cố A được coi là xảy ra trong một phép thử thì thực tế $P(A) \geq 1 - \alpha$, với α là xác suất nhỏ tùy thuộc vào tình huống thực tế.

Thí dụ :

*) Nguyên lý xác suất nhỏ : Biến cố B được coi là không xảy ra trong một phép thử thì thực tế $P(B) < \alpha$, với α là xác suất nhỏ tùy thuộc vào tình huống thực tế

Thí dụ :

1.6. Mối quan hệ giữa các biến cố

1.6.1. Tổng các biến cố

a) Tổng hai biến cố : Biến cố C được gọi là tổng của hai biến cố A và B, ký hiệu là C = A + B, khi đó biến cố C xảy ra nếu có ít nhất một trong hai biến cố A và B xảy ra

Thí dụ : Hai người cùng bắn vào bia một viên đạn, gọi A = (Người thứ nhất bắn trúng bia), gọi B = (Người thứ hai bắn trúng bia), C = (Bia bị trúng đạn). Khi đó

$$C = A + B$$

+) Mở rộng : Cho $A_1, A_2, ..., A_n$ là các biến cố, đặt biến cố $A = \sum_{i=1}^{n} A_i$, khi đó biến cố A xảy ra nếu có ít nhất một trong các biến cố $A_1, A_2, ..., A_n$ xảy ra

b) Hai biến cố xung khắc : Hai biến cố A và B được gọi là *xung khắc* với nhau nếu chúng không cùng xảy ra trong một phép thử. Trong trường hợp chúng có thể cùng xảy ra trong một phép thử thì gọi là hai biến cố không xung khắc.

Thí dụ 1 : Gieo một con xúc xắc, gọi A_1 = (Con xúc xắc xuất hiện mặt một chấm); A_2 = (Con xúc xắc xuất hiện mặt hai chấm), khi đó A_1, A_2 là hai biến cố xung khắc

Thí dụ 2 : Hai người cùng bắn một viên đạn vào bia, gọi B_1 = (Người thứ nhất bắn trúng bia); B_2 = (Người thứ hai bắn trúng bia), khi đó B_1, B_2 là hai biến cố không xung khắc

+) Mở rộng : Nhóm các biến cố $A_1; A_2; ...; A_n$ được gọi là *xung khắc với nhau từng đôi* nếu bất kỳ hai biến cố trong nhóm trên xung khắc với nhau

c) Nhóm đầy đủ các biến cố : Các biến cố H_1; H_2;.....; H_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. Hay nói khác đi các biến cố H_1; H_2;.....; H_n tạo thành một nhóm đầy đủ các biến cố nếu chúng đôi một xung khắc và $\sum_{i=1}^{n} H_i = U$

Thí dụ : Gieo một con xúc xắc cân đối và đồng chất, gọi A_i = (Con xúc xắc xuất hiện mặt *i* chấm), $1 \leq i \leq 6$ khi đó các biến cố A_1; A_2;.....; A_6 tạo thành một nhóm đầy đủ các biến cố

Nếu gọi H_C = (Con xúc xắc xuất hiện mặt có số chấm chẵn); H_L = (Con xúc xắc xuất hiện mặt có số chấm lẻ) thì các biến cố H_C, H_L cũng tạo thành một nhóm đầy đủ các biến cố

Chú ý: Với một phép thử có thể có nhiều nhóm đầy đủ

d) Hai biến cố đối lập : Hai biến cố A và $\overline{A}$ gọi là *đối lập* với nhau nếu chúng tạo thành một nhóm đầy đủ các biến cố

Thí dụ 1 : Bắn một viên đạn vào bia, gọi A = (Viên đạn trúng bia) và $\overline{A}$ = (Viên đạn không trúng bia) khi đó A và $\overline{A}$ là hai biến cố đối lập

Thí dụ 2 : Một hộp đựng 10 sản phẩm trong đó có 6 chính phẩm và 4 phế phẩm. Lấy ra 3 sản phẩm, gọi B = (Lấy được ít nhất một chính phẩm) và $\overline{B}$ = (Lấy được cả 3 phế phẩm) khi đó B và $\overline{B}$ là hai biến cố đối lập

1.6.2. Tích các biến cố

a) Tích hai biến cố : Biến cố C được gọi là tích của hai biến cố A và B, ký hiệu là C = A.B, khi đó biến cố C xảy ra khi đồng thời cả hai biến cố A và B xảy ra

Thí dụ : Hai người cùng bắn vào bia một viên đạn, gọi A = (Người thứ nhất bắn trúng bia), B = (Người thứ hai bắn trúng bia), gọi C = (Bia bị trúng 2 viên đạn) thì C = A.B

+) Mở rộng : Cho $A_1, A_2, ..., A_n$ là các biến cố, đặt biến cố $A = \prod_{i=1}^{n} A_i$, biến cố A xảy ra khi tất cả các biến cố $A_1, A_2, ..., A_n$ cùng xảy ra

b) Hai biến cố độc lập : Hai biến cố A và B được gọi là độc lập với nhau nếu việc xảy ra hay không xảy ra của biến cố A không làm thay đổi xác suất xảy ra của biến cố B và ngược lại. Trong trường hợp biến cố A xảy ra hay không xảy ra có làm thay đổi xác suất xảy ra của biến cố B thì A và B là hai biến cố phụ thuộc

Thí dụ : Một hộp đựng 10 sản phẩm trong đó có 7 chính phẩm và 3 phế phẩm, người ta lần lượt lấy ra 2 sản phẩm theo hai phương thức, thứ nhất có hoàn lại và thứ hai không hoàn lại. Gọi A = (Lấy được chính phẩm ở lần thứ nhất), B = (Lấy được chính phẩm ở lần thứ hai). Hỏi lấy theo phương thức nào hai biến cố A và B độc lập

Lời giải : Lấy theo phương thức thứ nhất

+) Mở rộng :

-) Các biến cố $A_1, A_2, ..., A_n$ được gọi là độc lập từng đôi với nhau nếu hai biến cố bất kỳ trong *n* biến cố trên độc lập với nhau

-) Các biến cố $A_1, A_2, ..., A_n$ được gọi là độc lập toàn phần với nhau nếu mỗi biến cố bất kỳ trong *n* biến cố trên độc lập với một tổ hợp bất kỳ của các biến cố còn lại

Thí dụ : Tung một đồng xu 3 lần, gọi A_i = (Đồng xu xuất hiện mặt ngửa ở lần tung thứ *i*), $i = \overline{1;3}$ khi đó các biến cố A_1; A_2; A_3 độc lập với nhau từng đôi.

1.7. Các định lý và công thức xác suất

1.7.1. Định lý cộng xác suất

+) Nếu A và B là hai biến cố xung khắc thì

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

+) Nếu A và B là hai biến cố không xung khắc thì

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

+) Nếu các biến cố $A_1, A_2, ..., A_n$ xung khắc với nhau từng đôi thì

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

+) Nếu các biến cố $H_1, H_2, ..., H_n$ tạo thành một nhóm đầy đủ các biến cố thì

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

+) Nếu A và $\overline{A}$ là hai biến cố đối lập thì $P(A) + P(\overline{A}) = 1$

+) Nếu A_1, A_2, A_3 là ba biến cố không xung khắc thì

$P(A_1+A_2+A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1A_2)-P(A_2A_3)-P(A_3A_1) + P(A_1A_2A_3)$

+) Nếu $A_1, A_2, ..., A_n$ là các biến cố không xung khắc và độc lập toàn phần với nhau thì $P\left(\sum_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} P\left(\overline{A}_i\right)$

+) Một số công thức khác

Cho A và B là hai biến cố, khi ấy ta có

- $AB + A\overline{B} + \overline{A}B + \overline{A}\overline{B} = U$
- $A + B = AB + A\overline{B} + \overline{A}B$
- $\overline{A} + \overline{B} = A\overline{B} + \overline{A}B + \overline{A}\overline{B}$
- $\overline{A}\overline{B} = \overline{A + B}$
- $\overline{AB} = \overline{A} + \overline{B}$

1.7.2. Xác suất có điều kiện, định lý nhân xác suất

a) Xác suất có điều kiện

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

Thí dụ : Một hộp đựng 10 sản phẩm trong đó có 6 chính phẩm và 4 phế phẩm, lấy ra lần lượt hai sản phẩm. Tính xác suất để lần thứ hai lấy được chính phẩm biết rằng lần thứ nhất lấy được phế phẩm.

Lời giải : Gọi A = (Lấy được chính phẩm ở lần thứ hai), B = (Lấy được phế phẩm ở lần thứ nhất). Theo đầu bài ta có biến cố B đã xảy ra với P(B) = 0,4 do vậy

$$P(A / B) = \frac{6}{9} = \frac{2}{3}$$

b) Tính chất

Nếu A và B là hai biến cố độc lập thì P(A/B) = P(A) và P(B/A) = P(B).

c) Định lý nhân xác suất

+) Nếu A và B là hai biến cố độc lập thì điều kiện cần và đủ là

$$P(AB) = P(A)P(B)$$

+) Nếu $A_1, A_2, ..., A_n$ là các biến cố độc lập toàn phần thì

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

+) Cho A và B là hai biến cố ta có

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

$$P(A/B) = \frac{P(AB)}{P(B)} \quad \text{với} \quad P(B) > 0$$

$$P(B/A) = \frac{P(AB)}{P(A)} \quad \text{với} \quad P(A) > 0$$

+) Nếu $A_1, A_2, ..., A_n$ là *n* biến cố phụ thuộc thì ta có công thức

$$P(A_1.A_2.....A_n) = P(A_1).P(A_2/A_1).P(A_3/A_1A_2).....P(A_n/A_1A_2...A_{n-1})$$

1.7.3. Công thức Bernoulli

a) Công thức Bernoulli : Giả sử ta thực hiện *n* phép thử độc lập, với mỗi phép thử chỉ có 2 trường hợp hoặc biến cố A xảy ra với P(A) = *p* hoặc biến cố $\overline{A}$ xảy ra với $P(\overline{A}) = 1 - p$. Gọi B = (Trong *n* phép thử độc lập nói trên biến cố A xuất hiện *k* lần), $0 \le k \le n$. Khi đó ta có

$$P(B) = P_n(k) = C_n^k p^k (1-p)^{n-k} \quad \text{(công thức Bernoulli).}$$

b) Thí dụ : Một xạ thủ có xác suất bắn trúng vòng mười là 0,8 cho mỗi lần bắn. Anh ta được phát 5 viên đạn để lần lượt bắn vào bia, gọi B = (Anh ta bắn trúng vòng mười 3 viên đạn trong 5 viên được phát). Tính P(B) = ?

Lời giải : Áp dụng công thức Bernoulli với $p = 0,8$ $n = 5$ $k = 3$ ta có

$$P(B) = C_5^3 0,8^3 0,2^2 = 0,2048$$

1.7.4. Công thức xác suất đầy đủ, công thức Bayes

a) Công thức xác suất đầy đủ

Giả sử các biến cố $H_1, H_2, ...,H_n$ tạo thành một nhóm đầy đủ các biến cố, nếu biến cố A xảy ra đồng thời với một trong các biến cố $H_1, H_2, ...,H_n$ thì ta có công thức

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i) \quad \text{(Công thức xác suất đầy đủ)}$$

Thí dụ 1 : Một nhà máy có hai dây chuyền sản xuất A và B, dây chuyền A sản xuất ra 60% số sản phẩm của nhà máy, dây chuyền B sản xuất ra 40% số sản phẩm của nhà máy. Biết rằng tỉ lệ phế phẩm do dây chuyền A sản xuất là 1,5% và tỉ lệ phế phẩm do dây chuyền B sản xuất là 2%. Lấy ngẫu nhiên một sản phẩm từ nhà máy, tính xác suất lấy được chính phẩm.

Lời giải : Gọi H_1 = (Lấy được sản phẩm do dây chuyền A sản xuất)

H_2 = (Lấy được sản phẩm do dây chuyền B sản xuất

A =(Lấy được chính phẩm của nhà máy)=> $\overline{A}$ = (Lấy được phế phẩm của nhà máy)

Theo giả thiết : $P(H_1) = 0,6$ $P(H_2) = 0,4$

$P(\overline{A}/H_1) = 0,015$; $P(\overline{A}/H_2) = 0,02$

=> $P(A/H_1) = 0,985$; $P(A/H_2) = 0,98$

Áp dụng công thức xác suất đầy đủ ta có

$$P(A) = P(H_1).P(A/H_1) + P(H_2).P(A/H_2)$$
$$= 0,6.0,985 + 0,4.0,98 = 0,983$$

Thí dụ 2 : Có hai hộp sản phẩm giống nhau, hộp thứ nhất đựng 10 sản phẩm trong đó có 8 chính phẩm và 2 phế phẩm, hộp thứ hai đựng 10 sản phẩm trong đó có 7 chính phẩm và 3 phế phẩm. Người ta chuyển 1 sản phẩm từ hộp thứ nhất sang hộp thứ hai sau đó lấy từ hộp hai ra 2 sản phẩm, tính xác suất lấy được 1 chính phẩm và 1 phế phẩm từ hộp thứ hai

Lời giải : Gọi H_1 = (Chuyển 1 chính phẩm từ hộp 1 sang hộp 2);

H_2 = (Chuyển 1 phế phẩm từ hộp 1 sang hộp 2)

A = (Lấy được 1 chính phẩm và 1 phế phẩm từ hộp 2)

Ta có : $P(H_1) = 0,8$ $P(H_2) = 0,2$ $P(A/H_1) = \frac{C_8^1.C_3^1}{C_{10}^2} = \frac{24}{45}$; $P(A/H_2) = \frac{C_7^1.C_4^1}{C_{10}^2} = \frac{28}{45}$

$P(A) = P(H_1).P(A/H_1) + P(H_2).P(A/H_2) = \frac{8}{10}.\frac{24}{45} + \frac{2}{10}.\frac{28}{45} = \frac{248}{450} = 0,551111$

b) Công thức Bayes

Giả sử các biến cố $H_1, H_2,.....,H_n$ tạo thành một nhóm đầy đủ các biến cố, nếu biến cố A xảy ra đồng thời với một trong các biến cố $H_1, H_2,.....,H_n$ thì ta có công thức

$$P(H_j/A) = \frac{P(H_j)P(A/H_j)}{\sum_{i=1}^{n} P(H_i)P(A/H_i)} \qquad j = \overline{1;n}$$

hay $$P(H_j/A) = \frac{P(H_j)P(A/H_j)}{P(A)} \;; \qquad j = \overline{1;n}$$

Thí dụ 1 : Có hai hộp sản phẩm giống hệt nhau, hộp I đựng 20 sản phẩm trong đó có 16 chính phẩm và 4 phế phẩm, hộp II đựng 20 sản phẩm trong đó có 18 chính phẩm và 2 phế phẩm. Người ta lấy ngẫu nhiên 1 hộp rồi từ hộp đó người ta lấy ngẫu nhiên ra 1 sản phẩm thì thấy nó là chính phẩm, tính xác suất để sản phẩm lấy ra là của hộp I

Lời giải : Gọi H_1 = (Lấy được hộp I); H_2 = (Lấy được hộp II); A = (Lấy được chính phẩm). Ta có $P(H_1) = P(H_2) = 0,5$ $P(A/H_1) = 0,8$ $P(A/H_2) = 0,9$

Áp dụng công thức xác suất đầy đủ ta có

$$P(A) = P(H_1).P(A/H_1) + P(H_2).P(A/H_2)$$
$$= 0,5(0,8 + 0,9) = 0,85$$

Áp dụng công thức Bayes ta có

$$P(H_1/A) = \frac{P(H_1)P(A/H_1)}{P(A)} = \frac{0,5.0,8}{0,85} = 0,47058824$$

Chú ý :

+) Các xác suất $P(H_1)$, $P(H_2)$,...., $P(H_n)$ gọi là các xác suất *tiên nghiệm* và các xác suất $P(H_1/A)$, $P(H_2/A)$,....., $P(H_n/A)$ gọi là các xác suất *hậu nghiệm*

+) Nhóm các biến cố (H_1/A), (H_2/A),....., (H_n/A) cũng tạo thành một nhóm đầy đủ các biến cố.

Thí dụ 2 : (Bài tập 1.64 sách bài tập xác suất và thống kê toán, đã có lời giải).

Chương II : Biến ngẫu nhiên và quy luật phân phối xác suất

2.1. Biến ngẫu nhiên

2.1.1. Định nghĩa biến ngẫu nhiên : Một biến số được gọi là ngẫu nhiên nếu trong kết quả của một 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.

+) Biến ngẫu nhiên thường ký hiệu bằng các chữ cái in hoa như X, Y, Z,....

+) Các giá trị có thể có của một biến ngẫu nhiên thường ký hiệu bằng các chữ thường như *x*, *y*, *z*,....

+) Biến ngẫu nhiên X nhận giá trị *x* ký hiệu là (X = *x*) thì thực chất đây là một biến cố ngẫu nhiên

Thí dụ : Gieo một con xúc xắc, nếu gọi A_1 = (Con xúc xắc xuất hiện mặt 1 chấm) thì A_1 là một biến cố ngẫu nhiên, nhưng nếu gọi X = (Số chấm xuất hiện) thì X là 1 biến ngẫu nhiên và $(X = 1) \equiv A_1$

2.1.2. Phân loại biến ngẫu nhiên

a) Biến ngẫu nhiên rời rạc : là biến ngẫu nhiên mà các giá trị có thể có của nó lập nên một tập hợp hữu hạn hoặc đếm được các phần tử

Hay nói cách khác : Biến ngẫu nhiên rời rạc là ta có thể liệt kê được tất cả các giá trị có thể có của nó

Thí dụ :

+) Gieo một con xúc xắc, gọi X = (Số chấm xuất hiện) khi đó X là biến ngẫu nhiên rời rạc và các giá trị có thể có của X là {1,2,...,6}

+) Một hộp đựng 10 viên bi trong đó có 6 bi đỏ và 4 bi trắng, lấy ngẫu nhiên từ hộp 3 viên bi. Gọi Y = (Số bi đỏ lấy được) khi đó Y là biến ngẫu nhiên rời rạc và các giá trị có thể có của Y là {0,1,2,3}

+) Một xạ thủ có xác suất bắn trúng bia cho mỗi lần bắn là 0,8 anh ta được phát từng viên đạn để lần lượt bắn vào bia cho đến khi anh ta bắn trúng bia thì dừng. Gọi Z = (Số viên đạn xạ thủ được nhận) khi đó Z là biến ngẫu nhiên rời rạc và các giá trị có thể có của Z là $\{1,2,....,n,......\}$ ($n \in N$).

b) Biến ngẫu nhiên liên tục : là biến ngẫu nhiên mà các giá trị có thể có của nó lấp đầy một khoảng trên trục số

Thí dụ :

+) Gọi X_1 = (Năng suất lúa vụ mùa của một tỉnh) thì X_1 là biến ngẫu nhiên liên tục

+) Gọi X_2 = (Chiều cao của thanh niên Việt nam tuổi từ 18 đến 22) thì X_2 là biến ngẫu nhiên liên tục

+) Gọi X_3 = (Giá của một loại cổ phiếu trong phiên giao dịch tháng tới) thì X_3 là biên ngẫu nhiên liên tục

2.2. Quy luật phân phối xác suất của biến ngẫu nhiên

2.2.1. Định nghĩa : Quy luật phân phối xác suất của một biến ngẫu nhiên là sự tương ứng giữa các giá trị có thể có của biến ngẫu nhiên với các xác suất tương ứng để biến ngẫu nhiên nhận các giá trị đó.

+) Để mô tả quy luật phân phối xác suất người ta thường dùng

-) Bảng phân phối xác suất (đối với biến ngẫu nhiên rời rạc)

-) Hàm phân bố xác suất

-) Hàm mật độ xác suất (đối với biến ngẫu nhiên liên tục)

2.2.2. Bảng phân phối xác suất của biến ngẫu nhiên rời rạc X

Giả sử biến ngẫu nhiên rời rạc X nhận một trong các giá trị $x_1, x_2, \ldots, x_n$ với các xác suất tương ứng là

$$p_1 = P(X = x_1),\ p_2 = P(X = x_2), \ldots, p_n = P(X = x_n)$$

Khi đó ta có bảng phân phối xác suất của X như sau

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

Bảng phân phối xác suất trên trở thành quy luật phân phối xác suất nếu các xác suất p_i (với $i=\overline{1;n}$) thỏa mãn

$$\begin{cases} 0 \le p_i \le 1 \\ \sum_{i=1}^{n} p_i = 1 \end{cases}$$

Thí dụ 1: Gieo một con xúc xắc cân đối và đồng chất, gọi X là số chấm xuất hiện. Khi đó ta có quy luật phân phối xác suất của X là

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

Thí dụ 2 : Một hộp đựng 10 sản phẩm trong đó có 7 chính phẩm và 3 phế phẩm, lấy ngẫu nhiên ra 2 sản phẩm từ hộp. Gọi Y là số chính phẩm lấy được khi đó bảng quy luật phân phối xác suất của Y là

Y	0	1	2
P	$\frac{C_3^2}{C_{10}^2}$	$\frac{C_7^1.C_3^1}{C_{10}^2}$	$\frac{C_7^2}{C_{10}^2}$

Hay

Y	0	1	2
P	$\frac{3}{45}$	$\frac{21}{45}$	$\frac{21}{45}$

Thí dụ 3 :
Một xạ thủ có xác suất bắn trúng bia cho mỗi lần bắn là 0,8. Anh ta bắn từng viên đạn vào bia cho đến khi trúng bia thì dừng. Gọi Z là số viên đạn xạ thủ được nhận để bắn vào bia khi đó quy luật phân phối xác suất của Z là

Z	1	2	3	.....	n	....
P	0,8	0,2.0,8	$0,2^2.0,8$	.....	$0,2^{n-1}.0,8$	....

Ta thấy $\sum_{n=1}^{+\infty} p_n = 0,8.\sum_{n=1}^{+\infty} 0,2^{n-1} = 0,8.\frac{1}{1-0,2} = 1$ (áp dụng công thức tổng cấp số nhân lùi vô hạn có công bội 0,2)

2.2.3. Hàm phân bố xác suất

a) Định nghĩa : Giả sử X là một biến ngẫu nhiên và x là một số thực. Hàm phân bố xác suất của biến ngẫu nhiên X ký hiệu là F(x) được định nghĩa

$$F(x) = P(X < x).$$

Nếu X là biến ngẫu nhiên rời rạc thì

$$F(x) = \sum_{x_i < x} p_i \qquad (*)$$

Nếu X là biến ngẫu nhiên liên tục thì hàm phân bố xác suất F(x) được cho dưới dạng hàm số có nhiều biểu thức.

Thí dụ : Cho X là biến ngẫu nhiên rời rạc có bảng quy luật phân phối xác suất như sau

X	1	2	4
P	0,25	0,5	0,25

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

Lời giải : Ta xét các trường hợp $x \le 1$, $1 < x \le 2$, $2 < x \le 4$, $x > 4$ và áp dụng công thức (*) ta có

$$F(x)=\begin{cases}0 & x\le 1\\ 0,25 & 1<x\le 2\\ 0,75 & 2<x\le 4\\ 1 & x>4\end{cases}$$

Nếu ta vẽ đồ thị của hàm F(x) thì đồ thị có dạng bậc thang.
b) <u>Các tính chất của hàm F(x)</u>
*) Nếu X là biến ngẫu nhiên bất kỳ ta có các tính chất sau
+) $\forall x\in R$ ta có : $0 \le F(x) \le 1$
+) $\forall x_1, x_2 \in R$ và $x_1 < x_2$ ta có : $F(x_1) \le F(x_2)$
+) $\forall a,b\in R$ và $a < b$ ta có : $P(a \le X < b) = F(b) - F(a)$
*) Nếu X là biến ngẫu nhiên liên tục thì ngoài các tính chất nêu trên còn có các tình chất sau
+) $P(X = x) = 0$ với x là số bất kỳ
+) $P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = P(a < X < b)$ với a, b là các số
thực bất kỳ $a < b$
+) Nếu biến ngẫu nhiên liên tục X chỉ nhận giá trị trong $[a; b]$ thì

$F(x) = 0$ với $x \le a$ và $F(x) = 1$ với $x > b$

+) $F(-\infty) = 0$; $F(+\infty) = 1$
+) F(x) là hàm liên tục
<u>Thí dụ</u> : Cho biến ngẫu nhiên liên tục X có hàm phân bố xác suất như sau

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

i) Tìm k
ii) Vẽ đồ thị của F(x) ứng với giá trị k tìm được.
iii) Tính $P(0 \le X < \frac{1}{3})$ và tính $P(\frac{1}{6} < X < 2)$

<u>Lời giải</u> :
i) k = 3
ii) Vẽ hình
iii) Ta có $F(x) = \frac{3}{4}x+\frac{3}{4}$ với $0 \le x < \frac{1}{3}$ do $[0; \frac{1}{3}) \subset (-1; \frac{1}{3}]$.

Vậy P($0 \le x < \frac{1}{3}$) = F($\frac{1}{3}$) - F(0) = $(\frac{3}{4}.\frac{1}{3}+\frac{3}{4})$ - $(\frac{3}{4}.0+\frac{3}{4})$ = 1 - $\frac{3}{4}=\frac{1}{4}=0,25$.

Chú ý : F(x) = P(X < x) phản ánh mức độ tập trung xác suất ở phía bên trái của 1 số thực x nào đó.

2.2.4.Hàm mật độ xác suất của biến ngẫu nhiên liên tục X

a) Định nghĩa : Cho X là một biến ngẫu nhiên liên tục có hàm phân bố xác suất F(x). Hàm mật độ xác suất của biến ngẫu nhiên liên tục X ký hiệu là *f(x)* được định nghĩa : *f(x)* = F$'$(x)

Nếu biết trước hàm mật độ *f(x)* thì hàm phân bố F(x) = $\int_{-\infty}^{x} f(t)dt$

b) Các tính chất của hàm f(x)

+) $f(x) \ge 0 \quad \forall x$

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

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

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

Thí dụ 1 : Cho biến ngẫu nhiên liên tục X có hàm phân bố xác suất như sau

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

i) Tìm hệ số k
ii) Tìm *f(x)*
iii) Tính P(0,25 < X < 0,75)

Lời giải :

i) Do F(x) là một hàm liên tục nên : $\lim\limits_{x\to 0}$ F(x) = F(0) và $\lim\limits_{x\to 1}$ F(x) = F(1)

=> k = 1

ii) Vì *f(x)* = F$'$(x) và k = 1 đã tính ở trên nên ta có

$$f(x) = \begin{cases} 2.x & \text{với } x \in [0;1] \\ 0 & \text{Với } x \notin [0;1] \end{cases}$$

iii) Vì (0,25; 0,75) $\subset$[0; 1] nên P(0,25 < X < 0,75) = $(0,75)^2$ - $(0,25)^2$ = 0,5.

<u>Thí dụ 2</u> : Cho biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau

$$f(x) = \begin{cases} m.x(1-x) & \text{với } x \in [0;1] \\ 0 & \text{với } x \notin [0;1] \end{cases}$$

i) Tìm hệ số *m*
ii) Tìm F(*x*)
iii) Tính P(X > 0,6)

<u>Lời giải</u> :

i) Áp dụng $\int_{-\infty}^{+\infty} f(x)dx = 1$ ta có $\int_{-\infty}^{+\infty} f(x) = \int_{-\infty}^{0} 0.dx + \int_{0}^{1} m.x(1-x)dx + \int_{1}^{+\infty} 0.dx = 1$

hay $m.\int_{0}^{1}(x - x^2)dx = 1$ <=> $m(\frac{x^2}{2} - \frac{x^3}{3})\Big|_0^1 = 1$

<=> $\frac{m}{6} = 1$ <=> $m = 6$

ii) Để tìm F(*x*) ta xét các trường hợp sau

-) Nếu $x \le 0$ thì F(*x*) = $\int_{-\infty}^{0} 0.dt = 0$

-) Nếu $0 < x \le 1$ thì

$$F(x) = \int_{-\infty}^{0} 0dt + \int_{0}^{x} 6t(1-t)dt = 6\left(\frac{t^2}{2} - \frac{t^3}{3}\right)\Bigg|_0^x = 3x^2 - 2x^3$$

-) Nếu $x > 1$ thì F(*x*) = $\int_{-\infty}^{0} 0.dt + \int_{0}^{1} 6.t.(1-t)dt + \int_{1}^{x} 0.dt = 1$

Vậy

$$F(x) = \begin{cases} 0 & \text{với } x \le 0 \\ 3x^2 - 2x^3 & \text{với } 0 < x \le 1 \\ 1 & \text{với } x > 1 \end{cases}$$

iii) P(X > 0,6) = P(0,6 < X < 1)
= F(1) - F(0,6) = 1 - $(3.0,6^2 - 2.0,6^3)$ = 0,352

<u>Bài tập tự giải tại lớp học</u>

1) Cho biến ngẫu nhiên liên tục X có hàm mật độ xác suất

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

Tìm xác suất để khi tiến hành 3 phép thử độc lập thì có 2 lần biến ngẫu nhiên X nhận giá trị trong khoảng (-1; 1).

2) Tuổi thọ của một loại sản phẩm là một biến ngẫu nhiên liên tục X có hàm mật độ xác suất

$$f(x) = \begin{cases} \dfrac{m}{x^2} & \text{với } x > 400 \text{ giờ} \\ 0 & \text{Với } x \leq 400 \text{ giờ} \end{cases}$$

i) Tìm *m*

ii) Tính P(X > 600)

2.3. Các tham số đặc trưng của biến ngẫu nhiên

2.3.1. Kỳ vọng toán của biến ngẫu nhiên

a) Kỳ vọng toán của biến ngẫu nhiên rời rạc : Cho biến ngẫu nhiên rời rạc X có bảng quy luật phân phối xác suất như sau

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

khi đó kỳ vọng toán của biến ngẫu nhiên X được tính bởi công thức

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

Trong trường hợp biến ngẫu nhiên rời rạc X có bảng quy luật phân phối xác suất

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

thì kỳ vọng toán của biến ngẫu nhiên X là

$$E(X) = \sum_{i=1}^{+\infty} x_i . p_i$$

b) Kỳ vọng toán của biến ngẫu nhiên liên tục : Cho biến ngẫu nhiên liên tục X có hàm mật độ xác suất *f(x)* khi đó kỳ vọng toán của biến ngẫu nhiên X được tính bởi công thức

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

Chú ý : *Bản chất của kỳ vọng toán của biến ngẫu nhiên là nó phản ánh giá trị trung tâm (trung bình) của phân phối xác suất của biến ngẫu nhiên đó (chi tiết xem trang 98 giáo trình)*

Các thí dụ :

Thí dụ 1 : Cho 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

Tính E(X)

<u>Lời giải</u> : Áp dụng công thức $E(X)=\sum_{i=1}^{n} x_i.p_i$ ta có E(X) = 1.0,1 + 3.0,5 + 4.0,4 = 3,2

<u>Thí dụ 2</u> : Cho biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau

$$f(x)=\begin{cases} 6x(1-x) & \text{với } x\in[0;1] \\ 0 & \text{với } x\notin[0;1] \end{cases}$$

Tính E(X)

<u>Lời giải</u> : Áp dụng công thức $E(X)=\int_{-\infty}^{+\infty} x.f(x)dx$ ta có

$$E(X)=\int_{-\infty}^{0} x.0dx+\int_{0}^{1} x.6x(1-x)dx+\int_{1}^{+\infty} x.0dx$$

$$E(X)=6\int_{0}^{1}(x^2-x^3)dx=6\left(\frac{x^3}{3}-\frac{x^4}{4}\right)\Big|_0^1=\frac{1}{2}=0,5$$

c) <u>Các tính chất của kỳ vọng toán</u>

+) Với C là hằng số và X là biến ngẫu nhiên ta có

+) E(C) = C

+) E(C.X) = C.E(X)

+) Cho X; Y là hai biến ngẫu nhiên bất kỳ ta có :

E(X + Y) = E(X) + E(Y)

Mở rộng tính chất trên : Nếu X_1; X_2;.....; X_n là các biến ngẫu nhiên bất kỳ ta có công thức sau :

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

+) Nếu X và Y là hai biến ngẫu nhiên độc lập thì E(X.Y) = E(X).E(Y)

Mở rộng tính chất trên : Nếu X_1; X_2;.....; X_n là các biến ngẫu nhiên độc lập lẫn nhau thì $E\left(\prod_{i=1}^{n} X_i\right)=\prod_{i=1}^{n} E(X_i)$

Chú ý: Hai biến ngẫu nhiên X và Y được gọi là độc lập với nhau nếu biến ngẫu nhiên X nhận bất kỳ giá trị nào trong số những giá trị có thể có của nó đều không làm thay đổi quy luật phân phối xác suất của biến ngẫu nhiên Y và ngược lại.

(Yêu cầu sinh viên đọc - hiểu các thí dụ 5; 6 và phần ứng dụng thực tế của kỳ vọng toán trong giáo trình Lý thuyết xác suất và thống kê toán của Tg : PGS. TS. Nguễn Cao Văn và TS. Trần Thái Ninh biên soạn, NXB Thống kê - 2005 (từ trang 100 đến 104)).

2.3.2. Các tham số trung vị (m_d) và mốt (m_0) :

a) Trung vị : Ký hiệu là m_d là giá trị chia phân phối của biến ngẫu nhiên thành hai phần bằng nhau.
Nếu X là biến ngẫu nhiên rời rạc thì x_i là trung vị m_d nếu $F(x_i) \le 0{,}5 < F(x_{i+1})$ với $F(x)$ là hàm phân phối xác suất của biến ngẫu nhiên X.
Nếu X là biến ngẫu nhiên liên tục có hàm mật độ xác suất *f(x)* thì trung vị m_d là giá trị thỏa mãn điều kiện : $\int_{-\infty}^{m_d} f(x)dx = 0{,}5$

b) Mốt : Ký hiệu là m_0 là giá trị của biến ngẫu nhiên tương ứng với
+) Xác suất lớn nhất nếu là biến ngẫu nhiên rời rạc
+) Giá trị cực đại của hàm mật độ xác suất nếu là biến ngẫu nhiên liên tục

Chú ý : Trong thực tế có thể gặp biến ngẫu nhiên không có giá trị m_0 hoặc ngược lại có thể có nhiều giá trị m_0 cùng một lúc.
Thí dụ : (Xem thí dụ 7 giáo trình trang 105)

2.3.3. Phương sai

a) Định nghĩa : Cho X là một biến ngẫu nhiên, phương sai của X ký hiệu là V(X) được định nghĩa

$$V(X) = E[X - E(X)]^2 \quad \text{hay} \quad V(X) = E(X^2) - [E(X)]^2$$

+) Phương sai của biến ngẫu nhiên rời rạc : Cho biến ngẫu nhiên rời rạc X có bảng quy luật phân phối xác suất như sau

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

khi đó phương sai của X được tính bởi công thức :

$$V(X) = \sum_{i=1}^{n}[x_i - E(X)]^2 p_i \quad \text{hay} \quad V(X) = \sum_{i=1}^{n} x_i^2 p_i - [E(X)]^2$$

+) Phương sai của biến ngẫu nhiên liên tục : Cho biến ngẫu nhiên liên tục X có hàm mật độ xác suất *f(x)* khi đó phương sai của X được tính bởi công thức :

$$V(X) = \int_{-\infty}^{+\infty}[x - E(X)]^2 f(x)dx \quad \text{hay} \quad V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - [E(X)]^2\ ]$$

Chú ý : *Bản chất phương sai của biến ngẫu nhiên là nó phản ánh mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình hay kỳ vọng toán của biến ngẫu nhiên đó (chi tiết xem trang 111 giáo trình)*

Thí dụ 1 : Cho 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

Tính V(X)

Lời giải : Áp dụng công thức $V(X)=\sum_{i=1}^{n} x_i^2 p_i - [E(X)]^2$ ta có

V(X) = (1^2.0,1 + 3^2.0,5 + 4^2.0,4) - $(3,2)^2$ = 11 - 10,24 = 0,76

Thí dụ 2 : Cho biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau

$$f(x)=\begin{cases} 6x(1-x) & \text{với } x\in[0;1] \\ 0 & \text{với } x\notin[0;1] \end{cases}$$

Tính V(X)

Lời giải : Áp dụng công thức $V(X)=\int_{-\infty}^{+\infty} x^2 f(x)dx - [E(X)]^2$ ta có

$$V(X)=(\int_{-\infty}^{0} x^2.0dx+\int_{0}^{1} x^2.6x(1-x)dx+\int_{1}^{+\infty} x^2.0dx)-[\mathrm{E(X)}]^2$$

$$= 6\int_0^1 (x^3-x^4)dx-(\frac{1}{2})^2 = 6(\frac{x^4}{4}-\frac{x^5}{5})\Big|_0^1-\frac{1}{4}=\frac{1}{20}$$

b) Các tính chất của phương sai

i) Với C là hằng số và X là biến ngẫu nhiên ta có

+) V(C) = 0

+) V(C X) = C^2 V(X)

+) V(C + X) = V(X)

ii) Cho X; Y là hai biến ngẫu nhiên độc lập ta có :

V(X + Y) = V(X) + V(Y) và V(X - Y) = V(X) + V(Y)

iii) Nếu X_1; X_2;.....; X_n là các biến ngẫu nhiên độc lập ta có công thức sau

$$V(\sum_{i=1}^{n} X_i)=\sum_{i=1}^{n} V(X_i)$$

*(Yêu cầu sinh viên đọc - hiểu các thí dụ 11; 12 và phần ứng dụng thực tế của phương sai trong giáo trình Lý thuyết xác suất và thống kê toán của Tg : PGS. TS. Nguễn Cao Văn và TS. Trần Thái Ninh biên soạn, NXB Thống kê - 2005 (từ trang 111 đến 114)).*Làm bài tập 2.58 sách bài tập tại lớp.

2.3.4. Độ lệch chuẩn

Độ lệch chuẩn của biến ngẫu nhiên X ký hiệu σ_X là căn bậc hai của phương sai :

$$\sigma_X=\sqrt{V(X)}$$

Ta nhận thấy đơn vị đo của phương sai bằng bình phương đơn vị đo của biến ngẫu nhiên. Vì vậy khi cần phải đánh giá mức độ phân tán của biến ngẫu nhiên theo đơn vị đo của nó người ta thương dùng độ lệch chuẩn chứ không dùng phương sai.

Các tham số khác như : Hệ số biến thiên; giá trị tới hạn; hệ số đối xứng; hệ số nhọn yêu cầu người học tham khảo giáo trình (như đã trích dẫn ở trên) từ trang 115 đến trang 117.

Chương III : Một số quy luật phân phối xác suất thông dụng

3.1. Quy luật không - một (0 - 1)

a) Định nghĩa : Biến ngẫu nhiên rời rạc X nhận một trong hai giá trị có thể có của nó là 0 hoặc 1 với các xác suất tương ứng được tính bởi công thức $P(X = x) = p^x(1-p)^{1-x}$ với $x = \overline{0;1}$, khi đó biến ngẫu nhiên X được gọi là tuân theo quy luật 0 - 1 với tham số *p* và ký hiệu là X ~ A(*p*).

Thí dụ : Tỷ lệ chính phẩm của một nhà máy được biết là 90%. Lấy ra một sản phẩm từ nhà máy và gọi X là dấu hiệu lấy được chính phẩm thì X ~ A(*p* = 0,9).

b) Bảng phân phối xác suất và các tham số đặc trưng của biến ngẫu nhiên X ~ A(*p*)

+) Bảng phân phối xác suất :

X	0	1
P	1- *p*	*p*

+) Các tham số đặc trưng

E(X) = *p* ; V(X) = *p*(1- *p*) ; $\sigma_X = \sqrt{p(1-p)}$

Chú ý : Quy luật 0 - 1 thường được dùng để đặc trưng cho các dấu hiệu nghiên cứu định tính có hai phạm trù luân phiên. Chẳng hạn giới tính, sản phẩm tốt - xấu, viên đạn trúng bia - không trúng bia,.v.v.

3.2. Quy luật nhị thức

a) Định nghĩa : Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có của nó là 0;1;2;....; *n* ($n \in N$) với các xác suất tương ứng được tính bởi công thức $P(X = x) = C_n^x p^x(1-p)^{n-x}$ với $x = \overline{0;n}$, khi đó biến ngẫu nhiên X được gọi là tuân theo quy luật nhị thức với các tham số *n*; *p* và ký hiệu là X ~ B(*n*; *p*).

Thí dụ : Một xạ thủ có xác suất bắn trúng vòng mười là 0,8 cho mỗi lần bắn. Anh ta được phát 5 viên đạn để lần lượt bắn vào bia, mỗi lần xạ thủ bắn gọi A là biến cố xạ thủ bắn trúng vòng mười, ta có P(A) = 0,8. Gọi X = (số lần biến cố A xảy ra trong 5 lần bắn), khi đó X ~ B(*n* = 5; *p* = 0,8).

b) Bảng phân phối xác suất và các tham số đặc trưng của biến ngẫu nhiên X ~ B(*n*; *p*)

+) Bảng phân phối xác suất : đặt *q* = 1- *p*

X	0	1	2	....	*x*	....	*n*
P	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	$C_n^2 p^2 q^{n-2}$	......	$C_n^x p^x q^{n-x}$	....	$C_n^n p^n q^0$

Chú ý : Theo công thức nhị thức Newton ta có :

$$1 = 1^n = (p+q)^n = \sum_{x=0}^{n} C_n^x p^x q^{n-x}$$

+) Các tham số đặc trưng

E(X) = np ; V(X) = npq ; $\sigma_X = \sqrt{npq}$; giá trị m_0 sao cho P(X = m_0) là lớn nhất được xác định : $np - p \le m_0 \le np + q$

c) <u>Quy luật phân phối xác suất của tần suất</u>

Giả sử ta thực hiện n phép thử độc lập, với mỗi phép thử biến cố A xảy ra với xác suất P(A) = p. Gọi X là số lần biến cố A xuất hiện trong n phép thử, nhưng ta muốn quan tâm tới tỉ lệ xuất hiện biến cố A trong n phép thử hơn là số lần xuất hiện biến cố A.

Đặt $f = \dfrac{X}{n}$

+) Bảng quy luật phân phối xác suất của f là

f	$\frac{0}{n}$	$\frac{1}{n}$	$\frac{2}{n}$	……..	$\frac{x}{n}$	…..	$\frac{n}{n}$
P	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	$C_n^2 p^2 q^{n-2}$	……..	$C_n^x p^x q^{n-x}$	…..	$C_n^n p^n q^0$

+) Các tham số đặc trưng của f

$$E(f) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$V(f) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}npq = \frac{pq}{n}$$

$$\sigma_f = \sqrt{\frac{pq}{n}}$$

Chú ý :

*) Nếu các biến ngẫu nhiên rời rạc X_i độc lập và X_i ~ A(p) $\forall i = \overline{1;n}$ thì biến ngẫu nhiên rời rạc $X = \sum_{i=1}^{n} X_i$ ~ B(n; p)

Quy luật A(p) là trường hợp riêng của quy luật B(n; p), quy luật A(p) là quy luật B(n; p) khi n = 1.

*) Nếu X_1; X_2 là các biến ngẫu nhiên rời rạc độc lập và X_1 ~ B(n_1; p); X_2 ~ B(n_2; p) thì biến ngẫu nhiên ($X_1 + X_2$) ~ B($n_1 + n_2$; p)

<u>Thí dụ 1</u> : Tung 1 đồng xu cân đối và đồng chất 3 lần liên tiếp. Tính xác suất để

i) Có 2 lần đồng xu xuất hiện mặt ngửa.

ii) Có ít nhất 1 lần đồng xu xuất hiện mặt ngửa.

iii) Có lần đồng xu xuất hiện mặt sấp, có lần đồng xu xuất hiệ mặt ngửa.

iv) Có không quá 1 lần đồng xu xuất hiện mặt sấp.

Lời giải : Gọi X = (Số lần đồng xu xuất hiện mặt ngửa), ta có X ~ B(n = 3; p = 0,5).

i) P(X = 2) = $C_3^2 0,5^2 0,5^1 = 0,375$

ii) P(X ≥ 1) = 1 - P(X = 0) = 1 - $C_3^0 0,5^0 0,5^3$ = 0,875

iii) P(Có cả sấp và ngửa) = 1 - ($C_3^0 0,5^0.0,5^3 + C_3^3 0,5^3.0,5^0$) = 0,75

iv) P(Có không quá 1 lần đồng xu xuất hiện mặt sấp) = P(X = 2) + P(X = 3) = 0,375 + 0,125 = 0,5.

Thí dụ 2 : Một phân xưởng dệt có 50 máy dệt hoạt động độc lập với nhau. Xác suất các máy dệt bị hỏng trong một ca sản xuất đều như nhau và bằng 0,07.

i) Tìm quy luật phân phối xác suất của số máy dệt bị hỏng trong một ca sản xuất

ii) Trung bình có bao nhiêu máy dệt bị hỏng trong một ca sản xuất ?

iii) Tìm xác suất để trong một ca sản xuất có trên 48 máy hoạt động tốt.

Lời giải :

i) Gọi Y = (Số máy dệt bị hỏng trong một ca sản xuất), ta có

Y ~ B(n = 50; p = 0,07).

ii) E(Y) = 50.0,07 = 3,5.

iii) P(Y < 2) = P(Y = 0) + P(Y = 1)

$$= C_{50}^0 0,07^0.0,93^{50} + C_{50}^1 0,07^1.0,93^{49} = 0,1265$$

3.3. Quy luật Poisson

Giả sử biến ngẫu nhiên rời rạc X ~ B(n; p), trong trường hợp n khá lớn và p khá nhỏ thì việc tính toán sẽ gặp nhiều khó khăn. Chẳng hạn n = 10000; p = 0,0002 thì P(X = 1000) = $C_{10000}^{1000} 0,0002^{1000}.0,9998^{9000} = ?$. Nhà toán học Poisson đã chứng minh được rằng khi n khá lớn; p khá nhỏ và $np = \lambda$ thì công thức Bernoulli $P_x = C_n^x p^x (1-p)^{n-x}$ sẽ xấp xỉ công thức $P_x = e^{-\lambda}\dfrac{\lambda^x}{x!}$ với x = 0; 1; 2;....; n;....

a) Định nghĩa : Biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có của nó là 0;1;2;....; n;....với các xác suất tương ứng được tính bởi công thức

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{với } x = 0;\ 1;\ 2;....;\ n;....$$

khi đó biến ngẫu nhiên X được gọi là tuân theo quy luật Poisson với tham số λ và ký hiệu là X ~ P(λ)

Chú ý : Ta có thể dùng bảng tính sẵn (xem phụ lục 2 từ trang 623 đến trang 628 giáo trình NXB Thống kê) để tính P(X ≤ x) tùy theo các giá trị của $x;\lambda$ đã cho.

b) Bảng quy luật phân phối xác suất và các tham số đặc trưng của biến ngẫu nhiên X ~ P(λ)

+) Bảng quy luật phân phối xác suất

X	0	1	2	....	n	....
P	$e^{-\lambda}.\frac{\lambda^0}{0!}$	$e^{-\lambda}.\frac{\lambda^1}{1!}$	$e^{-\lambda}.\frac{\lambda^2}{2!}$	....	$e^{-\lambda}.\frac{\lambda^n}{n!}$	....

Chú ý : sử dụng khai triển Taylor hàm $f(x)=e^x$ tại điểm x = 0 ta có

$$e^x=\sum_{i=0}^{+\infty}\frac{x^i}{i!} \Rightarrow e^\lambda=\sum_{x=0}^{+\infty}\frac{\lambda^x}{x!} \text{ do vậy } \sum_{x=0}^{+\infty}e^{-\lambda}.\frac{\lambda^x}{x!}=e^{-\lambda}\sum_{x=0}^{+\infty}\frac{\lambda^x}{x!}=e^{-\lambda}.e^{\lambda}=1$$

+) Các tham số đặc trưng

$E(X)=\lambda$; $V(X)=\lambda$; $\sigma_X=\sqrt{\lambda}$; giá trị m_0 sao cho P(X = m_0) là lớn nhất được xác định : $\lambda-1\le m_0\le\lambda$

<u>Thí dụ 1</u> : Xác suất trong khi vận chuyển mỗi chai rượu bị vỡ là 0,001. Người ta vận chuyển 3000 chai rượu đến cửa hàng.

i) Tìm số chai vỡ trung bình khi vận chuyển

ii) Tìm số chai vỡ có khả năng sảy ra nhiều nhất khi vận chuyển

<u>Lời giải</u> : Gọi X = (Số chai rượu bị vỡ khi vận chuyển), do n = 3000 (khá lớn)

p = 0,001 (khá nhỏ); nên đặt $np=\lambda$ => X ~ P($\lambda=3$)

i) Số chai vỡ trung bình là E(X) = 3.

ii) Gọi m_0 là số chai vỡ có khả năng xảy ra nhiều nhất,

ta có $\lambda-1\le m_0\le\lambda$ nên $2\le m_0\le 3$

<u>Thí dụ 2</u> : (Làm bài 3.25 sách bài tập)

<u>Lời giải</u> : Gọi X = (Số khách chờ đi ô tô buýt loại 6 chỗ ngồi), theo đầu bài cho ta có X ~ P($\lambda=2$)

i) P(X $\le$ 6) = P_0 + P_1 +...+ P_6 = 0,1351 + 0,2707 +...+ 0,012 = 0.9947.

ii)P(X $\ge$ 12) = 1 - P(X < 12)

= 1 - (0,9947 + 0,0034 + 0,0009 + 0,0002 + P_{10} + P_{11})

iii) P(X $\ge$ 8) = 1 - P(X < 8) = 1 - (0,9947 + 0,0034) = 0,0019 < 0,1 vậy không tăng thêm một xe chở khách nữa.

3.4. Quy luật phân phối đều

a) <u>Định nghĩa</u> : Biến ngẫu nhiên liên tục X được gọi là tuân theo quy luật phân phối đều trên [a; b] (với a; b $\in\mathbb{R}$, a < b) nếu hàm mật độ xác suất của X có dạng

$$f(x)=\begin{cases}\dfrac{1}{b-a} & \text{nếu } x\in[a;b]\\ 0 & \text{nếu } x\notin[a;b]\end{cases}$$

Ký hiệu X ~ U[a; b].

Hàm phân bố xác suất của X ~ U[a; b] là:

$$F(x)=\begin{cases}0 & \text{nếu } x\le a\\ \dfrac{x-a}{b-a} & \text{nếu } a<x\le b\\ 1 & \text{nếu } x>b\end{cases}$$

b) Các tham số đặc trưng : E(X) = $\frac{a+b}{2}$; V(X) = $\frac{(b-a)^2}{12}$.

Thí dụ : Khi thâm nhập một thị trường mới doanh nghiệp chỉ dự kiến được doanh số hàng tháng có thể đạt được tối thiểu 50 triệu đồng và tối đa 80 triệu đồng. Tuy nhiên để đảm bảo hoạt động kinh doanh, doanh nghiệp phải đạt tối thiểu 60 triệu đồng / một tháng. Vậy doanh nghiệp có nên thâm nhập thị trường đó hay không ?

Lời giải : Gọi X = (Doanh số hàng tháng doanh nghiệp có thể đạt được),
=> X ~ U[50; 80], (đơn vị triệu đồng).

Cách 1 : Ta có E(X) = $\frac{a+b}{2}=\frac{50+80}{2}=65>60$ => có thể thâm nhập được

Cách 2 : Hàm mật độ xác suất của X là :

$$f(x)=\begin{cases}\dfrac{1}{30} & \text{nếu } x\in[50;\,80]\\ 0 & \text{nếu } x\notin[50;\,80]\end{cases}$$

nên P(X ≥ 60) = $\int_{60}^{+\infty}\frac{1}{30}dx=\int_{60}^{80}\frac{1}{30}dx=\frac{1}{30}x\Big|_{60}^{80}=\frac{80-60}{30}=\frac{20}{30}>0,5$, doanh nghiệp có thể thâm nhập thị trường mới.

Chú ý : Trong thực tế quy luật phân phối đều được sử dụng khi không có thông tin về biến ngẫu nhiên.

3.5. Quy luật phân phối chuẩn

a) Định nghĩa : Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(-\infty;+\infty)$ được gọi là tuân theo quy luật phân phối chuẩn với các tham số μ và σ^2 và ký hiệu là

X ~ $N(\mu;\sigma^2)$ nếu hàm mật độ xác suất của X có dạng

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\qquad \forall x$$

+) Nếu tiến hành khảo sát hàm *f*(*x*) ta chú ý một số tính chất sau

-) Đường cong đối xứng qua đường thẳng $x=\mu$

-) Đồ thị nhận trục hoành làm đường tiệm cận ngang

-) Điểm cực đại có tọa độ $\left(\mu\ ;\ \frac{1}{\sigma\sqrt{2\pi}}\right)$

-) Có hai điểm uốn $\left(\mu+\sigma\,;\ \frac{1}{\sigma\sqrt{2\pi e}}\right)$ và $\left(\mu-\sigma\,;\ \frac{1}{\sigma\sqrt{2\pi e}}\right)$

(Vẽ đồ thị)

+) Hàm phân bố xác suất của X là

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$

+) Các tham số đặc trưng của biến ngẫu nhiên liên tục X ~ N($\mu;\sigma^2$)

E(X) = μ ; V(X) = σ^2; $\sigma_X=\sigma$

b) <u>Biến ngẫu nhiên có phân phối chuẩn hóa</u>

Cho biến ngẫu nhiên liên tục X ~ N($\mu;\sigma^2$); đặt U = $\frac{X-\mu}{\sigma}$ khi đó biến ngẫu nhiên U được gọi là tuân theo quy luật phân phối chuẩn hóa và ký hiệu là U ~ N(0; 1). Hàm mật độ xác suất của U có dạng

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \qquad \forall u$$ (vẽ đồ thị, chú ý $\varphi(u)$ là hàm chẵn)

Các giá trị của $\varphi(u)$ được tính sẵn thành bảng xem phụ lục 4 giáo trình : chẳng hạn

$\varphi(-1,51)=\varphi(1,51)=0,1276$

+) Các tham số đặc trưng của biến ngẫu nhiên U: E(U) = 0; V(U) = 1

+) Hàm phân bố xác suất của U là

$$\Phi(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{t^2}{2}}dt$$

Xét $u > 0$ ta có :

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{t^2}{2}}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-\frac{t^2}{2}}dt+\frac{1}{\sqrt{2\pi}}\int_{0}^{u} e^{-\frac{t^2}{2}}dt$$

đặt $\Phi_0(u)=\frac{1}{\sqrt{2\pi}}\int_{0}^{u} e^{-\frac{t^2}{2}}dt$

=> $\Phi(u) = 0,5 + \Phi_0(u)$

+) Giá trị của hàm $\Phi_0(u)$ được tính sẵn thành bảng (xem phụ lục 5 giáo trình)

<u>Thí dụ</u> : tra bảng phụ lục 5 ta có: $\Phi_0(1,96) = 0,475$;

$\Phi_0(-1,645) = -\Phi_0(1,645) = -0,45$

+) Tính chất của hàm $\Phi_0(u)$

-) $\Phi_0(-u) = -\Phi_0(u) \quad \forall u$

-) $\Phi_0(+\infty) = 0,5$

-) $\Phi_0(u) = 0,5 \ \forall u \ge 5$

(*) <u>***Công thức tính xác suất***</u>:

Giả sử biến ngẫu nhiên liên tục X ~ N($\mu;\sigma^2$) và $a;b\in\mathbb{R}$

ta có công thức :

$$P(a < X < b) = \Phi_0\left(\frac{b-\mu}{\sigma}\right) - \Phi_0\left(\frac{a-\mu}{\sigma}\right)$$

c) Giá trị tới hạn chuẩn mức xác suất α

Ký hiệu là u_α là giá trị thỏa mãn P(U > u_α) = α (vẽ hình)

Giá trị tới hạn chuẩn có tính chất : $u_{1-\alpha} = -u_\alpha$

Các giá trị tới hạn chuẩn được tính sẵn thành bảng (xem phụ lục 6 giáo trình)

Thí dụ : $u_{0,025} = 1,96$; $u_{0,05} = 1,645$; $u_{0,95} = -\ u_{0,05} - 1,645$

Ta có thể tìm $u_{0,025}$ bằng cách khác :

P(U > $u_{0,025}$) = 0,025 <=> 1 - P(U > $u_{0,025}$) = 1 - 0,025 <=> P(U < $u_{0,025}$) = 0,97 <=> Φ($u_{0,025}$) = 0,975 <=> $\Phi_0(u_{0,025})$ = 0,475

tra bảng phụ lục 5 có 0,475 = $\Phi_0(1,96)$

do đó : $\Phi_0(u_{0,025}) = \Phi_0(1,96)$ hay $u_{0,025}$ = 1,96.

d) Quy tắc 2σ và 3σ

Giả sử X là biến ngẫu nhiên và X ~ N(μ, σ^2). Cho ε là số dương cho trước khi đó

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

+) Nếu thay ε = 2σ ta có quy tắc 2σ như sau

$$P(|X - \mu| < 2\sigma) = 2\Phi_0(2) = 0,9544$$

+) Nếu thay ε = 3σ ta có quy tắc 3σ   như sau

$$P(|X - \mu| < 3\sigma) = 2\Phi_0(3) = 0,9974$$

Chú ý : Ta biết rằng biến ngẫu nhiên rời rạc X ~ B(*n*; *p)* mà khi *n* lớn *p* nhỏ và $np = \lambda$ thì ta coi X ~ P(λ). Nhưng trong trường hợp *n* lớn *p* không nhỏ thì sao ? Thực tế nếu *n* > 5 và $\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \frac{1}{\sqrt{n}} < 0,3$ thì ta có thể coi X ~ N[*np*; *np*(1- *p*)]

do đó : $P(X = x) = C_n^x p^x (1-p)^{n-x} \simeq \frac{1}{\sqrt{np(1-p)}} \varphi\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$

và $P(x \le X \le x+h) \simeq \Phi_0\left(\frac{x+h-np}{\sqrt{np(1-p)}}\right) - \Phi_0\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$ $h \in \mathbb{N}$

Thí dụ 1 : (Đọc thí dụ 3 trang 166 -> 167 giáo trình NXB thống kê)

Thí dụ 2 : Một trại chăn nuôi có hai giống lợn A và B có số lượng như nhau. Trọng lượng của giống lợn loại A có phân phối chuẩn với trung bình là 50kg và độ lệch chuẩn là 5kg, trọng lượng của giống lợn loại B có phân phối chuẩn với trung bình là 50kg và độ lệch chuẩn là 6kg.

Lợn đủ tiêu chuẩn xuất chuồng nếu có trọng lượng > 47kg.

i) Tìm xác suất để bắt được một con lợn đủ tiêu chuẩn xuất chuồng

ii) Bắt ngẫu nhiên ba con lợn từ trại chăn nuôi. Gọi X = (Số con đủ tiêu chuẩn xuất chuồng trong 3 con được bắt ra), tìm quy luật phân phối xác suất của X, tính E(X) và V(X).

iii) Tính xác suất để bắt được ít nhất một con đủ tiêu chuẩn xuất chuồng trong ba con bắt ra.

Lời giải :

i) Gọi H_1 = (Bắt được con lợn loại A); H_2 = (Bắt được con lợn loại B)

=> $P(H_1) = P(H_2) = 0,5$.

Gọi X_A = (Trọng lượng của lợn loại A); X_B = (Trọng lượng của lợn loại B)

=> $X_A \sim N(50; 5^2)$; $X_B \sim N(50; 6^2)$ (đơn vị kg).

Gọi A = (Bắt được con lợn đủ tiêu chuẩn xuất chuồng)

$$P(A/H_1) = P(X_A > 47) = \Phi_0\left(\frac{+\infty-50}{5}\right) - \Phi_0\left(\frac{47-50}{5}\right)$$

$$= 0,5 + \Phi_0(0,6) = 0,7257.$$

$$P(A/H_2) = P(X_B > 47) = \Phi_0\left(\frac{+\infty-50}{6}\right) - \Phi_0\left(\frac{47-50}{6}\right)$$

$$= 0,5 + \Phi_0(0,5) = 0,6915$$

Áp dụng công thức xác suất đầy đủ ta có

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

$$= 0,5.0,7257 + 0,5.0,6915 = 0,7086$$

ii) $X \sim B(n = 3; p = 0,7086)$; E(X) = 2,1258, V(X) = 0,61945812.

iii) Gọi B = (Bắt được ít nhất một con đủ tiêu chuẩn xuất chuồng trong ba con)

$$P(B) = P(X \geq 1) = 1 - P(X = 0) = 1 - C_3^3(1-0,7086)^3 = 0,975256.$$

3.6. Quy luật khi bình phương

a) Khái niệm :

Giả sử $X_1; X_2;...; X_n$ là các biến ngẫu nhiên liên tục và độc lập với nhau, giả sử các biến ngẫu nhiên $X_i \sim N(0; 1)$ $\forall i = \overline{1;n}$.

Đặt $\chi^2 = \sum_{i=1}^{n} X_i^2$ khi đó biến ngẫu nhiên χ^2 tuân theo quy luật khi bình phương với n bậc tự do và ký hiệu là $\chi^2 \sim \chi^2(n)$.

(Hàm mật độ xác suất của biến ngẫu nhiên có phân phối χ^2 xem giáo trình NXB thống kê trang 169).

b) Giá trị tới hạn khi bình phương : ký hiệu là $\chi_\alpha^2(n)$ là giá trị thỏa mãn

$$P[\chi^2 > \chi_\alpha^2(n)] = \alpha \qquad \text{(vẽ đồ thị hàm mật độ)}$$

Các giá trị $\chi^2_\alpha(n)$ được tính sẵn thành bảng xem phụ lục 7 giáo trình, chẳng hạn : $\chi^2_{0,025}(40) = 59,34$ hay $P[\chi^2(40) > 59,34]=0,025$.

3.7. Quy luật Student

a) Khái niệm : Giả sử U và V là các biến ngẫu nhiên liên tục và độc lập với nhau, giả sử U ~ N(0; 1), V ~ $\chi^2(n)$.

Đặt T = $\dfrac{U}{\sqrt{\dfrac{V}{n}}}$ khi đó biến ngẫu nhiên T được gọi là tuân theo quy luật Student với n bậc tự do và ký hiệu là T ~ T(n).

(Hàm mật độ xác suất của biến ngẫu nhiên có phân phối Student xem giáo trình NXB thống kê trang 171)

b) Giá trị tới hạn Student : ký hiệu là $t^{(n)}_\alpha$ là giá trị thỏa mãn

$$P[T > t^{(n)}_\alpha] = \alpha \qquad \text{(vẽ đồ thị hàm mật độ)}$$

+) Tính chất của giá trị tới hạn phân phối Student

$$t^{(n)}_{1-\alpha} = -t^{(n)}_\alpha$$

Các giá trị $t^{(n)}_\alpha$ được tính sẵn thành bảng xem phụ lục 8 giáo trình, chẳng hạn : $t^{(25)}_{0,05} = 1,708$; $t^{(30)}_{0,975} = -t^{(30)}_{0,025} = -\ 2,042$; $t^{(120)}_{0,025} = 1,98$

hay : $P[\mathrm{T}(25) > 1,708] = 0,05$

Chú ý : Khi số bậc tự do n tăng lên thì phân phối Student sẽ hội tụ rất nhanh về phân phối chuẩn hóa, do vậy với n lớn (thông thường n > 30) thì ta có thể thay thế giá trị tới hạn mức α của phân phối Student bằng giá trị tới hạn chuẩn mức α.

Chẳng hạn : $t^{(120)}_{0,025} \simeq u_{0,025} = 1,96$

3.8. Quy luật Fisher

a) Khái niệm : Giả sử U và V là các biến ngẫu nhiên liên tục và độc lập với nhau, giả sử U ~ $\chi^2(n_1)$, V ~ $\chi^2(n_2)$.

Đặt F = $\dfrac{U/n_1}{V/n_2} = \dfrac{n_2.U}{n_1.V}$ khi đó biến ngẫu nhiên F được gọi là tuân theo quy luật Fisher với n_1 và n_2 bậc tự do và ký hiệu là F ~ F(n_1;n_2).

(Hàm mật độ xác suất của biến ngẫu nhiên có phân phối Fisher xem giáo trình NXB thống kê trang 173 -> 174)

b) Giá trị tới hạn Fisher : ký hiệu là $f_\alpha(n_1;n_2)$ là giá trị thỏa mãn

$$P[F > f_\alpha(n_1;n_2)]=\alpha \qquad \text{(vẽ đồ thị hàm mật độ)}$$

+) Tính chất của giá trị tới hạn phân phối Fisher :

$$f_{1-\alpha}(n_1;n_2) = \frac{1}{f_\alpha(n_2;n_1)}$$

Chứng minh :

Ta có

$$P[F(n_2;n_1) > f_\alpha(n_2;n_1)] = \alpha \quad \Leftrightarrow \quad P\left(\frac{1}{F(n_2;n_1)} < \frac{1}{f_\alpha(n_2;n_1)}\right) = \alpha$$

$$\Leftrightarrow P\left(\frac{1}{F(n_2;n_1)} > \frac{1}{f_\alpha(n_2;n_1)}\right) = 1-\alpha$$

theo định nghĩa : $F(n_1;n_2)F(n_2;n_1) = 1 \Leftrightarrow \dfrac{1}{F(n_2;n_1)} = F(n_1;n_2)$

do đó : $$P\left(F(n_1;n_2) > \frac{1}{f_\alpha(n_2;n_1)}\right) = 1 - \alpha \qquad (1)$$

mặt khác theo định nghĩa : $P[F(n_1;n_2) > f_{1-\alpha}(n_1;n_2)] = 1 - \alpha$ (2)

từ (1) và (2) => $f_{1-\alpha}(n_1;n_2) = \dfrac{1}{f_\alpha(n_2;n_1)}$ đpcm.

Các giá trị $f_\alpha(n_1;n_2)$ được tính sẵn thành bảng xem phụ lục 9 giáo trình,

chẳng hạn : $f_{0,025}(7;6) = 5,7$; $f_{0,975}(6;7) = \dfrac{1}{f_{0,025}(7;6)} = \dfrac{1}{5,7}$

hay : $P[F(7;6) > 5,7] = 0,025$

Chương IV : Biến ngẫu nhiên hai chiều

4.1. Khái niệm về biến ngẫu nhiên nhiều chiều

+) Thí dụ : Một máy sản xuất một loại sản phẩm. Gọi X = (Chiều dài của sản phẩm); Y = (Chiều rộng của sản phẩm); Z = (Chiều cao của sản phẩm).

-) Nếu ta chỉ quan tâm đến chiều dài X của sản phẩm thì ta có một biến ngẫu nhiên, việc nghiên cứu một biến ngẫu nhiên ta đã thực hiện ở các chương trước.

-) Nếu ta quan tâm đồng thời đến chiều dài X và chiều rộng Y của sản phẩm thì ta có biến ngẫu nhiên đồng thời hai chiều (X, Y).

-) Tương tự nếu ta quan tâm tới cả ba chiều dài, rộng và cao của sản phẩm thì ta có biến ngẫu nhiên đồng thời ba chiều (X, Y, Z).

+) Tổng quát : Giả sử X_1, X_2,..., X_n là các biến ngẫu nhiên, ta gọi biến ngẫu nhiên

X = (X_1, X_2,..., X_n) là biến ngẫu nhiên *n* chiều và các biến ngẫu nhiên X_1, X_2,..., X_n là các biến ngẫu nhiên thành phần, chúng là các biến ngẫu nhiên một chiều.

+) Biến ngẫu nhiên hai chiều (X, Y) là rời rạc nếu các biến ngẫu nhiên X, Y là rời rạc và là liên tục nếu các biến ngẫu nhiên X, Y là liên tục.

4.2. Bảng phân phối xác suất của biến ngẫu nhiên rời rạc hai chiều

Giả sử biến ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có của nó là x_1, x_2,...,x_n và biến ngẫu nhiên rời rạc Y nhận một trong các giá trị có thể có của nó là y_1, y_2,..., y_m. Ký hiệu p(x_i, y_j) = P(X = x_i, Y = y_j) $\forall i = \overline{1;n}$

và $\forall j=\overline{1;m}$, khi đó ta có bảng phân phối xác suất của biến ngẫu nhiên hai chiều (X, Y) có dạng

Y \ X	x_1	x_2	....	x_n	**P_Y**
y_1	p(x_1,y_1)	p(x_2, y_1)	....	p(x_n, y_n)	**p(y_1)**
y_2	p(x_1,y_2)	p(x_2,y_2)	....	p(x_n,y_2)	**p(y_2)**
. . .	. . .	. . .	....	. . .	. . .
y_m	p(x_1,y_m)	p(x_2,y_m)	....	p(x_n,y_m)	**p(y_m)**
P_X	**p(x_1)**	**p(x_2)**	....	**p(x_n)**	**1**

bảng trên trở thành bảng quy luật phân phối xác suất đồng thời nếu các xác suất p(x_i, y_j) thỏa mãn

$$\begin{cases} 0\le p(x_i,y_j)\le 1 & \forall i=\overline{1;n} \text{ và } \forall j=\overline{1;m} \\ \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)=1 \end{cases}$$

+) Biết bảng phân phối xác suất đồng thời của biến ngẫu nhiên hai chiều (X, Y) bao giờ cũng có thể tìm được bảng phân phối xác suất biên của mỗi thành phần.

-) Bảng phân phối xác suất biên thành phần X

X	x_1	x_2	....	x_n
P	p(x_1)	p(x_2)	....	p(x_n)

trong đó : $$\begin{cases} p(x_i)=\sum_{j=1}^{m} p(x_i,y_j) \\ \sum_{i=1}^{n} p(x_i)=1 \end{cases} \qquad i=\overline{1;n}$$

-) Bảng phân phối xác suất biên thành phần Y

Y	y_1	y_2	....	y_m
P	p(y_1)	p(y_2)	....	p(y_m)

trong đó : $\begin{cases} p(y_j) = \sum_{i=1}^{n} p(x_i, y_j) \\ \sum_{j=1}^{m} p(y_j) = 1 \end{cases}$ $\quad j = \overline{1;m}$

+) Ta cũng có thể tìm được bảng phân phối xác suất có điều kiện, tùy theo yêu cầu bài toán. Chẳng hạn tìm quy luật phân phối xác suất của biến ngẫu nhiên (X / Y = y_j) với mỗi j

Ta có P(X = x_i / Y = y_j) = $\dfrac{p(x_i, y_j)}{p(y_j)}$, với $i = \overline{1;n}$

Do vậy ta có bảng phân phối xác suất của biến ngẫu nhiên
(X / Y = y_j) như sau :

X/Y=y_j	x_1	x_2	....	x_n
P	$\dfrac{p(x_1, y_j)}{p(y_j)}$	$\dfrac{p(x_2, y_j)}{p(y_j)}$	....	$\dfrac{p(x_n, y_j)}{p(y_j)}$

+) Tương tự ta có thể tìm được bảng phân phối xác suất có điều kiện
 (Y / X = x_i)

Chú ý : Nếu ta biết bảng phân phối xác suất đồng thời của biến ngẫu nhiên rời rạc hai chiều (X, Y) như trên ta có

i) P(X = x_i, Y = y_j) = P(Y = y_j).P[(X = x_i) / (Y = y_j)]
= P(X = x_i).P[(Y = y_j) / (X = x_i)] $\forall i = \overline{1;n}$ và $\forall j = \overline{1;m}$

ii) Nếu X, Y là hai biến ngẫu nhiên độc lập thì ta có
P(X = x_i, Y = y_j) = P(X = x_i).P(Y = y_j) $\forall i = \overline{1;n}$ và $\forall j = \overline{1;m}$

iii) Nếu tồn tại một cặp (x_i, y_j) mà P(X = x_i, Y = y_j) $\neq$ P(X = x_i).P(Y = y_j) thì X và Y là hai biến ngẫu nhiên phụ thuộc.

<u>Thí dụ 1</u> : Cho biến ngẫu nhiên rời rạc hai chiều (X, Y) có bảng quy luật phân phối xác suất đồng thời như sau :

Y \ X	0	1	2	**P_Y**
1	0,05	0,1	0,05	**0,2**
2	0,03	0,02	0,1	**0,15**
3	0,12	0,1	0,08	**0,3**
4	0,05	0,18	0,12	**0,35**
P_X	**0,25**	**0,4**	**0,35**	**1**

i) Tìm các bảng phân phối xác suất biên thành phần X, Y và (Y / X = 0)
ii) Các biến ngẫu nhiên thành phần X và Y độc lập hay phụ thuộc

<u>Lời giải</u> :

i) Ta có các bảng quy luật phân phối xác suất cần tìm như sau :

X	0	1	2
P	0,25	0,4	0,35

Y	1	2	3	4
P	0,2	0,15	0,3	0,35

Y/X=0	1	2	3	4
P	$\frac{0,05}{0,25}$	$\frac{0,03}{0,25}$	$\frac{0,12}{0,25}$	$\frac{0,05}{0,25}$

ii) Ta có : P(X = 1, Y = 1) = 0,1;   P(X = 1) = 0,4;   P(Y = 1) = 0,2 do đó
P(X = 1, Y = 1) $\neq$ P(X = 1).P(Y = 1) (vì 0,1 $\neq$ 0,08)
Vậy X và Y phụ thuộc.

4.3. Các tham số đặc trưng của hệ hai biến ngẫu nhiên

a) <u>Kỳ vọng toán của các biến ngẫu nhiên thành phần X và Y</u>

+) $E(X)=\sum_{i=1}^{n} x_i p(x_i)$

+) $E(Y)=\sum_{j=1}^{m} y_j p(y_j)$

b) <u>Phương sai và độ lệch chuẩn của các biến ngẫu nhiên thành phần X và Y</u>

+) $V(X)=\sum_{i=1}^{n} x_i^2 p(x_i)-[E(X)]^2$

+) $V(Y)=\sum_{j=1}^{m} y_j^2 p(y_j)-[E(Y)]^2$

+) $\sigma_X=\sqrt{V(X)}$ và $\sigma_Y=\sqrt{V(Y)}$

c) <u>Hiệp phương sai và hệ số tương quan của hai biến ngẫu nhiên X và Y</u>

+) Hiệp phương sai ký hiệu là Cov(X, Y) được xác định như sau :

$$\text{Cov}(X, Y) = E[(X - E(X)).(Y - E(Y))]$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p(x_i, y_j)-E(X).E(Y)$$

+) Hệ số tương quan ký hiệu là ρ_{xy} được xác định như sau :

$$\rho_{xy}=\frac{Cov(X,Y)}{\sigma_X.\sigma_Y}=\frac{Cov(X,Y)}{\sqrt{V(X)}.\sqrt{V(Y)}}$$

Các tính chất của hệ số tương quan

-) $\rho_{xy} = \rho_{yx}$

-) $-1 \le \rho_{xy} \le 1$

-) Nếu X và Y độc lập thì $\rho_{xy} = 0$, điều ngược lại không đúng.

-) Nếu $\rho_{xy} = \pm 1$ thì X và Y phụ thuộc hàm số (phụ thuộc tuyến tính)

-) $V(aX \pm bY) = a^2V(X) \pm 2abCov(X, Y) + b^2V(Y)$

-) $\rho_{xy} = 0$ thì ta gọi X và Y không tương quan với nhau, còn nếu $\rho_{xy} \ne 0$ thì ta gọi X và Y có tương quan với nhau

Thí dụ 2 : Có hai loại cổ phiếu A và B được rao bán trên thị trường chứng khoán, gọi X và Y lần lượt là lãi suất của cổ phiếu A và B (đơn vị %). Giả sử biến ngẫu nhiên rời rạc hai chiều (X, Y) có bảng phân phối xác suất đồng thời như sau.

Y \ X	-2	0	5	10
0	0	0,05	0,05	0,1
4	0,05	0,1	0,25	0,15
6	0,1	0,05	0,1	0

i) Nếu đầu tư toàn bộ vào cổ phiếu A thì lãi suất kỳ vọng và mức độ rủi ro là bao nhiêu ?
ii) Nếu mục tiêu là đạt được lãi suất kỳ vọng lớn nhất thì nên đầu tư vào hai loại cổ phiếu trên theo tỉ lệ nào ?
iii) Muốn hạn chế rủi ro về lãi suất đến mức thấp nhất thì nên đầu tư vào hai loại cổ phiếu trên theo tỉ lệ nào ?

Lời giải :

i) Ta có bảng phân phố xác suất của X là :

X	-2	0	5	10
P	0,15	0,2	0,4	0,25

do vậy : E(X) = -2.0,15 + 0.0,2 + 5.0,4 + 10.0,25 = 4,2 (%)

V(X) = $(-2)^2$.0,15 + 5^2.0,4 + 10^2.0,25 - $(4,2)^2$ = 17,96 => $\sigma_X = 4,238(\%)$.

ii) Gọi *p* (với $0 \le p \le 1$) là tỉ lệ đầu tư vào cổ phiếu A khi đó tỉ lệ đầu tư vào cổ phiếu B là 1 - *p*. Như vậy ta phải tìm *p* sao cho E[*p*.X + (1 - *p*).Y] đạt max. Tính E(Y) và V(Y) tương tự phần (i): E(Y) = 3,7 (%) và V(Y) = 4,11 => $\sigma_Y = 2,0273(\%)$.

E[p.X + (1 - p).Y] = p.E(X) + (1 - p).E(Y) = 4,2p + (1 - p).3,7 = 0,5p + 3,7 do $0 \le p \le 1$ và 0,5 > 0 do đó $3,7 \le 0,5.p + 3,7 \le 4,2$ => E[p.X + (1 - p).Y] đạt max bằng 4,2 khi p = 1, tức là đầu tư mua toàn bộ cổ phiếu A.

iii) Gọi p như phần (ii) tìm p sao cho V[p.X + (1 - p).Y] đạt min, ta có

$$V[p.X + (1 - p).Y] = p^2.V(X) + 2p(1 - p).Cov(X, Y) + (1 - p)^2.V(Y)$$

trong đó : V(X) = 17,96 và V(Y) = 4,11

$$\text{Cov}(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p(x_i, y_j) - E(X).E(Y)$$

$$= -8.0,05 - 12.0,1 + 20.0,25 + 30.0,1 + 40.0,15 - 4,2.3,7$$

$$= -3,14$$

$$\Rightarrow V[p.X + (1 - p).Y] = 17,96.p^2 - 6,28.(p - p^2) + 4,11 - 8,22.p + 4,11.p^2$$

$$= 28,35.p^2 - 14,5.p + 4,11$$

Ta có $f(p) = 28,35.p^2 - 14,5.p + 4,11$ đạt min khi $p = \frac{14,5}{56,7} = 0,255732$

Vậy để rủi ro thấp nhất thì ta đầu tư vào cổ phiếu A với tỷ lệ là 25,5732% và cổ phiếu B là 74,4268%

4.4. Kỳ vọng có điều kiện, khái niệm hàm hồi quy

a) Kỳ vọng có điều kiện : Cho x là một giá trị xác định của biến ngẫu nhiên rời rạc X, giả sử Y là biến ngẫu nhiên rời rạc có thể nhận một trong các giá trị có thể có là $y_1, y_2, \dots, y_m$. Khi đó

$$E(Y / X = x) = \sum_{j=1}^{m} y_j.P[(Y = y_j)/(X = x)]$$

Tương tự ta cũng có định nghĩa

$$E(X / Y = y) = \sum_{i=1}^{n} x_i.P[(X = x_i)/(Y = y)]$$

Thí dụ 3 : Quay lại phần (i) thí dụ 1 tìm E(Y / X = 0)

Ta có bảng phân phối xác suất của (Y / X = 0) là

Y / X = 0	1	2	3	4
P	$\frac{0,05}{0,25}$	$\frac{0,03}{0,25}$	$\frac{0,12}{0,25}$	$\frac{0,05}{0,25}$

do đó $E(Y / X = 0) = 1.\frac{0,05}{0,25} + 2.\frac{0,03}{0,25} + 3.\frac{0,12}{0,25} + 4.\frac{0,05}{0,25} = \frac{0,67}{0,25} = 2,68$

b) Hàm hồi quy : Ta đã biết mối quan hệ hàm số $y = f(x)$, đó là ứng với mỗi giá trị của x thuộc tập xác định cho ta duy nhất một giá trị của y thuộc tập giá trị - đó là quan hệ đơn trị. Nhưng trong thống kê kinh tế có nhiều mối quan hệ mà ứng với mỗi giá trị của X lại cho ta nhiều giá trị của Y.

Chẳng hạn những người cùng có mức thu nhập (X) là 3 triệu đồng / 1 tháng, nhưng mỗi người lại có mức tiêu dùng (Y) khác nhau, tức là ứng với mỗi giá trị của X cho ta nhiều giá trị của Y, đó là quan hệ đa trị. Để đưa quan hệ này về quan hệ đơn trị người ta dùng kỳ vọng có điều kiện và đưa ra khái niệm hàm hồi quy.

Vậy với mỗi giá trị x đặt $f(x)$ = E(Y / X = x) khi đó $f(x)$ là hàm hồi quy của Y đối với X, tương tự đặt $f(y)$ = E(X / Y = y) khi đó $f(y)$ là hàm hồi quy của X đối với Y

Chương V : Các định lý giới hạn

5.1. Bất đẳng thức Trêbưsép

Nếu X là biến ngẫu nhiên có kỳ vọng và phương sai hữu hạn thì với mọi $\varepsilon > 0$ tùy ý ta đều có

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

Chứng minh (xem giáo trình trang 234 - 235)

Chú ý giải thích trường hợp $V(X) \geq \varepsilon^2$

5.2. Định lý Trêbưsép

Nếu các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập từng đôi, có các kỳ vọng hữu hạn và các phương sai đều bị chặn trên bởi hằng số C ($V(X_i) \leq C$, $\forall i = \overline{1;n}$) thì với mọi $\varepsilon > 0$ bé tùy ý ta luôn có

$$\lim_{n \to \infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n}\right| < \varepsilon\right) = 1$$

Chứng minh (xem giáo trình trang 236 - 237)

Nếu các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập từng đôi, có các kỳ vọng hữu hạn và cùng bằmg *m* và các phương sai đều bị chặn trên bởi hằng số C ($V(X_i) \leq C$, $\forall i = \overline{1;n}$) thì với mọi $\varepsilon > 0$ bé tùy ý ta luôn có

$$\lim_{n \to \infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - m\right| < \varepsilon\right) = 1$$

Bản chất của định lý 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

5.3. 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 $\varepsilon > 0$ bé tùy ý ta luôn có

$$\lim_{n\to\infty} P\left(\left|f-p\right|<\varepsilon\right)=1$$

Chứng minh (xem giáo trình trang 240 - 241)

Định lý Bernoulli chứng tỏ sự ổn định của tần suất xung quanh giá trị xác suất của biến cố đó

5.4. Định lý giới hạn trung tâm

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 và phương sai hữu hạn

$$E(X_i)=\mu,\quad V(X_i)=\sigma^2,\quad \forall i=1,2,...,n,...$$

Đặt $U_n=\sum_{i=1}^{n}X_i$ và xét biến ngẫu nhiên $U_n^c=\dfrac{U_n-E(U_n)}{\sqrt{V(U_n)}}$ thì U_n^c sẽ hội tụ về biến ngẫu nhiên có phân phố chuẩn hóa khi $n\to\infty$. Tức là với n lớn thì

$$P(U_n^c<x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

Định lý giới hạn trung tâm còn gọi là định lý tiệm cận chuẩn.

Chương VI : Cơ sở lý thuyết mẫu

6.1. Khái niệm về phương pháp mẫu

+) Trong thực tế chúng ta thường phải nghiên cứu một hay một số dấu hiệu định tính hoặc định lượng trên một tập hợp các phần tử thuần nhất gọi là tổng thể. Thực chất các dấu hiệu này có thể coi là các biến ngẫu nhiên. Vì vậy việc nghiên cứu tổng thể thực chất là tìm các tham số đặc trưng của biến ngẫu nhiên đặc trưng cho tổng thể đó.

Thí dụ : khi nghiên cứu dân số Việt Nam ta có thể nghiên cứu một số dấu hiệu sau

-) X là thu nhập của người lao động Việt Nam
-) Y là tiêu dùng của người lao động Việt Nam
-) Z là giới tính của lao động Việt Nam.
-) .v.v.

+) Dấu hiệu nghiên cứu có thể là định lượng (thu nhập, tiêu dùng, năng suất của một giống lúa...) và có thể là định tính (giới tính, bậc học,sản phẩm tốt - xấu,...)

+) Nếu tổng thể với dấu hiệu nghiên cứu là định lượng thì ta quan tâm tới trung bình của tổng thể, ký hiệu là *m*, và phương sai (hay độ phân tán) của tổng thể ký hiệu là σ^2. Nếu tổng thể với dấu hiệu nghiên cứu là định tính thì ta quan tâm tới tham số *p* của tổng thể nó phản ánh cơ cấu của tổng thể

+) Để xác định các tham số *m*, σ^2, *p* của tổng thể người ta có thể dùng phương pháp điều tra toàn bộ tổng thể rồi sau đó xác định các tham số

trên của tổng thể, dựa vào các tham số đó chúng ta có thể đưa ra các đánh giá và phân tích tổng thể...
+) Phương pháp điều tra toàn bộ tổng thể ít được sử dụng vì có những hạn chế sau đây :
 -) Tốn kém về tài chính, mất nhiều thời gian
 -) Quy mô quá lớn dẫn đến không kiểm soát được
 -) Thông tin ban đầu thường kém tin cậy do người điều tra (khảo sát) và người trả lời không khớp
 -) Quy mô lớn dẫn đến có tính trùng hoặc bỏ sót phần tử (chẳng hạn do địa bàn hiểm trở gây khó khăn cho người điều tra)
 -) Trong nhiều trường hợp không thể có được danh sách toàn bộ tổng thể (chẳng hạn không thể có được danh sách chính xác những người nghiện ma túy)
+) Vì những lý do trên mà người ta đưa ra một phương pháp thường dùng trong thực tế đó là *phương pháp mẫu*
+) Phương pháp mẫu là phương pháp: Từ tổng thể ta lấy ra một số phần tử đại diện cho tổng thể, số phần tử lấy ra gọi là kích thước mẫu và ký hiệu là n trên cơ sở phân tích số liệu mẫu mà ta đưa ra các kết luận (hay suy đoán) về tổng thể
+) Phương pháp mẫu bao gồm các bước sau :
 Bước 1 : Từ tổng thể lấy ra một mẫu có kích thước n
 Bước 2 : Xác định các tham số đặc trưng mẫu, hay còn gọi là các thống kê mẫu
 Bước 3 : Xác định quy luật phân phối xác suất cho các thống kê mẫu
 Bước 4 : Xuất phát từ các thống kê mẫu ta đưa ra kết luận về các tham số của tổng thể

6.2. Tổng thể nghiên cứu

a) Định nghĩa : Toàn bộ tập hợp các phần tử đồng nhất theo một dấu hiệu nghiên cứu nào đó được gọi là tổng thể nghiên cứu hay tổng thể.
+) Dấu hiệu nghiên được đặc trưng bởi một biến ngẫu nhiên thì biến ngẫu nhiên đó gọi là biến ngẫu nhiên gốc
+) Số lượng các phần tử của tổng thể được gọi là kích thước của tổng thể và ký hiệu là **N**
+) **N** có thể là hữu hạn, có thể rất lớn, thậm chí là vô hạn
Thí dụ :
 -) Ta muốn nghiên cứu năng suất của một giống lúa của các tỉnh thuộc đồng bằng bắc bộ thì **N** là toàn bộ diện tích (ha) trồng giống lúa đó trên tất cả các tỉnh thuộc đồng bằng bắc bộ
 -) Ta muốn nghiên cứu toàn bộ những người nghiện ma túy trên toàn quốc thì **N** là tổng số tất cả những người nghiện ma túy của các tỉnh trên toàn quốc cộng lại, trong trường hợp này khó có được chính xác **N** là bao nhiêu
b) Mô tả tổng thể

+) Giả sử ta nghiên cứu tổng thể X với dấu hiệu nghiên cứu định lượng nhận các giá trị $x_1, x_2, \ldots, x_k$ với các tần số xuất hiện tương ứng là N_1, $N_2, \ldots, N_k$. Khi đó ta mô tả tổng thể bằng bảng phân phối tần số như sau :

Dấu hiệu nghiên cứu	x_1	x_2	....	x_k
Tần số	N_1	N_2	....	N_k

Trong đó :

$$\begin{cases} 0 \le N_i \le \mathrm{N} & \forall i = \overline{1;k} \\ \sum_{i=1}^{k} N_i = \mathrm{N} \end{cases}$$

+)Nếu đặt : $p_i = \frac{N_i}{\mathrm{N}}; \forall i = \overline{1;k}$ thì ta có bảng phân phối tần suất như sau

Dấu hiệu nghiên cứu	x_1	x_2	....	x_k
Tần suất	p_1	p_2	....	p_k

Trong đó :

$$\begin{cases} 0 \le p_i \le 1 & \forall i = \overline{1;k} \\ \sum_{i=1}^{k} p_i = 1 \end{cases}$$

+) Nếu đặt $\mathrm{W_i} = \sum_{x_j < x_i} N_j$ với $i; j = \overline{1;k}$ và $F(x_i) = \frac{\mathrm{W_i}}{\mathrm{N}} = \sum_{x_j < x_i} \frac{N_j}{\mathrm{N}}$

trong đó : $\mathrm{W_i}$ là tần số tích lũy theo x_i và $\mathrm{F}(x_i)$ là tần suất tích lũy theo x_i, $\mathrm{F}(x_i)$ là một hàm của x_i

+) Hàm $\mathrm{F}(x_i)$ có các tính chất giống như hàm phân bố xác suất $\mathrm{F}(x)$ của biến ngẫu nhiên rời rạc

c) <u>Các tham số đặc trưng của tổng thể</u>

+)Với dấu hiệu nghiên cứu định lượng có bảng phân phối tần số như trên ta có :

-) Trung bình *m* của tổng thể được xác định

$$m = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x_i = \frac{1}{\mathrm{N}}\sum_{i=1}^{k} N_i x_i$$

-) Phương sai (độ phân tán) σ^2 của tổng thể được xác định

$$\sigma^2 = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} (x_i - m)^2 = \frac{1}{\mathrm{N}}\sum_{i=1}^{k} N_i (x_i - m)^2$$

+) Với dấu hiệu nghiên cứu định tính, giả sử tổng thể có kích thước **N** trong đó có **M** phần tử của tổng thể mang dấu hiệu nghiên cứu, khi đó tham số *p* được xác định

$$p = \frac{\mathrm{M}}{\mathrm{N}}$$

6.3. Mẫu ngẫu nhiên

a) Định nghĩa : Mẫu ngẫu nhiên kích thước *n* là tập hợp *n* biến ngẫu nhiên độc lập $X_1, X_2,...,X_n$ được xây dựng (hay rút ra) từ biến ngẫu nhiên gốc X của tổng thể và có cùng phân phối xác suất với biến ngẫu nhiên gốc X.

+) Mẫu ngẫu nhiên kích thước *n* ký hiệu là $W_n = (X_1, X_2,...,X_n)$

+) Nếu biến ngẫu nhiên gốc $X \sim N(\mu, \sigma^2)$ thì $X_i \sim N(\mu, \sigma^2)$ $\forall i = \overline{1;n}$ đồng thời

$$E(X_i) = E(X) = \mu \text{ và } V(X_i) = V(X) = \sigma^2, \quad \forall i = \overline{1;n}$$

+) Nếu biến ngẫu nhiên gốc $X \sim A(p)$ thì $X_i \sim A(p)$ $\forall i = \overline{1;n}$ đồng thời

$$E(X_i) = E(X) = p \text{ và } V(X_i) = V(X) = p(1- p) \quad \forall i = \overline{1;n}$$

+) Nếu tiến hành một phép thử với mẫu ngẫu nhiên ta sẽ thu được một mẫu với các giá trị cụ thể (hay mẫu cụ thể) : $w_n = (x_1, x_2,..., x_n)$

b) Mô tả mẫu : Khi đã điều tra (khảo sát) được một mẫu đảm bảo tính đại diện cho tổng thể rồi thì ta dùng một trong các phương pháp sau để mô tả

+) Dùng đồ thị (hình đa giác, biểu đồ hình cột, hình hộp, hình bánh xe,.v.v.)

+) Dùng bảng phân phối tần số :

Dấu hiệu	x_1	x_2	....	x_k
Tần số	n_1	n_2	....	n_k

hoặc

Dấu hiệu	x_0 - x_1	x_1 - x_2	....	x_{k-1} - x_k
Tần số	n_1	n_2	....	n_k

Chú ý : ta ký hiệu x_{i-1} - x_i $\forall i = \overline{1;k}$ là các phần tử có dấu hiệu nghiên cứu thuộc đoạn giá trị $[x_{i-1}, x_i]$, nên muốn lấy giá trị đại diện cho đoạn đó thì ta lấy giá trị trung bình $\frac{x_{i-1} + x_i}{2}$.

+) Dùng bảng phân phối tần suất :

Dấu hiệu	x_1	x_2	....	x_k
Tần suất	f_1	f_2	....	f_k

Với $f_i = \frac{n_i}{n} \quad \forall i = \overline{1;k}$

6.4. Thống kê và các đặc trưng mẫu

6.4.1. Định nghĩa : Giả sử ta có tổng thể với dấu hiệu nghiên cứu được đặc trưng bởi biến ngẫu nhiên X. Để nghiên cứu về X thì từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2, ..., X_n)$.

Nếu chỉ dựa vào thông tin riêng lẻ của từng biến ngẫu nhiên $X_1, X_2, ..., X_n$ thì mới chỉ có được một vài kết luận sơ bộ và rời rạc về X. 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 quy luật mới làm cơ sở để nhận định về biến ngẫu nhiên X.

Việc tổng hợp mẫu $W_n = (X_1, X_2,..., X_n)$ được thực hiện dưới dạng một hàm nào đó đối với $X_1, X_2, ..., X_n$ của mẫu, thì hàm đó được gọi là *thống kê* và ký hiệu là G.

Như vậy $G = f(X_1, X_2,..., X_n)$, vì G là một hàm của các biến ngẫu nhiên do đó G cũng là một biến ngẫu nhiên, ta cần tìm quy luật phân phối xác suất và các tham số đặc trưng của thống kê G đó là E(G); V(G).

Khi mẫu ngẫu nhiên nhận một giá trị cụ thể $w_n = (x_1, x_2,..., x_n)$ thì G cũng nhận một giá trị cụ thể là $G_{qs} = f(x_1, x_2,..., x_n)$.

6.4.2. Một số thống kê đặc trưng của mẫu ngẫu nhiên

a) Trung bình mẫu

Giả sử từ tổng thể nghiên cứu với biến ngẫu nhiên gốc X ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2,..., X_n)$. Khi đó thống kê *trung bình mẫu* được xác định :

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

+) Nếu tổng thể nghiên cứu có trung bình là *m* và phương sai là σ^2 thì

$$E(\bar{X}) = m \text{ và } V(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow Se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

+) Nếu với mẫu cụ thể $w_n = (x_1, x_2,..., x_n)$ thì trung bình mẫu cũng nhận giá trị cụ thể, ta có :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{hay} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \quad \text{với} \quad \sum_{i=1}^{k} n_i = n$$

Chú ý : Các tham số trung vị (X_d), mốt (X_0), khoảng biến thiên ($R = X_{max} - X_{min}$), khoảng tứ phân vị yêu cầu người học tham khảo giáo trình NXB thống kê từ trang 292 đến trang 298

b) Độ lệch bình phương trung bình

Cho tổng thể với biến ngẫu nhiên gốc X, từ tổng thể rút ra một mẫu ngẫu nhiên $W_n = (X_1,X_2,...,X_n)$, khi đó *độ lệch bình phương trung bình* ký hiệu là MS được xác định như sau :

$$MS = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2 \quad \text{với} \quad \sum_{i=1}^{k} n_i = n$$

+) Với mẫu cụ thể $w_n = (x_1, x_2,..., x_n)$ ta có

$$\text{ms} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2$$

c) Phương sai mẫu và phương sai S^{*2}

+) Phương sai mẫu ký hiệu là S^2 được xác định như sau

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{hay} \quad S^2 = \frac{n}{n-1}MS$$

Với mẫu cụ thể $w_n = (x_1, x_2, ..., x_n)$ ta có

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{hay} \quad s^2 = \frac{n}{n-1}ms$$

+) Phương sai S^{*2} được xác định như sau

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$$

Với mẫu cụ thể ta có

$$s^{*2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

d) Tần suất mẫu

Giả sử ta có tổng thể với dấu hiệu nghiên cứu định tính có kích thước **N,** trong đó có **M** phần tử mang dấu hiệu nghiên cứu.

Từ tổng thể lấy ra một mẫu ngẫu nhiên có kích thước n

$$W_n = (X_1, X_2, ..., X_n)$$

Giả sử trong mẫu trên có X_A phần tử mang dấu hiệu nghiên cứu A, khi đó tần suất mẫu được xác định

$$f = \frac{X_A}{n}$$

Với mẫu cụ thể $w_n = (x_1, x_2, ..., x_n)$, giả sử có k phần tử mang dấu hiệu nghiên cứu A. Lúc đó giá trị tần suất mẫu được xác định :

$$f = \frac{k}{n}$$

Chú ý Nếu tổng thể với biến ngẫu nhiên gốc X ~ A(p)

=> E(X) = p và V(X) = $p(1 - p)$ thì với f là tần suất của mẫu ngẫu nhiên $W_n = (X_1, X_2, ..., X_n)$ ta có E(f) = p và V(f) = $\frac{p(1-p)}{n}$ do $X_A \sim B(n, p)$.

Thí dụ : hướng dẫn người học đọc hiểu các thí dụ 6,7,8 từ trang 304 đến trang 308 giáo trình.

6.5. Quy luật phân phối xác suất của các thống kê đặc trưng mẫu

Thông thường tổng thể nghiên cứu với biến ngẫu nhiên gốc X ta xét nó tuân theo quy luật phân phối chuẩn $N(\mu, \sigma^2)$ nếu là dấu hiệu định lượng và phân phối A(p) nếu là dấu hiệu định tính.

a) Giả sử biến ngẫu nhiên gốc X ~ $N(\mu, \sigma^2)$, thì từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2, ..., X_n)$, khi đó ta xác định được các thống kê $\bar{X}$ và S^2.

Khi đó ta có : E($\bar{X}$) = μ ; V($\bar{X}$) = $\frac{\sigma^2}{n}$ => $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

+) Nếu ta lập thống kê $U = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$

thì U là một biến ngẫu nhiên và U ~ N(0; 1)

Thật vậy:

$$E(U) = E\left(\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right) = \frac{\sqrt{n}}{\sigma}E\left(\bar{X}-\mu\right) = \frac{\sqrt{n}}{\sigma}[E(\bar{X})-\mu] = \frac{\sqrt{n}}{\sigma}[\mu-\mu]=0$$

$$V(U) = V\left(\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right) = \frac{n}{\sigma^2}V\left(\bar{X}-\mu\right) = \frac{n}{\sigma^2}V(\bar{X}) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1$$

+) Nếu ta lập thống kê $T = \frac{\bar{X}-\mu}{S}\sqrt{n}$

thì T là một biến ngẫu nhiên và T ~ T(n - 1)

+) Nếu ta lập thống kê $\chi^2 = \frac{nS^{*2}}{\sigma^2}$

thì χ^2 là một biến ngẫu nhiên và $\chi^2 \sim \chi^2(n)$

+) Nếu ta lập thống kê $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$

thì χ^2 là một biến ngẫu nhiên và $\chi^2 \sim \chi^2(n-1)$

b) Giả sử biến ngẫu nhiên gốc X ~ A(p), thì từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2, ..., X_n)$; $X_i = \{0;1\}\ \forall i = \overline{1;n}$, khi đó ta xác định được thống kê f

+) Nếu kích thước mẫu $n > 5$ và tham số p của tổng thể thỏa mãn

$$\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| . \frac{1}{\sqrt{n}} < 0,3$$

thì $f \sim N\left(p\,;\ \frac{p(1-p)}{n}\right)$

+) Nếu ta lập thống kê $U = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}}$ thì U là một biến ngẫu nhiên và

U ~ N(0; 1)

+) Nếu kích thước mẫu $n \geq 100$ và ta đặt $U = \frac{(f-p)\sqrt{n}}{\sqrt{f(1-f)}}$ thì U ~ N(0; 1)

c) Ta xét trường hợp có hai tổng thể với hai biến ngẫu nhiên gốc X_1, X_2 với $X_1 \sim N(\mu_1;\sigma_1^2)$ và $X_2 \sim N(\mu_2;\sigma_2^2)$:

+) Từ tổng thể X_1 ta rút ra một mẫu ngẫu nhiên $W_{n_1} = (X_{11}, X_{12}, ..., X_{1n_1})$ và từ mẫu này ta xác định được các thống kê mẫu

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \text{ và } MS_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2 \Rightarrow S_1^2 = \frac{n_1}{n_1-1}MS_1$$

+) Từ tổng thể X_2 ta rút ra một mẫu ngẫu nhiên $W_{n_2} = (X_{21}, X_{22}, ..., X_{2n_2})$ và từ mẫu này ta xác định được các thống kê mẫu

$$\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} \text{ và } MS_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2 \Rightarrow S_2^2 = \frac{n_2}{n_2 - 1} MS_2$$

+) Nếu ta lập thống kê $\bar{X}_1 - \bar{X}_2$ thì $\bar{X}_1 - \bar{X}_2$ là một biến ngẫu nhiên và

$$\left(\bar{X}_1 - \bar{X}_2\right) \sim N\left(\mu_1 - \mu_2 ;\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

+) Nếu ta lập thống kê $U = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ thì U ~ N(0; 1)

+) Nếu $n_1 > 30$; $n_2 > 30$ và ta lập thống kê

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \text{ thì T} \sim \text{N(0; 1)}$$

+) Nếu ta lập thống kê $F = \dfrac{S_1^2}{S_2^2}\dfrac{\sigma_2^2}{\sigma_1^2}$ thì F ~ F(n_1 - 1; n_2 - 2)

d) Ta xét trường hợp có hai tổng thể với hai biến ngẫu nhiên gốc X_1, X_2 với $X_1 \sim A(p_1)$ và $X_2 \sim A(p_2)$.

Từ tổng thể X_1 ta rút ra một mẫu ngẫu nhiên $W_{n_1} = (X_{11}, X_{12}, ..., X_{1n_1})$ với $X_{1i} = \{0;1\}\ \forall i = \overline{1; n_1}$, từ mẫu này ta xác định được thống kê mẫu f_1 . Từ tổng thể X_2 ta rút ra một mẫu ngẫu nhiên $W_{n_2} = (X_{21}, X_{22}, ..., X_{2n_2})$ với $X_{2i} = \{0;1\}\ \forall i = \overline{1; n_2}$, từ mẫu này ta xác định được thống kê mẫu f_2.

+) Nếu ta xét thống kê f_1 - f_2 thì f_1 - f_2 là một biến ngẫu nhiên hơn nữa

$$(f_1 - f_2) \sim N\left(p_1 - p_2;\ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

+) Nếu với $n_1 > 30$, $n_2 > 30$ và ta lập thống kê

$$U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \text{ thì U} \sim \text{N(0; 1).}$$

6.6. Suy diễn (suy đoán) thống kê

Giả sử tổng thể nghiên cứu với biến ngẫu nhiên gốc X. Giả sử đã biết quy luật phân phối xác suất của X và cũng biết các tham số đặc trưng của X . Từ tổng thể X ta rút ra một mẫu kích thước *n* là W_n, khi đó ta có thể suy đoán về các thống kê đặc trưng mẫu như thế nào ?. Đó là bài toán suy diễn thống kê, ta suy diễn một vài thống kê hay dùng trong thực tế

a) Suy đoán về giá trị trung bình mẫu

Giả sử tổng thể với biến ngẫu nhiên gốc X ~ N(μ, σ^2) đồng thời ta cũng giả sử μ và σ^2 đã biết. Từ tổng thể ta rút ra một mẫu $W_n = (X_1, X_2, ..., X_n)$, nếu ta không tính $\bar{X}$ bởi công thức $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ thì với mức xác suất 1 – α cho trước ta suy đoán xem $\bar{X}$ sẽ nằm trong khoảng nào ?

+) Ta đã biết thống kê $U = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ thì U ~ N(0; 1) do đó với mức xác suất 1 – α cho trước ta tìm được α_1, α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng phụ lục 6 tìm được các giá trị tới hạn chuẩn $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn :

$$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1-\alpha \iff P(-u_{\alpha_1} < U < u_{\alpha_2}) = 1-\alpha$$

$$\iff P\left(-u_{\alpha_1} < \frac{\bar{X}-\mu}{\sigma}\sqrt{n} < u_{\alpha_2}\right) = 1-\alpha \iff P\left(\mu - \frac{\sigma}{\sqrt{n}}u_{\alpha_1} < \bar{X} < \mu + \frac{\sigma}{\sqrt{n}}u_{\alpha_2}\right) = 1-\alpha$$

+) Thực tế ta hay xét ba trường hợp sau :

-) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ ta có $\mu - \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}} < \bar{X} < \mu + \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}$

-) $\alpha_1 = \alpha \Rightarrow \alpha_2 = 0$ ta có $\bar{X} > \mu - \frac{\sigma}{\sqrt{n}}u_{\alpha}$

-) $\alpha_1 = 0 \Rightarrow \alpha_2 = \alpha$ ta có $\bar{X} < \mu + \frac{\sigma}{\sqrt{n}}u_{\alpha}$

<u>Thí dụ</u> (người học đọc hiểu các thí dụ giáo trình)

b) <u>Suy đoán về giá trị phương sai mẫu</u>

Tương tự như suy đoán về trung bình mẫu, thì với mức xác suất 1 – α ta có thể suy đoán S^2 sẽ nằm trong khoảng nào ?

+) Ta đã biết thống kê $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ thì $\chi^2 \sim \chi^2(n-1)$ do đó với mức xác suất 1 – α cho trước ta tìm được α_1, α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng phụ lục 7 tìm được các giá trị tới hạn $\chi^2_{1-\alpha_1}(n-1)$ và $\chi^2_{\alpha_2}(n-1)$ thỏa mãn :

$$P\left(\chi^2_{1-\alpha_1}(n-1) < \chi^2 < \chi^2_{\alpha_2}(n-1)\right) = 1-\alpha$$

$$\iff P\left(\chi^2_{1-\alpha_1}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha_2}\right) = 1-\alpha$$

$$\iff P\left(\frac{\sigma^2}{n-1}\chi^2_{1-\alpha_1}(n-1) < S^2 < \frac{\sigma^2}{n-1}\chi^2_{\alpha_2}(n-1)\right) = 1-\alpha$$

+) Thực tế ta hay xét ba trường hợp sau :

-) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ thì ta có $\frac{\sigma^2}{n-1}\chi^2_{1-\frac{\alpha}{2}}(n-1) < S^2 < \frac{\sigma^2}{n-1}\chi^2_{\frac{\alpha}{2}}(n-1)$

-) $\alpha_1 = \alpha \Rightarrow \alpha_2 = 0$ thì ta có $\frac{\sigma^2}{n-1}\chi^2_{1-\alpha}(n-1) < S^2 < +\infty$

-) $\alpha_1 = 0 \Rightarrow \alpha_2 = \alpha$ thì ta có $0 < S^2 < \frac{\sigma^2}{n-1}\chi^2_{\alpha}(n-1)$

<u>Thí dụ</u> : (hướng dẫn người học làm bài tập 6.42 sách bài tập)

c) <u>Suy đoán về giá trị tần suất mẫu</u>

Giả sử biến ngẫu nhiên gốc X ~ A(p), thì từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2, ..., X_n)$; $X_i = \{0;1\}\ \forall i = \overline{1;n}$. Nếu ta biết tham số p của tổng thể thì với xác suất 1 – α cho trước ta sẽ suy đoán tần suất mẫu f sẽ nằm trong khoảng nào ?

+) Ta đã biết thống kê $U = \dfrac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}}$ thì U ~ N(0; 1) do đó với mức xác suất 1 – α cho trước ta tìm được α_1, α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng phụ lục 6 tìm được các giá trị tới hạn chuẩn $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn

$$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1-\alpha \quad \Leftrightarrow \quad P(-u_{\alpha_1} < U < u_{\alpha_2}) = 1-\alpha$$

$$\Leftrightarrow P\left(-u_{\alpha_1} < \frac{(f-p).\sqrt{n}}{\sqrt{p(1-p)}} < u_{\alpha_2}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(p - \frac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\alpha_1} < f < p + \frac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\alpha_2}\right) = 1-\alpha$$

+) Thực tế ta hay xét ba trường hợp sau :

-) $\alpha_1 = \alpha_2 = \dfrac{\alpha}{2}$ thì ta có $p - \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\frac{\alpha}{2}} < f < p + \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\frac{\alpha}{2}}$

-) $\alpha_1 = \alpha \Rightarrow \alpha_2 = 0$ thì ta có $f > p - \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\alpha}$

-) $\alpha_1 = 0 \Rightarrow \alpha_2 = \alpha$ thì ta có $f < p + \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}u_{\alpha}$

Thí dụ : (hướng dẫn người học làm bài tập 6.38 sách bài tập)

Chương VII : Ước lượng các tham số của biến ngẫu nhiên gốc

6.1. Khái niệm

Giả sử cần nghiên cứu tổng thể với dấu hiệu nào đó, 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 và cơ cấu của tổng thể theo dấu hiệu nghiên cứu. Nếu dấu hiệu nghiên cứu trong tổng thể được đặc trưng bởi một biến ngẫu nhiên X và giả sử đã biết dạng phân phối xác suất của X nhưng chưa biết các tham số đặc trưng của X. Như vậy việc xác định các tham số đặc trưng của tổng thể cần nghiên cứu được quy về bài toán xác định các tham số đặc trưng của biến ngẫu nhiên gốc X mà ta đã biết quy luật phân phối xác suất của X. Chẳng hạn nếu đã biết dấu hiệu nghiên cứu của tổng thể có thể xem như là biến ngẫu nhiên X ~ N(μ, σ^2) 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, thực chất chúng là trung bình và phương sai của tổng thể. Như vậy bài toán ước lượng tham số có thể phát biểu như sau : *Cho tổng thể với dấu hiệu nghiên cứu được đặc trưng bởi biến ngẫu nhiên X (ta gọi là biến ngẫu nhiên gốc), với giả thiết đã biết quy luật phân phối xác suất của X song chưa biết tham số θ nào đó của X. Ta cần ước lượng (hay xác định một cách gần đúng) giá trị θ.* Để ước lượng θ thì ta cần phải có thông tin, mà thông tin dựa vào đâu ? muốn có thông tin thì ta phải dựa vào mẫu. Từ tổng thể nghiên cứu ta rút ra một mẫu ngẫu nhiên có kích thước n dựa vào mẫu đó ta xây dựng một thống kê $\hat{\theta}$ dùng để ước lượng hay xác định một cách gần đúng tham số θ tương ứng của tổng thể. Có hai phương pháp sử dụng $\hat{\theta}$ để ước lượng θ đó là phương pháp *ước lượng điểm* và phương pháp *ước lượng bằng khoảng tin cậy*.

6.2. Phương pháp ước lượng điểm

Là phương pháp dùng một giá trị của thống kê mẫu thay cho một tham số tương ứng chưa biết của tổng thể.

6.2.1. Phương pháp hàm ước lượng

a) Khái niệm : Giả sử tổng thể với dấu hiệu nghiên cứu định lượng được đặc trưng bởi biến ngẫu nhiên X ~ N(μ, σ^2) thì ta cần ước lượng các tham số μ và σ^2, còn nếu tổng thể với dấu hiệu nghiên cứu định tính được đặc trưng bởi biến ngẫu nhiên X ~ A(p) thì ta cần ước lượng tham số p.

+) Các tham số μ, σ^2, p ta gọi chung là tham số θ của biến ngẫu nhiên gốc X

+) Các thống kê $\overline{X}, S^2, f$ ta gọi chung là các thống kê $\hat{\theta}$ của mẫu

+) Với mẫu ngẫu nhiên $W_n = (X_1, X_2,...,X_n)$

thì thống kê $\hat{\theta} = f(X_1, X_2,...,X_n)$ là một hàm của các biến ngẫu nhiên nên $\hat{\theta}$ được gọi là *hàm ước lượng* của θ.

Như vậy : $\overline{X}$ là ước lượng điểm của μ

S^2 là ước lượng điểm của σ^2

f là ước lượng điểm của p

Với mẫu cụ thể $w_n = (x_1, x_2,..., x_n)$ thì các giá trị $\overline{x}, s^2, f$ là các giá trị ước lượng điểm của các tham số μ, σ^2, p

b) Các tiêu chuẩn lựa chọn hàm ước lượng

+) Ước lượng không chệch : 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

+) Ước lượng hiệu quả nhất : Thống kê $\hat{\theta}$ 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.

Dựa vào bất đẳng thức Cramer – Rao để tìm ước lượng hiệu quả nhất

$$V(\hat{\theta}) \geq \frac{1}{nE\left[\dfrac{\partial \ln f(x,\theta)}{\partial \theta}\right]}$$

+) Ước lượng vững : 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

$$\lim_{n\to+\infty} P\left(\left|\hat{\theta}-\theta\right| < \varepsilon\right) = 1$$

Thí dụ : Làm bài 7.10 sách bài tập

6.2.2. Phương pháp ước lượng hợp lý tối đa

Giả sử ta đã biết quy luật phân phối xác suất của biến ngẫu nhiên gốc X dưới dạng hàm mật độ $f(x, \theta)$, ($f(x, \theta)$ cũng có thể là biểu thức xác suất nếu là biến ngẫu nhiên rời rạc). Khi đó để ước lượng các tham số đặc trưng của X bằng *phương pháp hợp lý tối đa* thì ta thực hiện các bước sau

Bước 1 : Lập mẫu ngẫu nhiên $W_n = (X_1, X_2,..., X_n)$

Bước 2 : Xây dựng một hàm với đối số θ tại một giá trị cụ thể của mẫu (ta gọi là hàm hợp lý)

$$L(x_1, x_2,..., x_n, \theta) = f(x_1, \theta).f(x_2, \theta).....f(x_n, \theta)$$

Bước 3 : Tìm hàm Ln(L) và $\dfrac{\partial Ln(L)}{\partial \theta}$

Bước 4 : Giải phương trình $\dfrac{\partial Ln(L)}{\partial \theta} = 0$ tìm được nghiệm

$$\theta = \hat{\theta} = f(x_1, x_2,..., x_n)$$

Bước 5 : Tìm $\dfrac{\partial^2 Ln(L)}{\partial \theta^2}$ và chứng tỏ $\left.\dfrac{\partial^2 Ln(L)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$

Khi đó $\hat{\theta}$ là ước lượng điểm của θ bằng phương pháp hợp lý tối đa

Thí dụ 1: Cho biến ngẫu nhiên gốc X có hàm mật độ

$$f(x) = \theta e^{-\theta x} \qquad \forall x > 0$$

Tìm $\hat{\theta}$ là ước lượng hợp lý tối đa của θ

Lời giải : (Ta thực hiện 5 bước đã trình bày ở trên)

Thí dụ 2: Cho biến ngẫu nhiên gốc X ~ N(μ, σ^2)

Tìm các ước lượng điểm của μ và σ^2 bằng phương pháp hợp lý tối đa

Lời giải : $\overline{X}$ và MS lần lượt là các ước lượng điểm của μ và σ^2 bằng phương pháp hợp lý tối đa

6.3. Phương pháp ước lượng bằng khoảng tin cậy

6.3.1.Khái niệm

+) Để ước lượng tham số θ của biến ngẫu nhiên gốc X bằng *phương pháp khoảng tin cậy*, thì từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2,..., X_n)$. Xuất phát từ thống kê $G = f(X_1, X_2,..., X_n)$ ta xây dựng một khoảng thống kê (G_1, G_2) sao cho khoảng thống kê này chứa tham số θ của tổng thể với một độ tin cậy (mức xác suất) cho trước.

+) Khoảng thống kê (G_1, G_2) được xây dựng từ G gọi là khoảng tin cậy của tham số θ nếu $P(G_1 < \theta < G_2) = 1- \alpha$

+) Giá trị 1- α được gọi là *độ tin cậy*, đặt $I = G_2 - G_1$ thì I được gọi là *độ dài khoảng tin cậy*, $\varepsilon = \frac{I}{2}$ gọi là *sai số* của ước lượng.

+) Với mẫu cụ thể $w_n = (x_1, x_2,..., x_n)$ thì ta có khoảng giá trị (G_{qs1}, G_{qs2}) chứa tham số θ của tổng thể với độ tin cậy 1- α

6.3.2. Ước lượng kỳ vọng toán của biến ngẫu nhiên gốc có phân phối chuẩn

(Ý nghĩa : tức là ta ước lượng tham số trung bình của tổng thể)

Giả sử ta có tổng thể với dấu hiệu nghiên cứu định lượng được đặc trưng bởi biến ngẫu nhiên $X \sim N(\mu, \sigma^2)$, tham số μ chưa biết, ta cần ước lượng μ bằng phương pháp khoảng tin cậy.

Từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2,..., X_n)$, từ mẫu W_n ta xác định các thống kê mẫu $\bar{X}$ và S^2. Để ước lượng μ bằng phương pháp khoảng tin cậy ta xét hai trường hợp

a) <u>Trường hợp 1</u> : Đã biết tham số σ^2 của biến ngẫu nhiên gốc X.

Ta chọn thống kê : $G = U = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$ thì $U \sim N(0; 1)$.

Với độ tin cậy 1 - α cho trước, ta tìm được α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng giá trị tới hạn chuẩn (phụ lục 6) tìm được các giá trị $u_{1-\alpha_1}, u_{\alpha_2}$ thỏa mãn

$$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) \Leftrightarrow P(-u_{\alpha_1} < U < u_{\alpha_2}) = 1 - \alpha$$

$$\Leftrightarrow P\left(-u_{\alpha_1} < \frac{\bar{X} - \mu}{\sigma}\sqrt{n} < u_{\alpha_2}\right) = 1 - \alpha \Leftrightarrow P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha_2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}u_{\alpha_1}\right) = 1 - \alpha \quad (1)$$

Công thức (1) đúng cho mọi cặp α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$, nhưng thực tế ta thường dùng ba trường hợp sau :

+) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ khi đó khoảng tin cậy đối xứng của μ là

$$\bar{X} - \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}$$

+) $\alpha_1 = \alpha; \alpha_2 = 0$ ta có $\mu < \bar{X} + \frac{\sigma}{\sqrt{n}}u_{\alpha}$ dùng để ước lượng giá trị trung bình tối đa

+) $\alpha_1 = 0; \alpha_2 = \alpha$ ta có $\mu > \bar{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha}$ dùng để ước lượng giá trị trung bình tối thiểu

Chú ý :

-) Trong trường hợp khoảng tin cậy đối xứng của μ ta có độ dài khoảng tin cậy

I = $2\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}$ và sai số của ước lượng $\varepsilon = \frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}$

-) Nếu cho trước độ dài khoảng tin cậy I_0 và với độ tin cậy 1 - α cho trước thì cần điều tra mẫu với kích thước tối thiểu n_0 bằng bao nhiêu để độ dài khoảng tin cậy của μ không vượt quá I_0.

Ta tìm n_0 qua bất đẳng thức :

$$2\frac{\sigma}{\sqrt{n_0}}u_{\frac{\alpha}{2}} \le I_0 \Leftrightarrow \sqrt{n_0} \ge 2\frac{\sigma}{I_0}u_{\frac{\alpha}{2}} \Leftrightarrow n_0 \ge 4\frac{\sigma^2}{I_0^2}[u_{\frac{\alpha}{2}}]^2$$

<u>Thí dụ</u> : (hướng dẫn người học đọc hiểu thí dụ 1; 2 giáo trình trang 360 - 362 và làm bài 7.24 sách bài tập)

b) <u>Trường hợp 2</u> : Chưa biết tham số σ^2 của biến ngẫu nhiên gốc X

Ta chọn thống kê : $G = T = \frac{\bar{X}-\mu}{S}\sqrt{n}$ khi đó $T \sim T(n-1)$

Với độ tin cậy 1 - α cho trước, ta tìm được α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng giá trị tới hạn Student (phụ lục 8) tìm được các giá trị $t_{1-\alpha_1}^{(n-1)}, t_{\alpha_2}^{(n-1)}$ thỏa mãn

$$P\left(t_{1-\alpha_1}^{(n-1)} < T < t_{\alpha_2}^{(n-1)}\right) = 1-(\alpha_1+\alpha_2) \Leftrightarrow P\left(-t_{\alpha_1}^{(n-1)} < \frac{\bar{X}-\mu}{S}\sqrt{n} < t_{\alpha_2}^{(n-1)}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\bar{X} - \frac{S}{\sqrt{n}}t_{\alpha_2}^{(n-1)} < \mu < \bar{X} + \frac{S}{\sqrt{n}}t_{\alpha_1}^{(n-1)}\right) = 1-\alpha \quad (2)$$

Công thức (2) đúng cho mọi cặp α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$, nhưng thực tế ta thường dùng ba trường hợp sau :

+) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ khi đó khoảng tin cậy đối xứng của μ là

$$\bar{X} - \frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}^{(n-1)} < \mu < \bar{X} + \frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}^{(n-1)}$$

+) $\alpha_1 = \alpha; \alpha_2 = 0$ ta có $\mu < \bar{X} + \frac{S}{\sqrt{n}}t_{\alpha}^{(n-1)}$ dùng để ước lượng giá trị trung bình tối đa

+) $\alpha_1 = 0; \alpha_2 = \alpha$ ta có $\mu > \bar{X} - \frac{S}{\sqrt{n}}t_{\alpha}^{(n-1)}$ dùng để ước lượng giá trị trung bình tối thiểu

Chú ý :

-) Trong trường hợp khoảng tin cậy đối xứng của μ ta có độ dài khoảng tin cậy

I = $2\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}^{(n-1)}$ và sai số của ước lượng $\varepsilon = \frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}^{(n-1)}$

-) Với một mẫu cụ thể đã điều tra (khảo sát) $w_n = (x_1, x_2, \ldots, x_n)$ thì

$$\mu \in \left(\bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}^{(n-1)} ;\quad \bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}^{(n-1)} \right)$$ với độ tin cậy 1 - α

Khi đó ta tính được độ dài khoảng tin cậy I = $2\frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}^{(n-1)}$. Nếu với độ tin cậy 1 - α và cho trước I_0 thì cần điều tra thêm tối thiểu bao nhiêu quan sát nữa để độ dài khoảng tin cậy của μ không vượt quá I_0.

Gọi m_0 là số quan sát cần điều tra thêm, khi đó ta tìm m_0 thông qua bất đẳng thức sau :

$$2\frac{s}{\sqrt{n+m_0}} t_{\frac{\alpha}{2}}^{(n-1)} \le I_0 \Leftrightarrow \sqrt{n+m_0} \ge 2\frac{s}{I_0} t_{\frac{\alpha}{2}}^{(n-1)} \Leftrightarrow m_0 \ge 4\frac{s^2}{I_0^2}[t_{\frac{\alpha}{2}}^{(n-1)}]^2 - n$$

Và lúc đó mẫu mới có kích thước là $n + m_0$ quan sát.

Thí dụ : Làm bài 7.19 sách bài tập

6.3.3. Ước lượng phương sai của biến ngẫu nhiên gốc có phân phối chuẩn

(Ý nghĩa : tức là ta ước lượng độ phân tán của tổng thể)

Giả sử tổng thể với dấu hiệu nghiên cứu định lượng được đặc trưng bởi biến ngẫu nhiên X ~ N(μ, σ^2), tham số σ^2 chưa biết, ta cần ước lượng σ^2 bằng phương pháp khoảng tin cậy.

Từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2,..., X_n)$, từ mẫu W_n ta xác định các thống kê mẫu $\bar{X}$ và S^2. Để ước lượng σ^2 bằng phương pháp khoảng tin cậy ta xét hai trường hợp

a) Trường hợp 1 : Đã biết tham số μ của biến ngẫu nhiên gốc X, như đã biết ta xác định được thống kê $S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ với $\sum_{i=1}^{k} n_i = n$. Khi đó ta chọn thống kê : $G = \chi^2 = \frac{nS^{*2}}{\sigma^2}$ thì $\chi^2 \sim \chi^2(n)$

Với độ tin cậy 1 - α cho trước, ta tìm được α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng giá trị tới hạn khi bình phương (phụ lục 7) tìm được các giá trị $\chi^2_{1-\alpha_1}(n), \chi^2_{\alpha_2}(n)$ thỏa mãn

$$P\left(\chi^2_{1-\alpha_1}(n) < \chi^2 < \chi^2_{\alpha_2}(n)\right) = 1 - (\alpha_1 + \alpha_2)$$

$$\Leftrightarrow P\left(\chi^2_{1-\alpha_1}(n) < \frac{nS^{*2}}{\sigma^2} < \chi^2_{\alpha_2}(n)\right) = 1-\alpha \quad \Leftrightarrow P\left(\frac{nS^{*2}}{\chi^2_{\alpha_2}(n)} < \sigma^2 < \frac{nS^{*2}}{\chi^2_{1-\alpha_1}(n)}\right) = 1-\alpha \quad (3)$$

Công thức (3) đúng cho mọi cặp α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$, nhưng thực tế ta thường dùng ba trường hợp sau :

+) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ khi đó ta có khoảng tin cậy hai phía của σ^2 là

$$\frac{nS^{*2}}{\chi^2_{\frac{\alpha}{2}}(n)} < \sigma^2 < \frac{nS^{*2}}{\chi^2_{1-\frac{\alpha}{2}}(n)}$$

+) $\alpha_1 = \alpha; \alpha_2 = 0$ ta có $0 < \sigma^2 < \frac{nS^{*2}}{\chi^2_{1-\alpha}(n)}$ dùng để ước lượng độ phân tán tối đa

+) $\alpha_1 = 0; \alpha_2 = \alpha$ ta có $\frac{nS^{*2}}{\chi^2_{\alpha}(n)} < \sigma^2 < +\infty$ dùng để ước lượng độ phân tán tối thiểu

b) <u>Trường hợp 2</u> : Chưa biết tham số μ của biến ngẫu nhiên gốc X. Khi đó ta chọn thống kê : $G = \chi^2 = \frac{(n-1)S^2}{\sigma^2}$ thì $\chi^2 \sim \chi^2(n-1)$

Với độ tin cậy 1 - α cho trước, ta tìm được α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng giá trị tới hạn khi bình phương (phụ lục 7) tìm được các giá trị $\chi^2_{1-\alpha_1}(n-1), \chi^2_{\alpha_2}(n-1)$ thỏa mãn

$$P\left(\chi^2_{1-\alpha_1}(n-1) < \chi^2 < \chi^2_{\alpha_2}(n-1)\right) = 1 - (\alpha_1 + \alpha_2)$$

$$\Leftrightarrow P\left(\chi^2_{1-\alpha_1}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha_2}(n-1)\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\frac{(n-1)S^2}{\chi^2_{\alpha_2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha_1}(n-1)}\right) = 1 - \alpha \qquad (4)$$

Công thức (4) đúng cho mọi cặp α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$, nhưng thực tế ta thường dùng ba trường hợp sau :

+) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ khi đó ta có khoảng tin cậy hai phía của σ^2 là

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}$$

+) $\alpha_1 = \alpha; \alpha_2 = 0$ ta có $0 < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}$ dùng để ước lượng độ phân tán tối đa (hay độ đồng đều tối thiểu)

+) $\alpha_1 = 0; \alpha_2 = \alpha$ ta có $\frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)} < \sigma^2 < +\infty$ dùng để ước lượng độ phân tán tối thiểu (hay độ đồng đều tối đa)

<u>Thí dụ</u> : Làm bài 7.53 sách bài tập.

6.3.4. Ước lượng tham số p của biến ngẫu nhiên gốc có phân phối không - một.

(Ý nghĩa : ước lượng tham số cơ cấu của tổng thể có dấu hiệu định tính)

Giả sử tổng thể với dấu hiệu nghiên cứu định tính được đặc trưng bởi biến ngẫu nhiên X ~ A(*p*), tham số *p* chưa biết, ta cần ước lượng *p* bằng phương pháp khoảng tin cậy.

Từ tổng thể ta rút ra một mẫu ngẫu nhiên $W_n = (X_1, X_2,..., X_n)$, với $n \geq 100$ và $X_i = \{0; 1\}$ $\forall i = \overline{1;n}$, từ mẫu W_n ta xác định được thống kê tần suất mẫu f. Khi đó ta chọn thống kê :

$$G = U = \frac{(f-p)\sqrt{n}}{\sqrt{f(1-f)}} \quad \text{thì } U \sim N(0; 1)$$

Với độ tin cậy 1 - α cho trước, ta tìm được α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$ và tra bảng giá trị tới hạn chuẩn (phụ lục 6) tìm được các giá trị $u_{1-\alpha_1}, u_{\alpha_2}$ thỏa mãn

$$P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1-(\alpha_1+\alpha_2) \Leftrightarrow P(-u_{\alpha_1} < U < u_{\alpha_2}) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(-u_{\alpha_1} < \frac{(f-p)\sqrt{n}}{\sqrt{f(1-f)}} < u_{\alpha_2}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_2} < p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha_1}\right) = 1-\alpha \qquad (5)$$

Công thức (5) đúng cho mọi cặp α_1, α_2 không âm sao cho $\alpha_1 + \alpha_2 = \alpha$, nhưng thực tế ta thường dùng ba trường hợp sau :

+) $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ khi đó khoảng tin cậy đối xứng của p là

$$f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\frac{\alpha}{2}} < p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\frac{\alpha}{2}}$$

+) $\alpha_1 = \alpha; \alpha_2 = 0$ ta có $p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha}$ dùng để ước lượng giá trị tỷ lệ tối đa

+) $\alpha_1 = 0; \alpha_2 = \alpha$ ta có $p > f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\alpha}$ dùng để ước lượng giá trị tỷ lệ tối thiểu

Chú ý :

-) Trong trường hợp khoảng tin cậy đối xứng của p ta có độ dài khoảng tin cậy

$I = 2\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\frac{\alpha}{2}}$ và sai số của ước lượng $\varepsilon = \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\frac{\alpha}{2}}$

-) Với một mẫu cụ thể đã điều tra (khảo sát) $w_n = (x_1, x_2,..., x_n)$, với $n \geq 100$ và $x_i = \{0; 1\}$ $\forall i = \overline{1;n}$, thì từ mẫu w_n ta xác định được tần suất mẫu f là một giá trị cụ thể và nếu cho trước độ tin cậy $1-\alpha$ thì

$$p \in \left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\frac{\alpha}{2}} ; \quad f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\frac{\alpha}{2}}\right)$$

với độ tin cậy 1 - α

Khi đó ta tính được độ dài khoảng tin cậy I = $2\frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{\frac{\alpha}{2}}$. Nếu với độ tin cậy 1 - α và cho trước I_0 thì cần điều tra thêm tối thiểu bao nhiêu quan sát nữa để độ dài khoảng tin cậy của *p* không vượt quá I_0.

Gọi m_0 là số quan sát cần điều tra thêm, khi đó ta tìm m_0 thông qua bất đẳng thức sau :

$$2\frac{\sqrt{f(1-f)}}{\sqrt{n+m_0}}u_{\frac{\alpha}{2}} \le I_0 \Leftrightarrow \sqrt{n+m_0} \ge 2\frac{\sqrt{f(1-f)}}{I_0}u_{\frac{\alpha}{2}} \Leftrightarrow m_0 \ge 4\frac{f(1-f)}{I_0^2}\left(u_{\frac{\alpha}{2}}\right)^2 - n$$

Và lúc đó mẫu mới có kích thước là $n + m_0$ quan sát.

<u>Thí dụ</u> : (hướng dẫn người học đọc hiểu thí dụ 9 giáo trình trang 384 - 385 và làm bài tập 7.46 sách bài tập).

Chương VIII : Kiểm định giả thuyết thống kê

8.1. Một số khái niệm

+) Giả thuyết thống kê là giả thuyết về

-) Tham số đặc trưng của biến ngẫu nhiên

-) Tính độc lập của các biến ngẫu nhiên

-) Dạng phân phối xác suất của biến ngẫu nhiên

+) Chỉ có giả thuyết thống kê mới kiểm định được

+) Giả thuyết ký hiệu là H_0, ứng với mỗi giả thuyết H_0 luôn tồn tại một giả thuyết đối ký hiệu là H_1. Khi đó chúng tạo thành một cặp giả thuyết thống kê.

<u>Thí dụ</u> :

1) H_0 : Nhu cầu X tuân theo quy luật phân phối chuẩn
   H_1 : Nhu cầu X không tuân theo quy luật phân phối chuẩn.

2) H_0 : Nhu cầu X và nhu cầu Y độc lập với nhau
   H_1 : Nhu cầu X và nhu cầu Y phụ thuộc nhau.

3) Một giống lúa trước đây có năng suất trung bình là 60(tạ / ha), nay ta nghi ngờ năng suất đó đã thay đổi, khi đó ta đưa ra cặp giả thuyết

H_0: $\mu = 60$; H_1: $\mu \neq 60$

Nếu ta nghi ngờ năng suất lúa đã giảm thì ta đưa ra cặp giả thuyết

H_0: $\mu = 60$; H_1: $\mu < 60$

Nếu ta cho rằng năng suất lúa đã tăng lên thì ta đưa ra cặp giả thuyết

H_0: $\mu = 60$; H_1: $\mu > 60$

+) Để kiểm định giả thuyết thống kê ta chọn thống kê G mà nếu giả thuyết H_0 là đúng thì quy luật phân phối xác suất của G sẽ biết và lúc đó G được gọi là tiêu chuẩn kiểm định giả thuyết thống kê

+) Giá trị của tiêu chuẩn kiểm định G được tính trên một mẫu cụ thể được gọi là giá trị quan sát của tiêu chuẩn kiểm định và ký hiệu là G_{qs}

+) Khi đã chọn được tiêu chuẩn kiểm định G thì các giá trị của nó có thể chia thành hai tập hợp không giao nhau đó là :

-) Miền bác bỏ giả thuyết H_0 (ký hiệu là W_α) bao gồm các giá trị của G mà tại đó giả thuyết H_0 bị bác bỏ

-) Miền không bác bỏ giả thuyết H_0 (ký hiệu là $\overline{W}_\alpha$) bao gồm các giá trị của G mà tại đó giả thuyết H_0 không bị bác bỏ

-) Điểm phân chia miền W_α và $\overline{W}_\alpha$ được gọi là giá trị tới hạn

+) Các bước xây dựng miền bác bỏ giả thuyết H_0 (ta gọi tắt là miền bác bỏ) W_α

Bước 1 : Xuất phát từ mức xác suất α (khá nhỏ) cho trước (thông thường α = 0,05) ta tìm miền W_α sao cho

$$P(G \in W_\alpha / H_0 \text{ đúng}) = \alpha$$

giá trị α gọi là *mức ý nghĩa* của kiểm định

Bước 2 : Nếu với mẫu cụ thể và với mức ý nghĩa α cho trước thì ta kết luận như sau

-) Nếu $G_{qs} \in W_\alpha$ thì ta bác bỏ giả thuyết H_0

-) Nếu $G_{qs} \notin W_\alpha$ thì ta chưa có cơ sở để bác bỏ giả thuyết H_0, thực tế ta chấp nhận H_0

+) Khi kiểm định một cặp giả thuyết thống kê ta có thể mắc sai lầm thuộc hai loại sau

-) Sai lầm loại 1 : Ta bác bỏ một giả thuyết đúng.

Với mức ý nghĩa α cho trước thỏa mãn $P(G \in W_\alpha / H_0 \text{ đúng}) = \alpha$ thì α chính là xác suất mắc sai lầm loại 1.

-) Sai lầm loại 2 : Ta thừa nhận một giả thuyết sai. Giả sử xác suất mắc sai lầm loại 2 là β thì

$$P(G \notin W_\alpha / H_1 \text{ đúng}) = \beta$$

Lúc đó xác suất không mắc sai lầm loại 2 là $P(G \in W_\alpha / H_1 \text{ đúng}) = 1 - \beta$, giá trị 1- β được gọi là *lực kiểm định* giả thuyết thống kê.

Người ta thường cho trước α và trong các miền W_α có thể lựa chọn, ta chọn miền W_α nào mà có β nhỏ nhất thì đó là miền bác bỏ tốt nhất

Việc chọn giá trị α tùy thuộc vào hậu quả mà sai lầm loại 1 và loại 2 mang lại.

8.2. Kiểm định tham số

8.2.1. Kiểm định giả thuyết về tham số μ của biến ngẫu nhiên gốc có phân phối chuẩn

(Ý nghĩa : Kiểm định giá trị trung bình của tổng thể)

Giả sử ta có tổng thể với dấu hiệu nghiên cứu định lượng được đặc trưng bởi biến ngẫu nhiên $X \sim N(\mu, \sigma^2)$. Ta chưa biết μ song có ý kiến cho rằng nó bằng μ_0 (μ_0 cho trước), khi đó ta đưa ra giả thuyết H_0: $\mu = \mu_0$

Để kiểm định giả thuyết trên thì từ tổng thể ta rút ra một mẫu ngẫu nhiên kích thước n : $W_n = (X_1, X_2, ..., X_n)$ và từ mẫu W_n ta xác định được các thống kê mẫu $\overline{X}, S^2$. Để chọn tiêu chuẩn kiểm định G thích hợp ta xét hai trường hợp

1) Trường hợp đã biết σ^2 của biến ngẫu nhiên gốc
Ta chọn tiêu chuẩn kiểm định

$$G = U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

Nếu giả thuyết H_0: $\mu = \mu_0$ là đúng thì $U \sim N(0; 1)$.
Với mức ý nghĩa α cho trước tùy thuộc vào giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :

+) Nếu cặp giả thuyết là H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}\ ,\quad |U| > u_{\frac{\alpha}{2}} \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu = \mu_0$; H_1: $\mu > \mu_0$ thì miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}\ ,\quad U > u_\alpha \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu = \mu_0$; H_1: $\mu < \mu_0$ thì miền bác bỏ bên trái giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}\ ,\quad U < -u_\alpha \right\}$$

Chú ý : Với mẫu cụ thể $w_n = (x_1, x_2,..., x_n)$ ta có $U_{qs} = \dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$

<u>Thí dụ</u> : Làm bài tập 8.5 sách bài tập
Gợi ý : kiểm định cặp giả thuyết H_0: $\mu = 453$; H_1: $\mu < 453$

2) Trường hợp chưa biết σ^2 của biến ngẫu nhiên gốc
Ta chọn tiêu chuẩn kiểm định

$$G = T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}$$

Nếu giả thuyết H_0: $\mu = \mu_0$ là đúng thì $T \sim T(n - 1)$.
Với mức ý nghĩa α cho trước tùy thuộc vào giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :

+) Nếu cặp giả thuyết là H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}\ ,\quad |T| > t^{(n-1)}_{\frac{\alpha}{2}} \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu = \mu_0$; H_1: $\mu > \mu_0$ thì miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}\,, \quad T > t_\alpha^{(n-1)} \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu = \mu_0$; H_1: $\mu < \mu_0$ thì miền bác bỏ bên trái giả thuyết H_0 là

$$W_\alpha = \left\{ T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}\,, \quad T < -t_\alpha^{(n-1)} \right\}$$

Chú ý : Với mẫu cụ thể $w_n = (x_1, x_2,..., x_n)$ ta có $T_{qs} = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$

<u>Thí dụ</u> : Làm bài tập 8.10 sách bài tập
Gợi ý : kiểm định cặp giả thuyết H_0: $\mu = 14$; H_1: $\mu \neq 14$

8.2.2. Kiểm định giả thuyết về tham số σ^2 của biến ngẫu nhiên gốc có phân phối chuẩn

(ý nghĩa : là kiểm định độ phân tán của tổng thể)

Giả sử ta có tổng thể với dấu hiệu nghiên cứu định lượng được đặc trưng bởi biến ngẫu nhiên $X \sim N(\mu, \sigma^2)$. Ta chưa biết σ^2 song có ý kiến cho rằng nó bằng σ_0^2 (σ_0^2 cho trước), khi đó ta đưa ra giả thuyết H_0: $\sigma^2 = \sigma_0^2$. Để kiểm định giả thuyết trên thì từ tổng thể ta rút ra một mẫu ngẫu nhiên kích thước n : $W_n = (X_1, X_2, ..., X_n)$ và từ mẫu W_n ta xác định được các thống kê mẫu $\overline{X}$, S^2

Ta chọn tiêu chuẩn kiểm định

$$G = \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

Nếu giả thuyết H_0: $\sigma^2 = \sigma_0^2$ là đúng thì thống kê $\chi^2 \sim \chi^2(n-1)$.

Với mức ý nghĩa α cho trước tùy thuộc vào giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :

+) Nếu cặp giả thuyết là H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 \neq \sigma_0^2$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}\,, \quad \left[\begin{array}{l} \chi^2 > \chi^2_{\frac{\alpha}{2}}(n-1) \\ 0 < \chi^2 < \chi^2_{1-\frac{\alpha}{2}}(n-1) \end{array}\right. \right\}$$

+) Nếu cặp giả thuyết là H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 > \sigma_0^2$ thì miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}\,, \quad \chi^2 > \chi^2_\alpha(n-1) \right\}$$

+) Nếu cặp giả thuyết là H_0: $\sigma^2 = \sigma_0^2$; H_1: $\sigma^2 < \sigma_0^2$ thì miền bác bỏ bên trái giả thuyết H_0 là

$$W_\alpha = \left\{ \chi^2 = \frac{(n-1)S^2}{\sigma_0^2}, \quad 0 < \chi^2 < \chi^2_{1-\alpha}(n-1) \right\}$$

Chú ý :

-)Với mẫu cụ thể $w_n = (x_1, x_2, ..., x_n)$ ta có $\chi^2_{qs} = \dfrac{(n-1)s^2}{\sigma_0^2}$

-) Khi nói độ đồng đều giảm sút tức là độ phân tán tăng lên, độ ổn định giảm sút tức là độ bất ổn định tăng lên, độ an toàn giảm sút tức là độ rủi ro tăng lên.

Thí dụ : Làm bài tập 8.46 sách bài tập
Gợi ý : kiểm định cặp giả thuyết H_0: $\sigma^2 = 10$; H_1: $\sigma^2 > 10$

8.2.3. Kiểm định giả thuyết về tham số p của biến ngẫu nhiên gốc có phân phối Không - Một

Giả sử ta có tổng thể với dấu hiệu nghiên cứu định tính được đặc trưng bởi biến ngẫu nhiên X ~ A(p).

Ta chưa biết p, song có ý kiến cho rằng $p = p_0$ (p_0 cho trước), khi đó ta đưa ra giả thuyết H_0: $p = p_0$

Để kiểm định giả thuyết trên thì từ tổng thể ta rút ra một mẫu ngẫu nhiên kích thước n

$W_n = (X_1, X_2, ..., X_n)$ với $X_i = \{0; 1\}$ $\forall i = \overline{1;n}$, và từ mẫu W_n ta xác định được thống kê tần suất mẫu f

Chọn tiêu chuẩn kiểm định

$$G = U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$$

Nếu giả thuyết H_0: $p = p_0$ là đúng thì U ~ N(0; 1).

Với mức ý nghĩa α cho trước, tùy thuộc giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :

+) Nếu cặp giả thuyết là H_0: $p = p_0$; H_1: $p \neq p_0$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}, \quad |U| > u_{\frac{\alpha}{2}} \right\}$$

+) Nếu cặp giả thuyết là H_0: $p = p_0$; H_1: $p > p_0$ thì ta có miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}, \quad U > u_\alpha \right\}$$

+) Nếu cặp giả thuyết là H_0: $p = p_0$; H_1: $p < p_0$ thì miền bác bỏ bên trái giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}, \quad U < -u_\alpha \right\}$$

Chú ý : Với mẫu cụ thể $w_n = (x_1, x_2,..., x_n)$ với $x_i = \{0; 1\}$ $\forall i = \overline{1;n}$ thì f là một giá trị cụ thể, khi đó ta có $U_{qs} = \dfrac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$

<u>Thí dụ</u> : Làm bài tập 8.29 sách bài tập
Gợi ý : kiểm định cặp giả thuyết H_0: $p = 0{,}03$; H_1: $p > 0{,}03$

8.2.4. Kiểm định giả thuyết về hai tham số p của hai biến ngẫu nhiên gốc có phân phối Không - Một

Giả sử ta có hai tổng thể với dấu hiệu nghiên cứu định tính được đặc trưng bởi hai biến ngẫu nhiên X_1 và X_2 với $X_1 \sim A(p_1)$, $X_2 \sim A(p_2)$. Ta chưa biết p_1, p_2 song có ý kiến cho rằng chúng bằng nhau, khi đó ta đưa ra giả thuyết H_0: $p_1 = p_2$.

Để kiểm định giả thuyết trên thì từ hai tổng thể X_1 và X_2 lần lượt ta rút ra hai mẫu ngẫu nhiên kích thước $n_1 > 100$, $n_2 > 100$

$W_{n_1} = (X_{11}, X_{12},..., X_{1n_1})$, $W_{n_2} = (X_{21}, X_{22},..., X_{2n_2})$ và từ hai mẫu W_{n_1}, W_{n_2} ta xác định được các thống kê mẫu (chú ý $X_{1i} = \{0; 1\}$ $\forall i = \overline{1;n_1}$, $X_{2j} = \{0; 1\}$ $\forall j = \overline{1;n_2}$) rồi từ hai mẫu W_{n_1}, W_{n_2} ta xác định được thống kê mẫu f_1 và f_2 .

Chọn tiêu chuẩn kiểm định

$$G = U = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$$

Nếu giả thuyết H_0: $p_1 = p_2$ là đúng (giả sử $p_1 = p_2 = p$ nào đó) thì thống kê $U \sim N(0; 1)$, vì p chưa biết mà $n_1 > 100$, $n_2 > 100$ nên ta có thể lấy xấp xỉ p bằng $\overline{f}$ với

$$\overline{f} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \frac{k_1 + k_2}{n_1 + n_2}$$

Với mức ý nghĩa α cho trước, tùy thuộc giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :

+) Nếu cặp giả thuyết là H_0: $p_1 = p_2$; H_1: $p_1 \neq p_2$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\overline{f}\left(1-\overline{f}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}},\quad |U| > u_{\frac{\alpha}{2}} \right\}$$

+) Nếu cặp giả thuyết là H_0: $p_1 = p_2$; H_1: $p_1 > p_2$ thì ta có miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\overline{f}\left(1-\overline{f}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} ,\quad U > u_\alpha \right\}$$

+) Nếu cặp giả thuyết là H_0: $p_1 = p_2$; H_1: $p_1 < p_2$ thì miền bác bỏ bên trái giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{f_1 - f_2}{\sqrt{\overline{f}\left(1-\overline{f}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} ,\quad U < -u_\alpha \right\}$$

<u>Thí dụ</u> : Làm bài tập 8.41 sách bài tập
Gợi ý : kiểm định cặp giả thuyết H_0: $p_1 = p_2$; H_1: $p_1 \neq p_2$

8.2.5. Kiểm định giả thuyết về hai tham số μ của hai biến ngẫu nhiên gốc có phân phối chuẩn

(Ý nghĩa : kiểm định sự bằng nhau của trung bình hai tổng thể)

Giả sử ta có hai tổng thể với dấu hiệu nghiên cứu định lượng được đặc trưng bởi hai biến ngẫu nhiên X_1 và X_2.
Giả sử ta đã biết $X_1 \sim N(\mu_1;\sigma_1^2)$ và $X_2 \sim N(\mu_2;\sigma_2^2)$ nhưng chưa biết μ_1 và μ_2 song có ý kiến cho rằng chúng bằng nhau, khi đó ta đưa ra giả thuyết

$$H_0 : \mu_1 = \mu_2$$

Để kiểm định giả thuyết trên thì từ hai tổng thể X_1 và X_2 lần lượt ta rút ra hai mẫu ngẫu nhiên kích thước n_1, n_2 :

$$W_{n_1} = (X_{11}, X_{12},..., X_{1n_1}),\ W_{n_2} = (X_{21}, X_{22},..., X_{2n_2})$$

và từ hai mẫu W_{n_1}, W_{n_2} ta xác định được các thống kê mẫu $\overline{X}_1,\ S_1^2$ và $\overline{X}_2,\ S_2^2$. Để chọn tiêu chuẩn kiểm định G thích hợp ta xét hai trường hợp

1) Trường hợp đã biết σ_1^2, σ_2^2 của hai biến ngẫu nhiên gốc
Ta chọn tiêu chuẩn kiểm định

$$G = U = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Nếu giả thuyết H_0: $\mu_1 = \mu_2$ là đúng thì $U \sim N(0; 1)$.
Với mức ý nghĩa α cho trước, tùy thuộc vào giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :
+) Nếu cặp giả thuyết là H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad |U| > u_{\frac{\alpha}{2}} \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 > \mu_2$ thì miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad U > u_\alpha \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$ thì miền bác bỏ bên trái giả thuyết H_0 là

$$W_\alpha = \left\{ U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad U < -u_\alpha \right\}$$

Chú ý : Với hai mẫu cụ thể $w_{n_1} = (x_{11}, x_{12}, ..., x_{1n_1})$ và $w_{n_2} = (x_{21}, x_{22}, ..., x_{2n_2})$ ta có

$$U_{qs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

2) Trường hợp chưa biết σ_1^2, σ_2^2 của hai biến ngẫu nhiên gốc (trường hợp này ta xét với $n_1 > 30$, $n_2 > 30$)

Ta chọn tiêu chuẩn kiểm định

$$G = T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Nếu giả thuyết H_0: $\mu_1 = \mu_2$ là đúng thì thống kê $T \sim N(0;1)$

(do $n_1 > 30$, $n_2 > 30$).

Với mức ý nghĩa α cho trước, tùy thuộc vào giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :

+) Nếu cặp giả thuyết là H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \quad |T| > u_{\frac{\alpha}{2}} \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 > \mu_2$ thì miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \quad T > u_\alpha \right\}$$

+) Nếu cặp giả thuyết là H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$ thì miền bác bỏ bên trái giả thuyết H_0 là

$$W_\alpha = \left\{ T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \quad T < -u_\alpha \right\}$$

Chú ý : Với hai mẫu cụ thể $w_{n_1} = (x_{11}, x_{12}, ..., x_{1n_1})$ và $w_{n_2} = (x_{21}, x_{22}, ..., x_{2n_2})$ ta có

$$T_{qs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Thí dụ : Làm bài tập 8.19 sách bài tập.
Gợi ý : kiểm định cặp giả thuyết H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$

8.2.6. Kiểm định giả thuyết về hai tham số σ^2 của hai biến ngẫu nhiên gốc có phân phối chuẩn

(Ý nghĩa : kiểm định sự bằng nhau về độ phân tán của hai tổng thể)

Giả sử ta có hai tổng thể với dấu hiệu nghiên cứu định lượng được đặc trưng bởi hai biến ngẫu nhiên X_1 và X_2. Giả sử ta đã biết X_1 ~ N($\mu_1;\sigma_1^2$) và X_2 ~ N($\mu_2;\sigma_2^2$) nhưng chưa biết σ_1^2 và σ_2^2 song có ý kiến cho rằng chúng bằng nhau khi đó ta đưa ra giả thuyết H_0 : $\sigma_1^2 = \sigma_2^2$. Để kiểm định giả thuyết trên thì từ hai tổng thể X_1 và X_2 lần lượt ta rút ra hai mẫu ngẫu nhiên kích thước n_1, n_2 : $W_{n_1} = (X_{11}, X_{12}, ..., X_{1n_1})$, $W_{n_2} = (X_{21}, X_{22}, ..., X_{2n_2})$ và từ hai mẫu W_{n_1}, W_{n_2} ta xác định được các thống kê mẫu $\overline{X}_1, S_1^2$ và $\overline{X}_2, S_2^2$.
Ta chọn tiêu chuẩn kiểm định

$$G = F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2}$$

Nếu giả thuyết H_0 là đúng thì F ~ F(n_1 - 1; n_2 - 1).
Với mức ý nghĩa α cho trước tùy thuộc vào giả thuyết đối H_1 mà ta xây dựng miền bác bỏ W_α tốt nhất tương ứng với các trường hợp sau :

+) Nếu cặp giả thuyết là H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 \neq \sigma_2^2$ thì miền bác bỏ hai phía giả thuyết H_0 là

$$W_\alpha = \left\{ F = \frac{S_1^2}{S_2^2},\quad \left[\begin{array}{l} F > f_{\frac{\alpha}{2}}^{(n_1-1;\, n_2-1)} \\ 0 < F < f_{1-\frac{\alpha}{2}}^{(n_1-1;\, n_2-1)} \end{array}\right. \right\}$$

+) Nếu cặp giả thuyết là H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 > \sigma_2^2$ thì miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ F = \frac{S_1^2}{S_2^2},\quad F > f_\alpha^{(n_1-1;\, n_2-1)} \right\}$$

+) Nếu cặp giả thuyết là H_0: $\sigma_1^2 = \sigma_2^2$; H_1: $\sigma_1^2 < \sigma_2^2$ thì miền bác bỏ bên phải giả thuyết H_0 là

$$W_\alpha = \left\{ F = \frac{S_1^2}{S_2^2},\quad 0 < F < f_{1-\alpha}^{(n_1-1;\, n_2-1)} \right\}$$

Với mẫu cụ thể ta có $F_{qs} = \frac{s_1^2}{s_2^2}$.

<u>Thí dụ</u> : Làm bài 8.49
Gợi ý : kiểm định cặp giả thuyết H_0: $\sigma_A^2 = \sigma_B^2$; H_1: $\sigma_A^2 \neq \sigma_B^2$

8.3.Kiểm định phi tham số

8.3.1. 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 $A_1, A_2,..., A_h$, dấu hiệu B có các phạm trù $B_1, B_2,..., B_k$. Nếu có ý kiến cho rằng A và B độc lập, khi đó ta đưa ra cặp giả thuyết

H_0: A và B độc lập
H_1: A và B phụ thuộc

Để kiểm định cặp giả thuyết trên thì từ tổng thể lập mẫu kích thước *n* và trình bày các số liệu mẫu dưới dạng bảng sau.

B \ A	B_1	B_2	...	B_k	Σ
A_1	n_{11}	n_{12}	...	n_{1k}	$\boldsymbol{n}_1$
A_2	n_{21}	n_{22}	...	n_{2k}	$\boldsymbol{n}_2$
. . .	. . .	. . .	...	. . .	. . .
A_h	n_{h1}	n_{h2}	...	n_{hk}	$\boldsymbol{n}_h$
Σ	$\boldsymbol{m}_1$	$\boldsymbol{m}_2$	...	$\boldsymbol{m}_k$	$\Sigma = \mathbf{n}$

Trong đó

n là kích thước mẫu

n_{ij} là tần số ứng với các phần tử đồng thời mang dấu hiệu A_i và B_j

$\boldsymbol{n}_i$ là tổng các tần số ứng với thành phần A_i

$\boldsymbol{m}_j$ là tổng các tần số ứng với thành phần B_j

Tiêu chuẩn kiểm định được chọn là

$$\chi^2 = n\left[\left(\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i \times m_j}\right)-1\right]$$

Nếu giả thuyết H_0 là đúng và với n khá lớn thì $\chi^2 \sim \chi^2([h-1]\times[k-1])$.

Do vậy với mức ý nghĩa α cho trước ta có miền bác bỏ giả thuyết H_0 là

$$W_\alpha = \left\{\chi^2 ;\quad \chi^2 > \chi^2_\alpha([h-1]\times[k-1])\right\}$$

<u>Thí dụ</u> : Làm bài tập 8.55 sách bài tập

<u>Lời giải</u> : bài 8.55

Ta kiểm định cặp giả thuyết

H_0: Quy mô của công ty và hiệu quả của quảng cáo là độc lập

H_1: Quy mô của công ty và hiệu quả của quảng cáo là phụ thuộc

Tiêu chuẩn kiểm định là

$$\chi^2 = n\left[\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{n_{ij}^2}{n_i \times m_j}\right)-1\right].$$

Với mẫu cụ thể ta có

$$\chi^2_{qs} = 356\left[\left(\frac{20^2}{140.104}+\frac{52^2}{131.104}+\cdots+\frac{25^2}{85.124}\right)-1\right]$$

$$= 356(1{,}08325252-1) = 29{,}6379.$$

Với mức ý nghĩa α = 0,1 ta có miền bác bỏ giả thuyết H_0 là

$$W_\alpha = (\chi^2_{0,1}[(3\text{-}1).(3\text{-}1)];+\infty) = (\chi^2_{0,1}(4);+\infty) = (7{,}779;+\infty)$$

$\Rightarrow \chi^2_{qs} \in W_\alpha \Rightarrow$ Bác bỏ giả thuyết H_0, vậy quy mô của công ty và hiệu quả của quảng cáo là phụ thuộc.

8.3.2. Kiểm định giả thuyết về dạng phân phối xác suất của biến ngẫu nhiên gốc

Giả sử chưa biết quy luật phân phối xác suất của biến ngẫu nhiên gốc X của tổng thể nghiên cứu. Song có cơ sở để giả thiết rằng X phân phối theo quy luật N(μ, σ^2). Lúc đó ta đưa ra cặp giả thuyết thống kê

H_0: X có phân phối theo quy luật chuẩn
H_1: X không phân phối theo quy luật chuẩn

Để kiểm định cặp giả thuyết trên thì từ tổng thể ta lấy ra mẫu

$$W_n = (X_1, X_2, \ldots, X_n)$$

từ mẫu W_n ta xác định được các thống kê sau

- Phương sai và độ lệch chuẩn

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \Rightarrow \quad S_X = \sqrt{S_X^2}$$

- Hệ số bất đối xứng S (Skewness)

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{S_X^3}$$

- Hệ số nhọn K (Kurtosis)

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{S_X^4}$$

Để kiểm định cặp giả thuyết trên ta chọn tiêu chuẩn kiểm định **Jarque – Bera**

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right]$$

Nếu giả thuyết H_0 là đúng và với *n* đủ lớn, (thông thường *n* > 100) thì

$$JB \sim \chi^2(2)$$

Do vậy với mức ý nghĩa α cho trước ta có miền bác bỏ giả thuyết H_0 là

$$W_\alpha = \left(\chi_\alpha^2(2); \quad +\infty\right)$$

Với mẫu cụ thể mà $\qquad JB_{qs} \in W_\alpha$

thì bác bỏ H_0, kết luận X không có phân phối chuẩn