

# Phần 2. Xác suất

## Chương 1

### Khái niệm và các phép toán

#### 1.1. Giải tích tổ hợp

##### 1.1.1. Quy tắc cộng, quy tắc nhân

###### a. Quy tắc cộng

Nếu một công việc được chia ra thành  $k$  trường hợp để thực hiện, trường hợp một có  $n_1$  cách thực hiện xong công việc, trường hợp hai có  $n_2$  cách thực hiện xong công việc, . . . , trường hợp  $k$  có  $n_k$  cách thực hiện xong công việc và không có một cách thực hiện nào ở trường hợp này lại trùng với một cách thực hiện ở trường hợp khác.

Khi đó ta có:  $n = n_1 + n_2 + \dots + n_k$  cách thực hiện công việc.

###### b. Quy tắc nhân

Giả sử một công việc nào đó được chia thành  $k$  giai đoạn. Có  $n_1$  cách thực hiện giai đoạn thứ nhất,  $n_2$  cách thực hiện giai đoạn thứ hai, . . . ,  $n_k$  cách thực hiện giai đoạn thứ  $k$ .

Khi đó ta có:  $n = n_1.n_2..n_k$  cách thực hiện công việc.

**Ví dụ 1.1.1.** Giả sử để đi từ A đến C ta bắt buộc phải đi qua điểm B. Có 3 đường khác nhau để đi từ A đến B và có 2 đường khác nhau để đi từ B đến C.

Vậy có  $n = 3.2 = 6$  cách khác nhau để đi từ A đến C

##### 1.1.2. Chỉnh hợp

###### a. Chỉnh hợp không lặp

**Định nghĩa 1.1.2.** Một chỉnh hợp chập  $k$  của  $n$  phần tử là một nhóm sắp thứ tự gồm  $k$  phần tử khác nhau lấy từ  $n$  phần tử đã cho ( $k \leq n$ ).

**Ký hiệu và công thức:**  $A_n^k = \frac{n!}{(n-k)!} = n(n-1)\dots(n-k+1)$

Như vậy: Hai chỉnh hợp chập  $k$  của  $n$  phần tử khác nhau nếu:

- Hoặc chúng có ít nhất một phần tử khác nhau;
- Hoặc chúng gồm  $k$  phần tử như nhau nhưng sắp xếp theo thứ tự khác nhau.

**Ví dụ 1.1.3.** Một lớp phải học 10 môn, mỗi ngày phải học 2 môn. Hỏi có bao nhiêu cách sắp xếp thời khóa biểu trong một ngày.

*Giải:* Vì mỗi cách sắp xếp thời khóa biểu trong ngày là việc ghép hai môn trong số 10 môn, các cách sắp xếp này khác nhau do có ít nhất một môn khác nhau hoặc chỉ do thứ tự sắp xếp trước sau giữa hai môn. Vì thế mỗi cách sắp xếp ứng với một chỉnh hợp chập 2 từ 10 phần tử.

Tức là có:  $A_{10}^2 = 10 \cdot 9 = 90$  (cách).

### **b. Chỉnh hợp có lặp**

**Định nghĩa 1.1.4.** Một chỉnh hợp có lặp chập  $k$  của  $n$  phần tử là một nhóm sắp thứ tự gồm  $k$  phần tử lấy từ  $n$  phần tử đã cho trong đó mỗi phần tử có thể có mặt đến  $n$  lần trong nhóm tạo thành.

*Ký hiệu và công thức:*  $\overline{A_n^k} = n^k$

**Ví dụ 1.1.5.** Để đăng ký xe máy người ta dùng 4 chữ số từ 0,1,...,9 cho một sêri. Hỏi mỗi sêri có thể đăng ký được bao nhiêu xe.

*Giải:* Số xe máy được đăng ký trong một sêri chính là chỉnh hợp có lặp chập 4 của 10 (trừ đi 1 vì trong thực tế không dùng 4 chữ số 0)

$$\overline{A_{10}^4} - 1 = 10^4 - 1 = 9999 \text{ (xe).}$$

**Ví dụ 1.1.6.** Để truyền tin bằng tín hiệu mooc-xơ gồm hai kí hiệu chấm (.) và vạch (-), người ta mã hóa mỗi chữ cái của bảng chữ cái thành một nhóm có thứ tự gồm không quá 4 kí hiệu. Biết rằng cùng một kí hiệu có thể có mặt nhiều lần trong nhóm có thứ tự tạo thành. Hỏi có thể mã hóa được bao nhiêu chữ cái ?

*Giải:* Một nhóm có thứ tự gồm  $k$  kí hiệu ( $1 \leq k \leq 4$ ) tạo nên chính là một chỉnh hợp lặp chập  $k$  từ 2 phần tử đã cho. Vì vậy số chữ cái mã hóa được là:

$$\overline{A_2^1} + \overline{A_2^2} + \overline{A_2^3} + \overline{A_2^4} = 1^2 + 2^2 + 2^3 + 2^4 = 30$$

Như vậy nếu bảng chữ cái của một thứ tiếng nào đó gồm không quá 30 chữ thì ta có thể mã hóa theo cách trên.

### **1.1.3 Hoán vị**

**Định nghĩa 1.1.7.** Hoán vị là một chỉnh hợp không lặp chập  $n$  của  $n$  phần tử. Hay hoán vị của  $n$  phần tử là một nhóm có thứ tự gồm đủ mặt  $n$  phần tử đã cho.

Vậy các hoán vị của  $n$  phần tử chỉ khác nhau bởi thứ tự sắp xếp giữa các phần tử đó.

*Kí hiệu và công thức:*  $P_n = n(n-1)(n-2)\dots 1 = n!$

**Ví dụ 1.1.8.** Có bao nhiêu cách xếp 5 người ngồi vào một chiếc ghế dài gồm 5 chỗ.

*Giải:* Số cách xếp 5 người vào một ghế dài gồm 5 chỗ chính là hoán vị của 5 phần tử, nên ta có

$$P_5 = 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120 \text{ (cách).}$$

**Ví dụ 1.1.9.** Có bao nhiêu cách xếp  $n$  đại biểu ngồi quanh một bàn tròn.

*Giải:* Do các chỗ ngồi quanh một bàn tròn không có phần tử thứ nhất và phần tử cuối cùng nên đại biểu thứ nhất được ngồi tự do. Các đại biểu còn lại có số cách chọn vị trí ngồi lần lượt là:  $(n-1), (n-2), \dots, 1$ .

Vậy cách xếp  $n$  đại biểu ngồi quanh bàn tròn là:  $(n-1)!$  (cách).

#### 1.1.4. Tổ hợp

**Định nghĩa 1.1.10.** Một tổ hợp chập  $k$  của  $n$  phần tử ( $k \leq n$ ) là một nhóm không phân biệt thứ tự gồm  $k$  phần tử khác nhau lấy ra từ  $n$  phần tử đã cho.

**Ký hiệu và công thức:**  $C_n^k = \frac{n!}{k!(n-k)!}$

Vậy: Mỗi tổ hợp gồm các phần tử khác nhau, hai tổ hợp khác nhau là do các phần tử chứa trong chúng khác nhau chứ không phải là do bộ khác nhau về thứ tự của các phần tử trong đó.

**Một vài tính chất của tổ hợp:**

$$+ C_n^k = C_n^{n-k}$$

$$+ C_n^0 = C_n^n = 1 \quad (\text{Quy ước: } 0! = 1)$$

$$+ C_n^1 = C_n^{n-1} = n$$

$$+ C_{n+1}^k = C_n^k + C_n^{k-1}$$

**Ví dụ 1.1.11.** 10 đội bóng thi đấu với nhau theo thể thức đấu vòng. Hỏi phải tổ chức bao nhiêu trận đấu?

*Giải:* Mỗi trận đấu ứng với một nhóm gồm 2 phần tử từ 10 đội (không phân biệt thứ tự). Vì vậy phải tổ chức tất cả:

$$C_{10}^2 = \frac{10 \cdot 9}{2} = 45 \text{ trận đấu.}$$

**Chú ý:**

\* Để nhận dạng một hoán vị của  $n$  phần tử ta thường dùng các dấu hiệu đặc trưng sau:

- +  $n$  phần tử đều phải có mặt;
- + Mỗi phần tử chỉ xuất hiện đúng một lần;
- + Có thứ tự giữa các phần tử.

\* Ta sử dụng khái niệm chỉnh hợp khi gặp tình huống:

- + Phải chọn  $k$  phần tử từ  $n$  phần tử;
- + Sắp thứ tự  $k$  phần tử đó.

\* Ta sử dụng khái niệm tổ hợp khi gặp tình huống:

- + Cần chọn ra từ một tập có  $n$  phần tử một tập con có  $k$  phần tử;
- + Lưu ý rằng trong tập con  $k$  phần tử đó ta không quan tâm đến thứ tự của các phần tử.

## 1.2. Phép thử và biến cố

### 1.2.1. Phép thử và biến cố

**Định nghĩa 1.2.1.** Khi thực hiện một nhóm các điều kiện nào đó ta nói rằng đã thực hiện một *phép thử*. Hiện tượng được xét trong phép thử gọi là *biến cố* (hay *sự kiện*)

**Ví dụ 1.2.2.** Tung một con xúc xắc là thực hiện một phép thử. Hiện tượng xúc xắc xuất hiện mặt 3 chấm; xúc xắc xuất hiện mặt 6 chấm; xúc xắc xuất hiện mặt có số chấm lớn hơn 6; xúc xắc xuất hiện mặt có số chấm nhỏ hơn 7 là các biến cố.

**Ví dụ 1.2.3.** Tung một đồng xu là thực hiện một phép thử. Hiện tượng: đồng xu xuất hiện mặt sấp; đồng xu xuất hiện mặt ngửa là các biến cố.

**Phân loại phép thử:** 2 loại

+ *Phép thử lặp:* Là phép thử được thực hiện trong những điều kiện như nhau.

+ *Phép thử không lặp:* Là những phép thử được thực hiện trong những điều kiện khác nhau.

**Phân loại biến cố:** 3 loại

+ *Biến cố ngẫu nhiên:* là biến cố có thể xảy ra hoặc không xảy ra khi thực hiện một phép thử.

**Ký hiệu:** A, B, C,...

**Ví dụ 1.2.4.** Trong ví dụ 1.2.2. ở trên, biến cố “xúc xắc xuất hiện mặt ba chấm”, “xúc xắc xuất hiện mặt 6 chấm” là các biến cố ngẫu nhiên.

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

**Ký hiệu:**  $\Omega$  (hoặc  $U$ )

**Ví dụ 1.2.5.** Trong ví dụ 1.2.2. ở trên, biến cố “xúc xắc xuất hiện mặt có số chấm nhỏ hơn 7” là biến cố chắc chắn.

+ *Biến cố không thể có:* là biến cố nhất định không xảy ra khi thực hiện một phép thử.

**Ký hiệu:**  $\emptyset$  (hoặc  $V$ )

**Ví dụ 1.2.6.** Trong ví dụ 1.2.2. ở trên, biến cố “xúc xắc xuất hiện mặt có số chấm lớn hơn 6” là biến cố không thể có.

## 1.2.2. Quan hệ giữa các biến cố

### a. Hợp (tổng) của các biến cố

**Định nghĩa 1.2.7.** Biến cố A được gọi là *hợp* của các biến cố  $A_1, A_2, \dots, A_n$  nếu A xảy ra khi và chỉ khi ít nhất một trong các biến cố  $A_i$  ( $i = 1, \dots, n$ ) xảy ra. Và ta viết:

$$A = A_1 \cup A_2 \cup \dots \cup A_n$$

**Ví dụ 1.2.8.** Tung một con xúc xắc, gọi  $A_i$  là biến cố “xuất hiện mặt  $i$  chấm” ( $i=1,2,\dots,6$ ); gọi A là biến cố “xuất hiện mặt có số chấm chẵn”. Khi đó ta có  $A = A_2 \cup A_4 \cup A_6$ .

### b. Giao (tích) của các biến cố

**Định nghĩa 1.2.9.** Biến cố B được gọi là *giao* của các biến cố  $A_1, A_2, \dots, A_n$  nếu B xảy ra khi và chỉ khi tất cả các biến cố  $A_i$  ( $i = 1, \dots, n$ ) xảy ra. Và ta viết:

$$B = A_1 \cap A_2 \cap \dots \cap A_n$$

**Ví dụ 1.2.10.** Một mạch điện gồm 2 bóng đèn mắc song song. Gọi A là biến cố “bóng thứ nhất bị cháy khi điện quá tải”; B là biến cố “bóng thứ 2 bị cháy khi điện

quá tải”; C là biến cố “mạch điện bị ngắt khi điện quá tải”, vậy  $C = A \cap B$  và C chỉ xảy ra khi cả 2 biến cố A và B cùng đồng thời xảy ra.

### c. Biến cố xung khắc

**Định nghĩa 1.2.11.** Hai biến cố A và B được gọi là *xung khắc* với nhau nếu chúng không đồng thời xảy ra trong một phép thử.

*Như vậy, nếu A và B xung khắc thì:  $A \cap B = V$*

**Ví dụ 1.2.12.** Tung một con xúc xắc, gọi A là biến cố “xuất hiện mặt 2 chấm”; B là biến cố “xuất hiện mặt 3 chấm”, khi đó A và B là xung khắc nhau.

### d. Biến cố đối lập

**Định nghĩa 1.2.13.** Biến cố không xảy ra biến cố A được gọi là biến cố *đối lập* của A. Kí hiệu biến cố đối lập của biến cố A là  $\bar{A}$

Như vậy ta có: 
$$\begin{cases} A \cup \bar{A} = U \\ A \cap \bar{A} = V \end{cases}$$

**Ví dụ 1.2.14.** Bắn một viên đạn vào bia, biến cố “bắn trúng bia” và biến cố “bắn trượt bia” là hai biến cố đối lập.

### e. Hệ đầy đủ các biến cố

**Định nghĩa 1.2.15.** Các biến cố  $A_1, A_2, \dots, A_n$  được gọi là *hệ đầy đủ* các biến cố nếu trong kết quả của phép thử sẽ xảy ra một và chỉ một trong các biến cố đó.

Nghĩa là ta có 
$$\begin{cases} A_i A_j = V & (\forall i \neq j) \\ \sum_{i=1}^n A_i = U \end{cases}$$

**Ví dụ 1.2.16.** Gieo một con xúc xắc. Gọi  $A_i$  là biến cố “xúc xắc xuất hiện mặt i chấm” ( $i=1,2,\dots,6$ ) thì các biến cố  $A_1; A_2; A_3; A_4; A_5; A_6$  tạo nên một hệ đầy đủ các biến cố.

Nếu khả năng xảy ra các biến cố đó là như nhau ta gọi là hệ đầy đủ đồng khả năng.

## 1.2.3. Biến cố sơ cấp và không gian các biến cố sơ cấp

Một biến cố ngẫu nhiên được gọi là phức hợp nếu nó có thể biểu diễn được dưới dạng hợp của hai biến cố không đồng nhất với nó. Một biến cố không là phức hợp được gọi là biến cố sơ cấp (Nói cách khác: Các kết quả có thể có khi một phép thử được thực hiện gọi là các biến cố sơ cấp-hoặc các biến cố cơ bản). Vậy một biến cố phức hợp có thể xuất hiện theo nhiều cách khác nhau. Biến cố sơ cấp chỉ xuất hiện theo một cách duy nhất. Các biến cố sơ cấp từng đôi xung khắc. Tập hợp mọi biến cố sơ cấp của một phép thử được gọi là *không gian các biến cố sơ cấp*.

## 1.3. Các định nghĩa về xác suất

Mọi biến cố ngẫu nhiên đều giống nhau ở chỗ chúng không chắc chắn, nhưng khả năng xảy ra của mỗi biến cố lại có thể khác nhau. Với mỗi một biến cố ngẫu

nhiên, người ta dùng một con số để đặc trưng cho khả năng xảy ra của biến cố đó nhiều hay ít, số đó được gọi là *xác suất* của biến cố A.

**Ký hiệu:**  $P(A)$  (P là viết tắt từ chữ *Probability*)

### 1.3.1. Định nghĩa cổ điển về xác suất

**a. Định nghĩa 1.3.1.** Giả sử trong một phép thử có  $n$  kết quả đồng khả năng có thể xảy ra. Khi đó *xác suất* xuất hiện biến cố A trong một phép thử là tỷ số giữa số kết cục thuận lợi cho A và tổng số các kết cục duy nhất đồng khả năng có thể xảy ra khi thực hiện phép thử đó.

Nghĩa là:  $p(A) = \frac{m}{n}$ , trong đó:  $m$  là số kết cục thuận lợi cho A;  $n$  là tổng số các kết cục duy nhất đồng khả năng có thể xảy ra.

#### b. Tính chất:

(a)  $0 \leq P(A) \leq 1$ .

(b)  $P(U) = 1$ .

(c)  $P(V) = 0$ .

**Ví dụ 1.3.2.** Trong một lô xỏ số có 100 vé trong đó có 8 vé có thưởng. Mua ngẫu nhiên 5 vé. Tính xác suất để trong 5 vé đã mua có 2 vé trúng thưởng.

*Giải:* Gọi A là biến cố “trong 5 vé đã mua có 2 vé có thưởng”

$$\text{Số cách mua 5 vé là: } n = C_{100}^5$$

$$\text{Số kết cục thuận lợi cho A là } m = C_8^2 \cdot C_{92}^3.$$

$$\text{Vậy } P(A) = \frac{C_8^2 \cdot C_{92}^3}{C_{100}^5}$$

**Ví dụ 1.3.3.** Một người khi gọi điện thoại quên mất 2 số cuối cùng của số điện thoại và chỉ nhớ được rằng chúng khác nhau. Tìm xác suất để quay ngẫu nhiên một lần được đúng số cần gọi.

*Giải:* Gọi A là biến cố “quay ngẫu nhiên một lần được đúng số cần gọi”.

$$\text{Số kết cục đồng khả năng là: } n = A_{10}^2 = 90;$$

$$\text{Số kết cục thuận lợi cho A là: } m = 1.$$

$$\text{Vậy } P(A) = \frac{1}{90}$$

**Ví dụ 1.3.4.** Trên 7 tấm bìa kích thước như nhau trên mỗi tấm có ghi các chữ cái : 1 tấm ghi chữ H, 2 tấm ghi chữ O; 1 tấm ghi chữ C; 1 tấm ghi chữ A; 1 tấm ghi chữ T và 1 tấm ghi chữ N. Tính xác suất để xếp ngẫu nhiên 7 tấm bìa đó thành hàng ngang đọc được chữ "HOCTOAN".

*Giải:* Gọi A là biến cố “đọc thành chữ HOCTOAN” .

$$\text{Số kết cục đồng khả năng } n = 7!$$

$$\text{Số kết cục thuận lợi } m = 1.2!.1...1 = 2 .$$

$$\text{Vậy } P(A) = \frac{2}{7!}$$

**Ví dụ 1.3.5.** Một hộp chứa 7 cầu trắng và 3 cầu đen cùng kích thước. Rút ngẫu nhiên cùng lúc 4 cầu. Tính xác suất để trong bốn cầu rút được có:

- a/ 2 cầu đen
- b/ Ít nhất 2 cầu đen
- c/ Toàn cầu trắng

*Giải:*

Ta thấy rút ngẫu nhiên cùng lúc 4 trong 10 cầu có  $n = C_{10}^4$  cách

a/ Gọi A là sự kiện "trong 4 cầu rút ra có 2 cầu đen"

$$\text{Số kết cục thuận lợi A là } m = C_3^2 \cdot C_7^2$$

$$\text{Vậy } P(A) = \frac{C_3^2 \cdot C_7^2}{C_{10}^4} = 0,3$$

b/ Gọi B là sự kiện "trong 4 cầu rút ra có ít nhất 2 cầu đen"

$$\text{Số kết cục thuận lợi B là } m = C_3^2 \cdot C_7^2 + C_7^1 \cdot C_3^3$$

$$\text{Vậy } P(B) = \frac{C_7^2 \cdot C_3^2 + C_7^1 \cdot C_3^3}{C_{10}^4} = \frac{1}{3}$$

c/ Gọi C là sự kiện "4 cầu rút ra toàn cầu trắng"

$$\text{Số kết cục thuận lợi C là } m = C_7^4$$

$$\text{Vậy } P(C) = \frac{C_7^4}{C_{10}^4} = \frac{1}{6}$$

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

+ Ưu điểm là tìm xác suất của biến cố ta không phải tiến hành phép thử. (Phép thử chỉ tiến hành một cách giả định).

+ Hạn chế: Nó đòi hỏi là số kết cục duy nhất đồng khả năng có thể xảy ra trong phép thử phải là hữu hạn. Hạn chế lớn nhất của định nghĩa cổ điển là trong thực tế nhiều khi không thể biểu diễn kết quả của phép thử dưới dạng tập hợp các kết cục duy nhất và đồng khả năng.

## 1.3.2. Định nghĩa thống kê về xác suất

### a. Định nghĩa tần suất

**Định nghĩa 1.3.6.** Tần suất xuất hiện biến cố A trong n phép thử là tỷ số giữa số phép thử trong đó biến cố xuất hiện và tổng số phép thử được thực hiện.

$$\text{Kí hiệu: } f(A) = \frac{m}{n}$$

**Ví dụ 1.3.7.** Để nghiên cứu khả năng xuất hiện mặt sấp khi tung một đồng xu, người ta tiến hành tung đồng xu nhiều lần và thu được kết quả sau:

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

### b. Định nghĩa xác suất theo thống kê:

**Định nghĩa 1.3.8.** Xác suất xuất hiện biến cố A trong một phép thử là một số p không đổi mà tần suất  $f$  xuất hiện biến cố đó trong  $n$  phép thử sẽ hội tụ theo xác suất về p khi số phép thử tăng lên vô hạn.

### c. Ưu điểm và hạn chế của định nghĩa xác suất theo thống kê

+ Ưu điểm: Nó không đòi hỏi những điều kiện áp dụng như với định nghĩa cổ điển. Nó hoàn toàn dựa trên các quan sát thực tế để làm cơ sở kết luận về xác suất xảy ra của một biến cố.

+ Hạn chế: Chỉ áp dụng được đối với các hiện tượng ngẫu nhiên mà tần suất của nó có tính ổn định. Để xác định một cách tương đối chính xác giá trị của xác suất ta phải tiến hành trên thực tế một số đủ lớn các phép thử.

### 1.3.3. Nguyên lý xác suất lớn và xác suất nhỏ

Trong nhiều bài toán thực tế ta thường gặp các biến cố có xác suất rất nhỏ, tức là gần bằng không. Trong trường hợp đó ta không thể cho rằng những biến cố này sẽ không xảy ra khi thực hiện một phép thử. Thậm chí một biến cố có xác suất bằng không vẫn chưa chắc chắn đã là biến cố không thể có, tức là vẫn có thể xảy ra.

Qua quan sát người ta thấy rằng các biến cố có xác suất nhỏ gần như sẽ không xảy ra khi tiến hành một phép thử. Trên cơ sở đó có thể đưa ra "Nguyên lý thực tế không thể có của các biến cố có xác suất nhỏ" sau đây: *Nếu một biến cố có xác suất rất nhỏ thì thực tế có thể cho rằng trong một phép thử biến cố đó sẽ không xảy ra.* (Lưu ý: việc qui định mức xác suất được coi là rất nhỏ sẽ tùy thuộc vào từng bài toán cụ thể).

Một xác suất khá nhỏ mà với nó có thể cho rằng biến cố thực tế sẽ không xảy ra được gọi là mức ý nghĩa.

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

## 1.4. Các định lý cơ bản về xác suất

### 1.4.1. Định lý cộng xác suất

**Định lý 1.4.1.** Nếu A và B là 2 biến cố bất kì thì :

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

**Định lý 1.4.2.** Nếu A, B và C là 3 biến cố bất kì thì:



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

**Hệ quả 1.4.3.**

(a) Xác suất của tổng hai biến cố xung khắc bằng tổng xác suất của các biến cố đó:  $P(A + B) = P(A) + P(B)$

(b) Xác suất của tổng các biến cố xung khắc từng đôi một bằng tổng xác suất của các biến cố đó:  $P\left(\sum_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$

(c) Nếu các biến cố  $A_1, A_2, \dots, A_n$ . tạo nên một hệ đầy đủ các biến cố thì tổng xác suất của chúng bằng 1, tức là:  $\sum_{i=1}^n P(A_i) = 1$ .

(d) Tổng xác suất của hai biến cố đối lập nhau bằng 1, tức là:

$$P(A) + P(\bar{A}) = 1$$

**Ví dụ 1.4.4.** Trong một thùng đựng 30 quả cầu gồm 10 đỏ, 5 xanh và 15 trắng. Rút hủ họa một quả. Hãy tính xác suất xuất hiện quả đỏ hoặc xanh.

*Giải:* Gọi A là sự kiện "xuất hiện quả màu đỏ", B là sự kiện "xuất hiện quả màu xanh". Rõ ràng các sự kiện A, B xung khắc nên:

$$P(A + B) = P(A) + P(B) = \frac{10}{30} + \frac{5}{30} = 0,5$$

**Ví dụ 1.4.5.** Trong một cỗ bài có 52 quân (trong đó có 4 quân K). Lấy ngẫu nhiên ra 4 quân. Tính xác suất để trong 4 quân lấy ra có:

a/ Đúng 2 quân là K.

b/ Ít nhất 2 quân là K.

c/ Không quá 2 quân là K.

*Giải:* Gọi  $A_i$  là biến cố "lấy được đúng i quân là K", (i=0,1,2,3,4).

a/ Gọi A là sự kiện "lấy được đúng 2 quân là K", ta có

$$P(A) = P(A_2) = \frac{C_4^2 \cdot C_{48}^2}{C_{52}^4} \approx 0,025$$

b/ Gọi B là sự kiện "trong 4 quân lấy ra có ít nhất 2 quân là K"

$$B = A_2 + A_3 + A_4$$

$$\begin{aligned} P(B) &= P(A_2) + P(A_3) + P(A_4) \\ &= \frac{C_4^2 \cdot C_{48}^2}{C_{52}^4} + \frac{C_4^3 \cdot C_{48}^1}{C_{52}^4} + \frac{C_4^4}{C_{52}^4} \\ &= 0,025 + 0,0007 + 0,0000036 \\ &= 0,026 \end{aligned}$$

c/ Gọi C là sự kiện "trong 4 quân lấy ra có không quá 2 quân là K"

$$C = A_0 + A_1 + A_2 \quad \Rightarrow \quad \bar{C} = A_3 + A_4$$

$$P(\bar{C}) = 0,0007. \text{ Vậy } P(C) = 1 - P(\bar{C}) \approx 0,9993$$

### 1.4.2. Định lý nhân xác suất

#### a. Định nghĩa xác suất có điều kiện

**Định nghĩa 1.4.6.** Xác suất của biến cố A được tính với điều kiện biến cố B đã xảy ra gọi là *xác suất có điều kiện* của A.

**Ký hiệu:**  $P\left(\frac{A}{B}\right)$

**Ví dụ 1.4.7.** 5 người rút 5 thăm trong đó có 2 vé đi xem đá bóng. Trước lúc bắt thăm, xác suất rút được vé của anh A (cũng như của anh B) là  $P(A) = \frac{2}{5}$ . Nếu cho

biết thêm điều kiện trước đó B đã rút được một vé thì xác suất để A rút được vé là 1/4. Rõ ràng sự xuất hiện của B (anh B rút được vé) đã thay đổi khả năng rút được vé của A.

$$\text{Vậy } P(A) = \frac{2}{5}; \quad P(A/B) = \frac{1}{4}; \quad P(A/\bar{B}) = \frac{2}{4}$$

#### b. Định lý nhân xác suất

**Định lý 1.4.8.** Xác suất của tích 2 biến cố A và B bằng tích xác suất của một trong 2 biến cố đó với xác suất có điều kiện của biến cố còn lại.

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

#### Hệ quả 1.4.9.

$$(a) \text{ Nếu } P(B) > 0 \text{ thì } P\left(\frac{A}{B}\right) = \frac{P(AB)}{P(B)}$$

$$(b) P(A_1 A_2 \dots A_n) = P(A_1).P\left(\frac{A_2}{A_1}\right) \dots P\left(\frac{A_n}{A_1 A_2 \dots A_{n-1}}\right)$$

**Ví dụ 1.4.10.** Trong một bình kín có 10 quả cầu kích thước như nhau trong đó có 6 quả trắng và 4 quả đen lấy ngẫu nhiên liên tiếp không hoàn lại 2 lần mỗi lần 1 quả. Tìm xác suất để:

- a/ Cả 2 quả đều là trắng
- b/ 2 quả cùng một màu.
- c/ Có ít nhất 1 quả màu trắng.

**Giải:**

Gọi  $A_i$  là biến cố "quả lấy lần thứ i có màu trắng", ( $i=1,2$ ).

a/ Gọi A là biến cố "cả 2 quả đều là màu trắng".

$$A = A_1 A_2 \Rightarrow P(A) = P(A_1 A_2) = P(A_1).P(A_2 / A_1) = \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{3}$$

b/ Gọi B là sự kiện "2 quả lấy ra cùng màu", ta có:  $B = A_1 A_2 + \overline{A_1 A_2}$

$$\begin{aligned} P(B) &= P(A_1).P\left(\frac{A_2}{A_1}\right) + P(\overline{A_1}).P\left(\frac{\overline{A_2}}{\overline{A_1}}\right) \\ &= \frac{6}{10} \cdot \frac{5}{9} + \frac{4}{10} \cdot \frac{3}{9} = \frac{7}{15} \end{aligned}$$

c/ Gọi C là sự kiện "trong 2 quả lấy ra có ít nhất 1 quả màu trắng", thì  $\overline{C} = \overline{A_1 A_2}$

$$P(\bar{C}) = P(\bar{A}_1) \cdot P\left(\frac{\bar{A}_2}{\bar{A}_1}\right) = \frac{4}{10} \cdot \frac{3}{9} = \frac{2}{15}$$

$$\text{Vậy: } P(C) = 1 - P(\bar{C}) = \frac{13}{15}$$

**Ví dụ 1.4.11.** Trong một lớp có  $n$  học sinh, trong đó có  $m$  nữ ( $m \leq n$ ) gọi liên tiếp (không gọi lại) 3 lần mỗi lần 1 học sinh. Tìm xác suất để cả 3 học sinh được gọi đều là nữ.

*Giải:* Gọi  $A_i$  là sự kiện "học sinh được gọi lần thứ  $i$  là nữ", ( $i=1,2,3$ )

$A$  là sự kiện "cả 3 học sinh được gọi là nữ".

$$A = A_1 A_2 A_3$$

$$\begin{aligned} P(A) &= P(A_1) \cdot P\left(\frac{A_2}{A_1}\right) \cdot P\left(\frac{A_3}{A_1 A_2}\right) \\ &= \frac{m}{n} \cdot \frac{m-1}{n-1} \cdot \frac{m-2}{n-2} \end{aligned}$$

**Ví dụ 1.4.12.** Một người có 3 con gà mái, 2 con gà trống nhốt chung trong một lồng. Một người đến mua, người bán gà bắt ngẫu nhiên ra một con. Người mua chấp nhận mua con gà đó.

a/ Tìm xác suất để người đó mua được con gà mái.

b/ Người thứ hai đến mua, người bán gà lại bắt ngẫu nhiên ra một con. Tìm xác suất người thứ hai mua được gà trống.

c/ Xác suất này sẽ bằng bao nhiêu nếu người bán gà quên mất rằng con gà bán cho người thứ nhất là gà trống hay mái?

*Giải:* Gọi  $B_i$  là sự kiện "người thứ  $i$  mua được gà mái", ( $i=1, 2$ ).

a/  $P(B_1) = 3/5 = 0,6$ .

b/ Vì người thứ hai mua sau khi người thứ nhất đã mua xong cho nên:

$$P\left(\frac{\bar{B}_2}{B_1}\right) = \frac{2}{4} = 0,5.$$

c/ Ta có  $\bar{B}_2 = (B_1 \cup \bar{B}_1)\bar{B}_2 = B_1\bar{B}_2 \cup \bar{B}_1\bar{B}_2$ .

$$\text{Vậy } P(\bar{B}_2) = P(B_1)P\left(\frac{\bar{B}_2}{B_1}\right) + P(\bar{B}_1)P\left(\frac{\bar{B}_2}{\bar{B}_1}\right) = \frac{3}{5} \cdot \frac{2}{4} + \frac{2}{5} \cdot \frac{1}{4} = 0,4.$$

### c. Sự kiện độc lập

**Định nghĩa 1.4.13.** Hai biến cố  $A$  và  $B$  gọi là *độc lập* với nhau nếu việc xảy ra hay không xảy ra của biến cố này không làm thay đổi xác suất xảy ra của biến cố kia. Trong trường hợp ngược lại thì 2 biến cố đó được gọi là phụ thuộc. Nếu  $A$  và  $B$  là độc lập với nhau thì:

$$+ P\left(\frac{A}{B}\right) = P(A);$$

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

+ Các cặp  $(\bar{A}, B), (A, \bar{B}), (\bar{A}, \bar{B})$  cũng độc lập.

**Tổng quát:**

+ Các biến cố  $A_1, A_2, \dots, A_n$  gọi là độc lập từng đôi với nhau nếu mỗi cặp 2 trong  $n$  biến cố đó độc lập với nhau.

+ Các biến cố gọi là độc lập toàn phần với nhau nếu mỗi biến cố độc lập với tổ hợp bất kỳ các biến cố còn lại.

**Ví dụ 1.4.14.** Một phòng điều trị có 3 bệnh nhân bệnh nặng với xác suất cần cấp cứu trong cùng một giờ của các bệnh nhân tương ứng là 0,7 ; 0,8 và 0,9. Tìm xác suất sao cho trong cùng một giờ :

a/ Có 2 bệnh nhân cần cấp cứu.

b/ Có ít nhất 1 bệnh nhân không cần cấp cứu.

*Giải:* Gọi  $A_i$  là sự kiện "bệnh nhân thứ  $i$  cần cấp cứu", ( $i=1,2,3$ ). Theo giả thiết ta có:  $P(A_1) = 0,7$ ;  $P(A_2) = 0,8$ ;  $P(A_3) = 0,9$

a/ Gọi  $A$  là sự kiện "có 2 bệnh nhân cần cấp cứu"

$$A = A_1 A_2 \bar{A}_3 + A_1 \bar{A}_2 A_3 + \bar{A}_1 A_2 A_3 \quad \text{Vì } A_i \text{ là độc lập nên}$$

$$\begin{aligned} P(A) &= P(A_1).P(A_2).P(\bar{A}_3) + P(A_1).P(\bar{A}_2).P(A_3) + P(\bar{A}_1).P(A_2).P(A_3) \\ &= 0,7.0,8.0,1 + 0,7.0,2.0,9 + 0,3.0,8.0,9 \\ &= 0,398 \end{aligned}$$

b/ Gọi  $B$  là sự kiện "có ít nhất 1 bệnh nhân không cần cấp cứu", vậy  $\bar{B}$  là sự kiện "không có bệnh nhân không cần cấp cứu".

$$\begin{aligned} \bar{B} &= A_1.A_2.A_3 \quad \Rightarrow \quad P(\bar{B}) = 0,7.0,8.0,9 \\ P(B) &= 1 - P(\bar{B}) = 0,496 \end{aligned}$$

**Ví dụ 1.4.15.** Hai máy bay cùng ném bom một mục tiêu, mỗi máy bay ném 1 quả bom với xác suất trúng mục tiêu tương ứng là 0,7 và 0,8. Tìm xác suất để mục tiêu trúng bom.

*Giải:* Gọi  $A_i$  là sự kiện "máy bay thứ  $i$  bắn trúng mục tiêu", ( $i=1,2$ ).

$A$  là sự kiện "mục tiêu trúng bom"

$A = A_1 + A_2$ . Vì  $A_1; A_2$  là không xung khắc và độc lập nên

$$\begin{aligned} P(A) &= P(A_1) + P(A_2) - P(A_1 A_2) \\ &= 0,7 + 0,8 - 0,7.0,8 = 0,94 \end{aligned}$$

*Cách khác:* Vì  $A_1; A_2$  là không xung khắc và độc lập nên:

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

### 1.4.3. Công thức xác suất toàn phần - Công thức Bayes

**Bài toán 1.4.16.** Giả sử trong một phép thử có một hệ đầy đủ các biến cố  $A_1, A_2, \dots, A_n$ ,  $B$  là một biến cố nào đó.  $B$  chỉ xảy ra khi và chỉ khi một trong các biến cố  $A_i (i = \overline{1, n})$  của hệ trên xảy ra. Biết các xác suất  $P(A_i)$  và  $P\left(\frac{B}{A_i}\right)$ . Hãy tìm  $P(B)$ .

*Giải:* Theo giả thiết B xảy ra khi và chỉ khi một trong các biến cố  $A_i$  của hệ trên xảy ra ta có:  $B = A_1B \cup A_2B \cup \dots \cup A_nB$ . Vì  $A_i$  lập thành hệ đầy đủ các biến cố nên các biến cố  $A_i$  xung khắc từng đôi, do đó  $A_iB$  cũng xung khắc từng đôi và:

$$P(B) = P(A_1).P\left(\frac{B}{A_1}\right) + P(A_2).P\left(\frac{B}{A_2}\right) + \dots + P(A_n).P\left(\frac{B}{A_n}\right).$$

Hay: 
$$P(B) = \sum_{i=1}^n P(A_i).P\left(\frac{B}{A_i}\right).$$

Công thức này được gọi là *công thức xác suất toàn phần* (hay *công thức xác suất đầy đủ*).

Cùng với giả thiết của bài toán trên nhưng thêm điều kiện là phép thử đã được thực hiện và sự kiện B đã xảy ra và do vậy phải có duy nhất một sự kiện  $A_k$  nào đó xảy ra từ đó ta có:

$$P\left(\frac{A_k}{B}\right) = \frac{P(A_k).P\left(\frac{B}{A_k}\right)}{\sum_{i=1}^n P(A_i).P\left(\frac{B}{A_i}\right)}. \quad \text{Đây là công thức Bayes.}$$

**Chú ý 1.4.17.** Công thức xác suất đầy đủ thường được gọi là công thức xác suất tiên nghiệm. Công thức Bayes được xác định sau khi phép thử đã tiến hành và biến cố B đã xảy ra do đó được gọi là công thức xác suất hậu nghiệm.

**Ví dụ 1.4.17.** Một trại lợn nhận lợn giống từ 3 cơ sở theo tỷ lệ 25%; 35% và 40%. Biết tỷ lệ lợn giống không đủ tiêu chuẩn ở mỗi cơ sở lần lượt là: 5%; 4% và 2%. Bất ngẫu nhiên một con lợn của trại.

a/ Tìm xác suất để ta bắt được con lợn đủ tiêu chuẩn.

b/ Giả sử ta bắt được con lợn không đủ tiêu chuẩn. Theo anh (chị) con lợn đó có khả năng thuộc cơ sở nào nhất.

*Giải:* Gọi  $A_i$  là sự kiện "bắt được lợn giống của cơ sở thứ i", ( $i = 1..3$ )

A là sự kiện "bắt được lợn đủ tiêu chuẩn"

$$A = AA_1 + AA_2 + AA_3$$

a/  $P(A) = P(A_1)P(A/A_1) + P(A_2)P(A/A_2) + P(A_3)P(A/A_3) = 0,9655$

b/  $P(\bar{A}) = 1 - P(A) = 0,0345$

Theo Bayes:  $P(A_1/\bar{A}) = (0,25.0,05)/0,0345 = 0,362$

$$P(A_2/\bar{A}) = (0,35.0,04)/0,0345 = 0,406$$

$$P(A_3/\bar{A}) = (0,4.0,02)/0,0345 = 0,229$$

Vậy con lợn không đủ tiêu chuẩn đó khả năng thuộc cơ sở 2 cung cấp

**Ví dụ 1.4.18.** Một thiết bị gồm ba loại linh kiện: Loại I chiếm 35%, loại II chiếm 25%, loại III chiếm 40% tổng số linh kiện của toàn thiết bị. Xác suất hư hỏng sau khoảng thời gian làm việc nào đó của các loại linh kiện tương ứng là: 15%; 25% và

5%. Máy đang hoạt động bỗng bị hỏng, Hãy tính xác suất để từng loại linh kiện bị hỏng (giả thiết các loại linh kiện không cùng hỏng đồng thời).

*Giải:* Gọi  $A_i$  là sự kiện "kiểm tra linh kiện loại  $i$ ", ( $i = 1..3$ )

$A$  là sự kiện "máy bị hỏng".

$$A = AA_1 + AA_2 + AA_3$$

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

$$= 0,35 \cdot 0,15 + 0,25 \cdot 0,25 + 0,4 \cdot 0,05 = 0,135$$

Theo Bayes:  $P(A_1/A) = (0,35 \cdot 0,15) / 0,135 = 0,389$

$$P(A_2/A) = (0,25 \cdot 0,25) / 0,135 = 0,463$$

$$P(A_3/A) = (0,4 \cdot 0,05) / 0,135 = 0,148$$

#### 1.4.4. Công thức Bernoulli

**a. Dãy phép thử Bernoulli :** Xét một dãy  $n$  phép thử độc lập, mỗi phép thử chỉ có 2 biến cố  $A$  hoặc  $\bar{A}$  với  $P(A) = p; P(\bar{A}) = 1 - p = q$  không đổi, không phụ thuộc vào thứ tự phép thử. Dãy phép thử đó gọi là *dãy phép thử Bernoulli*.

**b. Công thức Bernoulli:** Với dãy phép thử Bernoulli ta có bài toán.

**Bài toán 1.4.19.** Tìm xác suất sao cho trong  $n$  phép thử đó sự kiện  $A$  xuất hiện đúng  $k$  lần không phân biệt thứ tự.

**Bài toán 1.4.20.** Tìm xác suất sao cho trong  $n$  phép thử đó sự kiện  $A$  xuất hiện từ  $k_1$  đến  $k_2$  lần không phân biệt thứ tự.

*Giải:* Gọi  $B$  là biến cố trong  $n$  phép thử Bernoulli đó biến cố  $A$  xuất hiện đúng  $k$  lần không phân biệt thứ tự. Ta thấy biến cố  $B$  có thể xuất hiện theo nhiều cách khác nhau. Giả sử ta xét một trường hợp của  $B$  là biến cố  $A$  xuất hiện ở  $k$  phép thử đầu còn  $(n-k)$  phép thử tiếp theo xuất hiện  $\bar{A}$ . Vậy

$$\begin{aligned} P(A \dots A \bar{A} \dots \bar{A}) &= P(A \dots A) \cdot P(\bar{A} \dots \bar{A}) \\ &= p^k \cdot q^{n-k} \end{aligned}$$

Ta có  $C_n^k$  cách để  $A$  xuất hiện với cùng một xác suất như trên.

Vậy:  $P_n(k) = P(B) = C_n^k \cdot p^k \cdot q^{n-k}$ . Đây là công thức Bernoulli.

Hoàn toàn tương tự như trên Gọi  $H$  là biến cố trong  $n$  phép thử Bernoulli biến cố  $A$  xuất hiện từ  $k_1$  đến  $k_2$  lần không phân biệt thứ tự.

$$H = \bigcup_{k=k_1}^{k_2} B_k.$$

Vì các  $B_k$  xung khắc từng đôi một do đó:

$$P_n(k_1, k_2) = P(H) = \sum_{k=k_1}^{k_2} C_n^k \cdot p^k \cdot q^{n-k}.$$

**Ví dụ 1.4.21.** Xác suất tiêu thụ điện năng trong mỗi ngày không vượt quá quy định tại một xí nghiệp là  $p = 0,75$ . Tính xác suất sao cho trong 6 ngày liên tiếp có 4 ngày lượng điện năng tiêu thụ không vượt quá mức quy định.

*Giải:* Bài toán thỏa mãn dãy phép thử Bernoulli

Ta có  $n = 6, k = 4, p = 0,75, q = 0,25$

$$\text{Vậy: } P_6(4) = C_6^4 (0,75)^4 \cdot (0,25)^2 = 0,3$$

**Ví dụ 1.4.22.** Bắn liên tiếp 5 viên đạn vào một mục tiêu xác suất trúng đích của mỗi viên là 0,2 .

a/ Tính xác suất để có đúng 2 viên trúng;

b/ Để mục tiêu bị tiêu diệt cần phải có từ 3 viên trúng mục tiêu trở lên. Tính xác suất để mục tiêu bị tiêu diệt.

*Giải:* Bài toán thỏa mãn dãy phép thử Bernoulli, do đó

$$\text{a/ } P_5(2) = C_5^2 p^2 q^3 = C_5^2 \cdot (0,2)^2 \cdot (0,8)^3 = 0,2048$$

$$\text{b/ } p_5(3,5) = C_5^3 p^3 q^2 + C_5^4 p^4 q + C_5^5 p^5 = 0,0575$$

**Ví dụ 1.4.23.** Một bác sĩ chữa bệnh có xác suất chữa khỏi là 0,8 có người nói rằng cứ 5 người đến chữa thì có chắc chắn 4 người khỏi bệnh, người khác lại cho rằng trong 10 người đến chữa có chắc chắn 8 người khỏi bệnh. Điều đó có đúng không.

*Giải:* Cả 2 người khẳng định đều sai, vì bài toán trên thỏa mãn dãy phép thử Bernoulli nên xác suất xảy ra trong các trường hợp là:

$$P_5(4) = C_5^4 \cdot (0,8)^4 \cdot 0,2 = 0,4096$$

$$P_{10}(8) = C_{10}^8 \cdot (0,8)^8 \cdot (0,2)^2 = 0,3018$$

### c. Giá trị khả năng nhất

Trị số của  $P_n(k)$  nói chung phụ thuộc vào  $k$  (với  $n$  cố định), ta cần tìm trị số  $k = k_0$  sao cho  $P_n(k)$  đạt giá trị lớn nhất. Số  $k_0$  được gọi là số lần xuất hiện chắc chắn nhất (có khả năng nhất) của biến cố A trong  $n$  phép thử đã cho. Qua khảo sát  $P_n(k)$  ta có kết quả sau:

+ Nếu  $(np + p - 1)$  là số nguyên thì  $P_n(k)$  đạt cực đại tại 2 giá trị  $k_0 = np + p - 1$  và  $k_0 = np + p$

+ Nếu  $(np + p - 1)$  là số không nguyên thì  $P_n(k)$  đạt giá trị lớn nhất tại  $k_0 = [np + p - 1] + 1$

**Ví dụ 1.4.24.** Tỷ lệ mắc một loại bệnh A ở một vùng là 10%. Trong đợt khám bệnh cho vùng đó người ta đã khám 100 người. Tìm xác suất để trong 100 người có:

a/ 6 người bị bệnh A;

b/ 95 người không bị bệnh A;

c/ Ít nhất 1 người bị bệnh A;

d/ Tìm số người bị bệnh A có khả năng nhất? Tính xác suất tương ứng.

*Giải:* Bài toán thỏa mãn dãy phép thử Bernoulli với  $P(A) = 0,10$  do đó:

$$\text{a/ } P_{100}(6) = C_{100}^6 \cdot 0,1^6 \cdot 0,9^{94};$$

$$\text{b/ } P_{100}(95) = C_{100}^{95} \cdot 0,9^{95} \cdot 0,1^5;$$

$$c/ P_{100}(k \geq 1) = 1 - P_{100}(k = 0) = 1 - C_{100}^0 \cdot 0,1^0 \cdot 0,9^{100} = 1 - 0,9^{100}$$

d/ Theo bài ra ta có  $np + p - 1 = 100 \cdot 0,1 + 0,1 - 1 = 9,1$ . Vậy số người bị bệnh A có khả năng nhất khi khám 100 người là 10 người và

$$P_{100}(10) = C_{100}^{10} \cdot 0,1^{10} \cdot 0,9^{90}$$

## BÀI TẬP CHƯƠNG 1

### Dạng 1: Công thức xác suất cổ điển

- 1.** Thang máy của một tòa nhà 7 tầng xuất phát từ tầng một với 3 khách. Tìm xác suất để:
  - a/ Tất cả cùng ra ở tầng bốn.
  - b/ Tất cả cùng ra ở một tầng.
  - c/ Mỗi người ra ở một tầng khác nhau.
- 2.** Xếp ngẫu nhiên 4 khách lên 9 toa tàu hỏa. Tìm xác suất để:
  - a/ 4 người lên toa đầu.
  - b/ 4 người lên cùng một toa.
  - c/ 4 người lên 4 toa khác nhau.
- 3.** Có 2 lô hàng, lô 1 có 90 chính phẩm và 10 phế phẩm, lô 2 có 80 chính phẩm và 20 phế phẩm. Lấy ngẫu nhiên từ mỗi lô hàng ra 1 sản phẩm. Tính xác suất để:
  - a/ Lấy được 1 chính phẩm;
  - b/ Lấy được ít nhất 1 chính phẩm.
  - c/ Lấy được 2 chính phẩm.
- 4.** Có hai chuồng lợn giống, chuồng 1 có 7 con cái và 3 con đực, chuồng 2 có 6 con cái và 4 con đực. Bắt ngẫu nhiên từ mỗi chuồng ra một con. Tính xác suất để:
  - a/ Cả 2 con bắt ra đều là con cái.
  - b/ Bắt được một con cái và một con đực.
  - c/ Bắt được ít nhất một con đực.
- 5.** Một kĩ sư nông nghiệp có hai hộp hạt giống cùng loại: Hộp 1 có 12 hạt giống trong đó 8 hạt đủ tiêu chuẩn, hộp 2 có 12 hạt giống trong đó có 9 hạt đủ tiêu chuẩn. Chọn ngẫu nhiên từ mỗi hộp ra 1 hạt giống. Tìm xác suất để trong hai hạt lấy ra:
  - a/ Có một hạt đủ tiêu chuẩn, một hạt không đủ tiêu chuẩn.
  - b/ Lấy được ít nhất 1 hạt đủ tiêu chuẩn.
  - c/ Lấy được 2 hạt đủ tiêu chuẩn.
- 6.** Trong một hòm đựng 8 chi tiết là chính phẩm và 5 chi tiết là phế phẩm. Lấy đồng thời ra 3 chi tiết. Tính xác suất để:
  - a/ Cả 3 chi tiết lấy ra là chính phẩm.
  - b/ Trong 3 chi tiết lấy ra có 2 chính phẩm.
  - c/ Trong 3 chi tiết lấy ra có ít nhất 1 chính phẩm.
- 7.** Trong một lớp học có 15 học sinh nam và 10 học sinh nữ. Gọi ngẫu nhiên 4 học sinh lên bảng làm bài tập. Tính xác suất để:



- a/ có 2 học sinh nam.  
 b/ Có ít nhất 2 học sinh nam.  
 c/ Có cả nam và nữ.
- 8.** Một hộp đựng 7 quả cầu trắng và 8 quả cầu đen cùng kích cỡ. Lấy ngẫu nhiên ra 4 quả cầu. Tìm xác suất để:  
 a/ Trong 4 quả lấy ra có 3 quả trắng?  
 b/ Có 4 quả cùng màu?  
 c/ Có ít nhất 1 quả màu đen?
- 9.** Trong một hộp bút có 10 chiếc bút bi cùng kích cỡ, trong đó có 6 chiếc bút mực đen và 4 chiếc bút mực xanh. Lấy ngẫu nhiên ra 3 chiếc bút. Tìm xác suất trong 3 chiếc lấy ra có:  
 a/ 2 chiếc bút mực xanh?  
 b/ ít nhất 2 chiếc bút mực xanh:  
 c/ 2 chiếc cùng màu:
- 10.** Một chiếc hộp đựng 6 quả cầu trắng, 4 quả cầu đỏ và 2 quả cầu đen. Chọn ngẫu nhiên ra 6 quả cầu. Tìm xác suất trong 6 quả lấy ra có:  
 a/ 3 quả trắng, 2 quả đỏ và 1 quả đen?  
 b/ 4 quả đỏ?  
 c/ Không có quả nào màu trắng?

**Dạng 2: Công thức xác suất tổng, công thức xác suất đầy đủ, Bayss, Bernouly**

- 11.** Một nhà máy sản xuất bóng đèn. Máy A sản xuất 25% số bóng đèn ,máy B sản xuất 35% số bóng đèn,còn máy C sản xuất 40% số bóng đèn.Tỉ lệ sản phẩm hỏng của các máy tương ứng là 5% (máy A),4% (máy B) và 2% (máy C).  
 a/ Lấy ngẫu nhiên một bóng đèn.Tìm xác suất để gặp bóng đèn xấu.  
 b/ Khi lấy ngẫu nhiên một bóng đèn ta được bóng đèn tốt. Tìm xác suất để bóng tốt lấy được đó do máy B sản xuất.
- 12.** Một dự án trồng cây lâm nghiệp nhận giống cây trồng từ 3 cơ sở sản xuất giống cây trồng. Trung bình cơ sở 1 cung cấp 35%, cơ sở 2 cung cấp 40%, cơ sở 3 cung cấp 25% tổng số giống cây trồng của dự án. Trong đó khoảng 90% cây giống do cơ sở 1 cung cấp là đủ tiêu chuẩn, 85% cây giống do cơ sở 2 cung cấp là đủ tiêu chuẩn, 80% cây giống do cơ sở 3 cung cấp là đủ tiêu chuẩn. Lấy ngẫu nhiên một cây trồng của dự án để kiểm tra.  
 a/ Tính xác suất để cây trồng lấy ra đủ tiêu chuẩn.  
 b/ Giả sử cây lấy ra đủ tiêu chuẩn, theo anh (chị) cây đó có khả năng do cơ sở nào cung cấp.
- 13.** Một trại lợn nhận lợn giống từ 3 cơ sở theo tỷ lệ 20% ; 35% và 45% . Biết tỷ lệ lợn giống không đủ tiêu chuẩn ở mỗi cơ sở lần lượt là 2% ; 3% và 4% . Bất ngẫu nhiên một con lợn của trại.  
 a/ Tìm xác suất để bắt được con lợn đủ tiêu chuẩn.  
 b/ Giả sử bắt được con lợn không đủ tiêu chuẩn. Theo bạn con lợn đó có khả năng thuộc cơ sở nào nhất?

**14.** Trong một bệnh viện, tỷ lệ bệnh nhân các tỉnh như sau: Tỉnh A: 25% , tỉnh B: 35% và tỉnh C: 40% . Biết tỷ lệ bệnh nhân là kỹ sư của các tỉnh tương ứng là 2,5% ; 3% và 4,5% . Chọn ngẫu nhiên một bệnh nhân.

a/ Tính xác suất để bệnh nhân đó là kỹ sư.

b/ Giả sử bệnh nhân được chọn không phải là kỹ sư. Theo bạn bệnh nhân đó có khả năng thuộc tỉnh nào nhất?

**15.** Có 3 cửa hàng I, II và III cùng kinh doanh sản phẩm Y. Tỷ lệ sản phẩm loại A trong 3 cửa hàng I, II, III lần lượt là 70%, 75% và 50%. Một khách hàng chọn ngẫu nhiên một cửa hàng và từ đó mua một sản phẩm.

a/ Tính xác suất để khách hàng đó mua được sản phẩm loại A.

b/ Giả sử khách hàng đã mua được sản phẩm loại A, theo bạn sản phẩm đó có khả năng thuộc cửa hàng nào?

**16.** Một cửa hàng bán máy tính với 40% máy tính của hãng IBM, 60% máy tính của hãng Acer. Biết rằng tỷ lệ máy sản xuất tại chính hãng IBM và Acer lần lượt là 0,8; 0,9. Một khách hàng mua máy tính tại cửa hàng.

a/ Tính xác suất để khách hàng mua được máy tính sản xuất tại chính hãng.

b/ Giả sử khách hàng mua được máy tính sản xuất tại chính hãng, theo bạn máy tính đó có khả năng do hãng nào sản xuất?

**17.** Có 20 kiện hàng mỗi kiện hàng có 10 sản phẩm. Trong số đó có 8 kiện loại 1, mỗi kiện hàng có 1 phế phẩm; 7 kiện hàng loại 2, mỗi kiện hàng có 2 phế phẩm và 5 kiện hàng loại 3, mỗi kiện có 3 phế phẩm. Lấy ngẫu nhiên một kiện hàng, rồi từ đó lấy ngẫu nhiên một sản phẩm.

a/ Tính xác suất để sản phẩm lấy ra là phế phẩm.

b/ Nếu lấy được sản phẩm là phế phẩm, theo bạn sản phẩm đó có khả năng thuộc kiện hàng loại nào nhiều hơn cả?

**18.** Trong một lớp học, tỷ lệ học sinh thích chơi game là 70%. Biết rằng nếu ham chơi game thì tỷ lệ học sinh đạt học lực khá là 30%, còn nếu không chơi game thì tỷ lệ học sinh đạt học lực khá là 60%. Gọi một học sinh lên bảng.

a/ Tính xác suất để học sinh đó có học lực khá.

b/ Giả sử học sinh đó có học lực khá. Tính xác suất để học sinh đó chơi game.

**19.** Ở một vùng dân cư cứ 100 người có 20 người hút thuốc lá. Biết rằng tỷ lệ người viêm họng trong số người hút thuốc lá là 65%, còn trong số người không hút thuốc là 35%. Khám ngẫu nhiên một người thì thấy anh ta viêm họng, tìm xác suất để người đó hút thuốc. Nếu người đó không viêm họng thì xác suất để người đó không hút thuốc là bao nhiêu.

**20.** Có 2 hộp như nhau đựng các mẫu hàng xuất khẩu. Hộp thứ nhất có 10 mẫu trong đó có 6 mẫu loại A và 4 mẫu loại B. Hộp thứ 2 có 10 mẫu trong đó có 3 mẫu loại A và 7 mẫu loại B. Chọn ngẫu nhiên 1 hộp và từ đó lấy ngẫu nhiên 1 mẫu.

a/ Tính xác suất để mẫu lấy ra là loại B.

b/ Giả sử mẫu lấy ra loại A. Hỏi mẫu đó có khả năng thuộc hộp loại nào nhiều hơn?

**21.** Trong 1 bệnh viện bỏng: 80% bệnh nhân bị bỏng do nóng, 20% bệnh nhân bị bỏng do hóa chất. Trong số những bệnh nhân bị bỏng do nóng thì có 30% bị biến chứng, còn với bỏng do hóa chất thì có 60% bị biến chứng. Từ tập bệnh án rút ngẫu

nhiên ra 1 hồ sơ thấy đó là của bệnh nhân bị biến chứng. Tìm xác suất để bệnh nhân đó bị bỏng do hóa chất gây ra?

**22.** Có 20 hộp sản phẩm cùng loại, trong đó có 10 hộp của xí nghiệp I, 6 hộp của xí nghiệp II, 4 hộp của xí nghiệp III. Tỷ lệ sản phẩm tốt của các xí nghiệp tương ứng lần lượt là 50%, 65% và 75%. Lấy ngẫu nhiên ra một hộp và chọn ngẫu nhiên ra một sản phẩm.

a/ Tính xác suất để sản phẩm đó là tốt.

b/ Nếu sản phẩm đó là tốt, theo bạn sản phẩm đó có khả năng thuộc xí nghiệp nào là nhiều hơn cả?

**23.** Có 18 học sinh thi học sinh giỏi chia làm 4 nhóm: nhóm I có 5 học sinh, nhóm II có 7 học sinh, nhóm III có 4 học sinh và nhóm IV có 2 học sinh. Xác suất để một học sinh trong nhóm đạt giải tương ứng lần lượt là 0,8; 0,7; 0,6; 0,5.

a/ Tính xác suất để một học sinh bất kỳ đạt giải.

b/ Nếu học sinh đó đạt giải hãy tính xác suất để học sinh đó thuộc nhóm I?

**24.** Trong một làng tỷ lệ nam là 60% và nữ là 40%. Khả năng mắc bệnh bạch tạng ở nam là 0,6% và ở nữ là 0,35%. Gặp một người trong làng thấy người đó mắc bệnh. Tìm xác suất để người đó là nam? Nếu người đó không mắc bệnh xác suất để người đó là nam là bao nhiêu?

**25.** Hai máy cùng sản xuất một loại sản phẩm. Tỷ lệ phế phẩm của máy I là 3% của máy II là 2%. Từ một kho gồm  $\frac{2}{3}$  sản phẩm của máy I và  $\frac{1}{3}$  sản phẩm của máy II ta lấy một sản phẩm. Tính xác suất để:

a/ Sản phẩm lấy ra là tốt.

b/ Giả sử sản phẩm lấy ra là sản phẩm tốt. Tính xác suất để sản phẩm lấy ra là của máy I sản xuất.

**26.** Có 10 sinh viên đi thi, trong đó có 3 sinh viên thuộc loại giỏi, 4 khá và 3 trung bình. Trong số 20 câu hỏi thi qui định thì sinh viên loại giỏi trả lời được tất cả, sinh viên khá trả lời được 16 câu, còn sinh viên trung bình chỉ trả lời được 10 câu. Gọi ngẫu nhiên 1 sinh viên và phát 1 phiếu thi có 4 câu hỏi thì anh ta trả lời được cả 4 câu hỏi. Tính xác suất để sinh viên đó thuộc loại khá.

## Chương 2

### Đại lượng ngẫu nhiên – Quy luật phân phối xác suất

#### 2.1. Đại lượng ngẫu nhiên

##### 2.1.1. Định nghĩa:

**Định nghĩa 2.1.1.** *Đại lượng ngẫu nhiên (hay biến ngẫu nhiên)* là đại lượng mà trong kết quả của phép thử sẽ nhận một và chỉ một trong các giá trị có thể của nó với một xác suất tương ứng xác định.

Người ta thường dùng các chữ cái in  $X, Y, Z, \dots$  hoặc  $X_1, X_2, \dots$  để chỉ đại lượng ngẫu nhiên và các chữ cái thường  $x, y, z, \dots$  hoặc  $x_1, x_2, \dots$  để chỉ các giá trị có thể có của nó.

Đại lượng ngẫu nhiên được phân làm 2 loại: Đại lượng ngẫu nhiên rời rạc và đại lượng ngẫu nhiên liên tục.

+ Đại lượng ngẫu nhiên rời rạc là đại lượng ngẫu nhiên mà các giá trị có thể có của nó lập nên một tập hợp hữu hạn hoặc vô hạn đếm được phân tử.

**Ví dụ 2.1.2.** Gọi  $X$  là số viên đạn bắn trúng bia khi bắn 3 viên, khi đó  $X$  là đại lượng ngẫu nhiên rời rạc nhận một trong các giá trị 0, 1, 2, 3.

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

**Ví dụ 2.1.3.** Gọi  $X$  là khoảng cách từ điểm chạm của viên đạn đến tâm bia, thì  $X$  là đại lượng ngẫu nhiên liên tục.

**Chú ý 2.1.4.** Có thể nói rằng gần như tất cả các đại lượng chỉ về trọng lượng, độ dài đều là đại lượng ngẫu nhiên liên tục.

##### 2.1.2. Các quy luật phân phối xác suất của đại lượng ngẫu nhiên

###### a. Bảng phân phối xác suất của đại lượng ngẫu nhiên

Người ta thường biểu thị qui luật phân phối xác suất của đại lượng ngẫu nhiên rời rạc dưới dạng bảng gồm 2 dòng: Dòng 1 ghi các giá trị mà đại lượng ngẫu nhiên có thể nhận được:  $x_1, x_2, \dots, x_n$ . Dòng 2 ghi các giá trị xác suất tương ứng.

X	$x_1$	$x_2$	.....	$x_i$	.....	$x_n$
p	$p_1$	$p_2$	.....	$p_i$	.....	$p_n$

$$\text{với } \sum_{i=1}^n p_i = 1.$$

**Ví dụ 2.1.5.** Một tổ sản xuất có 3 mô tơ chạy độc lập với nhau, với xác suất để mỗi mô tơ chạy tốt trong ngày là 0,7. Gọi X là số mô tơ chạy tốt trong ngày. Lập bảng phân phối xác suất của X.

*Giải:* X là đại lượng ngẫu nhiên rời rạc với các giá trị có thể có 0, 1, 2, 3. với xác suất tương ứng được tính theo công thức Bernoulli

$$P(X = i) = p_3(i) = C_3^i p^i q^{3-i}, \text{ với } p = 0,7; q = 0,3.$$

Ta có bảng phân phối xác suất của đại lượng ngẫu nhiên X là:

X	0	1	2	3
P	0,027	0,189	0,441	0,343

**Ví dụ 2.1.6.** Một xạ thủ có 3 viên đạn được yêu cầu bắn lần lượt từng viên cho đến khi trúng mục tiêu hoặc hết cả 3 viên thì thôi. Tìm bảng phân phối xác suất của số đạn đã bắn, biết rằng xác suất bắn trúng đích của mỗi lần là 0,8.

*Giải:* Gọi X là số đạn đã dùng, ta có X là đại lượng ngẫu nhiên rời rạc gồm 3 giá trị là 1, 2, 3 với

$$P(X = 1) = 0,8; P(X = 2) = 0,2 \cdot 0,8 = 0,16; P(X = 3) = 0,2 \cdot 0,2 = 0,04.$$

Vậy

X	1	2	3
P	0,8	0,16	0,04

## b. Hàm phân phối xác suất

**Định nghĩa 2.1.7.** Hàm phân phối xác suất của đại lượng X ngẫu nhiên, ký hiệu là  $F(x)$ , là xác suất để đại lượng ngẫu nhiên nhận giá trị nhỏ hơn x, với x là một số thực bất kỳ.

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

**Chú ý 2.1.8.** Nếu X là đại lượng ngẫu nhiên rời rạc thì  $F(x) = \sum_{x_i < x} p_i$ .

**Ví dụ 2.1.9.** Bắn liên tiếp 3 viên đạn vào bia, với xác suất trúng đích của mỗi viên là 0,4. Gọi X là số viên đạn trúng bia. Hãy tìm hàm phân phối xác suất của X.

*Giải:* X là đại lượng ngẫu nhiên rời rạc nhận 4 giá trị 0, 1, 2, 3, với xác suất được tính theo công thức Bernoulli

$$P_3(i) = C_3^i p^i q^{3-i}. \quad (i = 1..3)$$

Ta có bảng phân phối xác suất:

X	0	1	2	3
P	0,216	0,432	0,288	0,064

- + Nếu  $x \leq 0$ , biến cố  $(X < x)$  là biến cố không thể có, do đó  $F(x) = 0$ ;
- + Nếu  $0 < x \leq 1$ , biến cố  $(X < x)$  sẽ xảy ra khi  $X = 0$ , do đó  $F(x) = 0,216$ ;
- + Nếu  $1 < x \leq 2$ , biến cố  $(X < x)$  sẽ xảy ra khi  $X = 0$  hoặc  $X = 1$ ,  
do đó  $F(x) = 0,216 + 0,432 = 0,648$ ;

+ Nếu  $2 < x \leq 3$  biến cố ( $X < x$ ) sẽ xảy ra khi  $X = 0$  hoặc  $X = 1$ , hoặc  $X = 2$ , do đó  $F(x) = 0,936$ ;

+ Nếu  $x > 3$  biến cố ( $X < x$ ) sẽ xảy ra khi  $X = 0$ ;  $X = 1$ ;  $X = 2$ , hoặc  $X = 3$ , do đó  $F(x) = 1$ .

Vậy hàm phân phối xác suất của  $X$  là:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 0,216 & 0 < x \leq 1 \\ 0,648 & 1 < x \leq 2 \\ 0,936 & 2 < x \leq 3 \\ 1 & x > 3 \end{cases}$$

**Ví dụ 2.1.10.** Có 2 lồng nhốt gà: Lồng thứ nhất có 6 gà mái và 2 gà trống, lồng thứ 2 có 5 gà mái và 3 gà trống. Từ lồng thứ nhất bắt 2 gà bỏ sang lồng 2, rồi từ lồng 2 bắt ngẫu nhiên ra 2 con.

a/ Tìm qui luật phân phối xác suất chỉ số gà mái được bắt ra.

b/ Tìm hàm phân phối xác suất chỉ số gà mái được bắt ra.

*Giải:* Gọi  $A_i$  là sự kiện bắt được  $i$  gà mái từ lồng 1 vào lồng 2 ( $i = 0..2$ )

$X$  là số gà mái có trong 2 con được bắt từ lồng 2:  $X = 0; 1; 2$

$X = A_0X + A_1X + A_2X$

$$P(X) = P(A_0)P(X/A_0) + P(A_1)P(X/A_1) + P(A_2)P(X/A_2)$$

$$P(X=0) = \frac{C_2^2}{C_8^2} \cdot \frac{C_5^2}{C_{10}^2} + \frac{C_6^1 C_2^1}{C_8^2} \cdot \frac{C_4^2}{C_{10}^2} + \frac{C_6^2}{C_8^2} \cdot \frac{C_3^2}{C_{10}^2} = \frac{127}{1260} \approx 0,1008$$

Tương tự:  $P(X=1) = \frac{628}{1260} \approx 0,4984$        $P(X=2) = \frac{505}{1260} \approx 0,4008$

a/ Qui luật phân phối xác suất của  $X$ :

$X$	0	1	2
$P$	0,1008	0,4984	0,4008

b/ Hàm phân phối xác suất của  $X$ :  $F(X) = \begin{cases} 0 & \text{Khi } x \leq 0 \\ 0,1008 & \text{Khi } 0 < x \leq 1 \\ 0,5992 & 1 < x \leq 2 \\ 1 & x > 2 \end{cases}$

### Các tính chất của hàm phân phối xác suất

#### Tính chất 2.1.11.

(a)  $0 \leq F(x) \leq 1$ ;

(b) Hàm phân phối xác suất là hàm không giảm tức là với  $x_2 > x_1$  thì  $F(x_2) \geq F(x_1)$ .

(c)  $F(-\infty) = 0$ ;  $F(+\infty) = 1$

#### Hệ quả 2.1.12.

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

(b) Xác suất để đại lượng ngẫu nhiên liên tục  $X$  nhận một giá trị xác định bằng 0:  $P(X = x) = 0$

(c) Đối với đại lượng ngẫu nhiên liên tục  $X$  ta có các đẳng thức:  
 $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$

(d) Nếu đại lượng ngẫu nhiên  $X$  chỉ nhận giá trị trong  $[a, b]$  thì với  $x \leq a, F(x) = 0$  và với  $x > b, F(x) = 1$ .

**Ý nghĩa của hàm phân phối xác suất:** Hàm phân phối xác suất phản ánh mức độ tập trung xác suất ở về phía bên trái một số thực  $x$  nào đó.

### c. Hàm mật độ xác suất

**Định nghĩa 2.1.13.** Hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục  $X$ , ký hiệu  $f(x)$ , là đạo hàm bậc nhất của hàm phân phối xác suất của đại lượng ngẫu nhiên đó.

$$f(x) = F'(x)$$

### Các tính chất của hàm mật độ xác suất

#### Tính chất 2.1.14.

(a)  $f(x) \geq 0, \quad \forall x$

(b)  $P(a < X < b) = \int_a^b f(x) dx$

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

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

**Ví dụ 2.1.15.** Đại lượng ngẫu nhiên liên tục  $X$  có hàm mật độ xác suất như sau:

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

a/ Tìm hệ số  $a$ .

b/ Tìm hàm phân phối xác suất  $F(x)$ .

c/ Tìm  $P(0 < X < \frac{\pi}{4})$

**Giải:**

a/ Theo tính chất của hàm mật độ xác suất ta có  $a \geq 0$  và

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

b/ Để tìm  $F(x)$  ta áp dụng tính chất:  $F(x) = \int_{-\infty}^x f(x) dx$ .

+ Với  $x < -\frac{\pi}{2}$ ;  $F(x) = \int_{-\infty}^x 0 dx = 0$

+ Với  $-\frac{\pi}{2} < x < \frac{\pi}{2}$ , ta có:

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

+ Với  $x > \frac{\pi}{2}$  ta có

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

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

$$c/ P(0 < X < \frac{\pi}{4}) = \int_0^{\pi/4} \frac{1}{2} \cos x dx = \frac{\sqrt{2}}{4}.$$

**Ví dụ 2.1.16.** Cho hàm phân phối của biến ngẫu nhiên X có dạng:

$$F(x) = a + b \arctg x \quad (-\infty < x < +\infty)$$

Tìm a, b và  $f(x)$ .

*Giải:* Áp dụng tính chất 3 của hàm phân phối xác suất

$$\begin{cases} \lim_{x \rightarrow -\infty} (a + b \arctg x) = 0 \\ \lim_{x \rightarrow +\infty} (a + b \arctg x) = 1 \end{cases} \Leftrightarrow \begin{cases} a - \frac{b\pi}{2} = 0 \\ a + \frac{b\pi}{2} = 1 \end{cases} \Rightarrow a = \frac{1}{2}; \quad b = \frac{1}{\pi}$$

Vậy  $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctg x$  và do đó  $f(x) = F'(x) = \frac{1}{\pi(1+x^2)}$ .

**Ví dụ 2.1.17.** Đại lượng ngẫu nhiên liên tục X có hàm mật độ xác suất:

$$f(x) = \frac{A}{e^x + e^{-x}}$$

a/ Hãy xác định hệ số A.

b/ Tìm hàm phân phối xác suất F(x).

c/ Tìm xác suất để trong 3 phép thử độc lập có 1 lần X nhận giá trị trong khoảng (-1; 1)

*Giải:*



$$a/ \text{ Áp dụng tính chất } \begin{cases} f(x) \geq 0 \\ \int_{-\infty}^{+\infty} f(x) dx = 1 \end{cases} \Leftrightarrow \int_{-\infty}^{+\infty} \frac{A}{e^x + e^{-x}} dx = 1 \Leftrightarrow A = \frac{2}{\pi}$$

$$b/ F(x) = \int_{-\infty}^x \frac{2/\pi}{e^x + e^{-x}} dx = \frac{2}{\pi} \lim_{b \rightarrow -\infty} \int_b^x \frac{de^x}{e^{2x} + 1} = \frac{2}{\pi} \arctg e^x$$

$$c/ P(-1 < X < 1) = F(1) - F(-1) = \frac{2}{\pi} (\arctg e - \arctg e^{-1}) = p$$

$$P_3(1) = C_3^1 p(1-p)^2$$

**Ý nghĩa của hàm mật độ xác suất:** Hàm mật độ xác suất của đại lượng ngẫu nhiên  $X$  tại mỗi điểm  $x$  cho biết mức độ tập trung xác suất tại điểm đó.

## 2.2. Các tham số đặc trưng của đại lượng ngẫu nhiên

### 2.2.1. Kỳ vọng toán

#### a. Các định nghĩa

**Định nghĩa 2.2.1.**  $X$  là đại lượng ngẫu nhiên rời rạc nhận một trong các giá trị có thể có  $x_1, x_2, \dots, x_n$  với các xác suất tương ứng  $p_1, p_2, \dots, p_n$ . Kỳ vọng toán của đại

lượng ngẫu nhiên rời rạc  $X$ , ký hiệu  $E(X)$  được xác định  $E(X) = \sum_{i=1}^n x_i p_i$

**Định nghĩa 2.2.2.** Nếu  $X$  là đại lượng ngẫu nhiên liên tục với hàm mật độ xác suất

$f(x)$  và miền giá trị là  $[a, b]$  thì  $E(X) = \int_a^b x \cdot f(x) dx$ .

**Ví dụ 2.2.3.** Tìm kỳ vọng toán của đại lượng ngẫu nhiên rời rạc  $X$  có bảng phân phối xác suất như sau:

X	1	2	3
P	0,8	0,16	0,04

*Giải:*

$$E(X) = 1 \cdot 0,8 + 2 \cdot 0,16 + 3 \cdot 0,04 = 1,24$$

**Ví dụ 2.2.4.** Tìm kỳ vọng toán của đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \frac{3}{4}(x^2 + 2x) & x \in (0,1) \\ 0 & x \notin (0,1) \end{cases}$$

$$\text{Giải: } E(x) = \int_0^1 \frac{3}{4}(x^2 + 2x) x dx = \frac{11}{16}$$

## Các tính chất của kỳ vọng toán

### Tính chất 2.2.5.

- (a)  $E(C) = C$  ( $c = const$ ).
- (b)  $E(CX) = CE(X)$  ( $C = const$ )
- (c)  $E(X + Y) = E(X) + E(Y)$
- (d)  $E(X.Y) = E(X).E(Y)$  Nếu X và Y độc lập.

### Bản chất và ý nghĩa của kỳ vọng toán

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

$$\bar{X} = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{n} = x_1 \frac{n_1}{n} + x_2 \frac{n_2}{n} + \dots + x_k \frac{n_k}{n}$$

$$\bar{X} = x_1 f_1 + x_2 f_2 + \dots + x_k f_k$$

Theo định nghĩa thống kê về xác suất với  $n$  đủ lớn ta có:

$$\bar{X} = x_1 p_1 + x_2 p_2 + \dots + x_k p_k = E(X)$$

*Vậy: Kỳ vọng toán của đại lượng ngẫu nhiên gần bằng trung bình số học của các giá trị quan sát của đại lượng ngẫu nhiên. Nó phản ánh giá trị trung tâm của phân phối xác suất của đại lượng ngẫu nhiên.*

**Ví dụ 2.2.6.** Có 5000 người xét nghiệm máu để tìm ký sinh trùng sốt rét. Tỷ lệ mắc bệnh ở địa phương theo thống kê là 10%. Có thể làm xét nghiệm theo hai phương pháp.

+ Phương pháp 1: Xét nghiệm từng người

+ Phương pháp 2: Lấy máu 10 người một trộn lẫn làm một xét nghiệm. Nếu kết quả xét nghiệm là âm tính (vô trùng) thì thấy qua 10 người không ai mắc bệnh. Nếu kết quả xét nghiệm là dương tính thì chứng tỏ trong 10 người đó có ít nhất một người mắc bệnh. Lúc đó phải làm thêm 10 xét nghiệm lẻ để phát hiện người có bệnh cụ thể. Hỏi làm theo cách nào lợi hơn.

*Giải:*

Theo phương pháp 1 thì phải làm 5000 xét nghiệm.

Theo phương pháp 2: Gọi X là số xét nghiệm phải làm đối với mỗi nhóm 10 người, X = 1 (Nếu kết quả là âm tính); X = 11 (Nếu kết quả là dương tính)

$$p_1 = P(X = 1) = (1 - 0,1)^{10} = (0,9)^{10}$$

$$p_2 = P(X = 11) = 1 - p_1 = 1 - (0,9)^{10}$$

X	1	11
P	$(0,9)^{10}$	$1 - (0,9)^{10}$

$$E(X) = (0,9)^{10} + 11 \cdot (1 - (0,9)^{10}) \approx 7,51$$

Tức là trung bình phải làm 7,51 ca xét nghiệm cho mỗi nhóm 10 người. Vậy theo cách 2 phải làm  $7,51 \cdot 500 = 3755$  xét nghiệm.

Kết luận: Làm theo phương pháp 2 lợi hơn.

### b. Phương sai

**Định nghĩa 2.2.7.** Phương sai của đại lượng ngẫu nhiên X, ký hiệu D(X), là kỳ vọng toán của bình phương sai lệch của đại lượng ngẫu nhiên so với kỳ vọng toán của nó:  $D(X) = E[X - E(X)]^2$ .

+ Nếu X là đại lượng ngẫu nhiên rời rạc thì

$$D(X) = \sum_{i=1}^n [X_i - E(X)]^2 \cdot p_i.$$

+ Nếu X là đại lượng ngẫu nhiên liên tục thì

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

Trong thực tế khi có một mẫu cụ thể ta thường dùng các công thức:

+ Nếu X là đại lượng ngẫu nhiên rời rạc:

$$D(X) = E(X^2) - [E(X)]^2 = \sum_{i=1}^n x_i^2 p_i - \left( \sum_{i=1}^n x_i p_i \right)^2$$

+ Nếu X là đại lượng ngẫu nhiên liên tục:

$$D(X) = \int_{-\infty}^{+\infty} x^2 f(x) dx - \left[ \int_{-\infty}^{+\infty} x f(x) dx \right]^2$$

**Ví dụ 2.2.8.** Trong một hộp kín có 17 quả cầu kích thước như nhau, trong đó có 9 quả màu trắng, 8 quả màu đen, lấy ngẫu nhiên ra 2 quả. Gọi X là số cầu đen được lấy ra:

a/ Lập dãy phân phối xác suất của X.

b/ Tìm hàm phân phối xác suất của X.

c/ Tính E(X), D(X).

*Giải:* Vì X là số cầu đen được lấy ra nên X là ĐLNN rời rạc nhận các giá trị 0;1;2, với xác suất tương ứng:

$$P(X = 0) = \frac{C_9^2}{C_{17}^2} = \frac{36}{136} \approx 0,265 \quad ; \quad P(X = 1) = \frac{C_9^1 C_8^1}{C_{17}^2} = \frac{72}{136} \approx 0,529$$

$$P(X = 2) = \frac{C_8^2}{C_{17}^2} = \frac{28}{136} \approx 0,206$$

a/ Dãy phân phối xác suất của X:

X	0	1	2
P	0,265	0,529	0,206

$$\text{b/ Hàm phân phối } F(x) = \begin{cases} 0 & \text{Khi } x \leq 0 \\ 0,265 & \text{Khi } 0 < x \leq 1 \\ 0,794 & \text{Khi } 1 < x \leq 2 \\ 1 & \text{Khi } x > 2 \end{cases}$$

$$\text{c/ } E(X) = 1,0,529 + 2,0,206 = 0,941;$$

$$D(X) = 0,529 + 4,0,206 - (0,941)^2 = 0,468$$

**Ví dụ 2.2.9.** Cho X là đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất

$$f(x) = \begin{cases} \frac{3}{4}x(2-x) & \text{khi } x \in [0; 2] \\ 0 & x \notin [0; 2] \end{cases}$$

a/ Tính E(X), D(X).

b/ Tính P{0,9 < X < 1,1} và P{X > 1,5}

*Giải:*

$$\text{a/ } E(X) = \int_0^2 \frac{3}{4}x^2(2-x)dx = \frac{3}{4} \left( \frac{2x^3}{3} - \frac{x^4}{4} \right) \Big|_0^2 = 1$$

$$D(X) = \int_0^2 \frac{3}{4}x^3(2-x)dx - \left( \int_0^2 \frac{3}{4}x^2(2-x)dx \right)^2 = \frac{1}{5}$$

$$\text{b/ } P(1,5 < X < 2) = \int_{1,5}^2 \frac{3}{4}x(2-x)dx = \frac{3}{4} \left( x^2 - \frac{x^3}{3} \right) \Big|_{1,5}^2 = 0,15625$$

$$P(0,9 < X < 1,1) = \int_{0,9}^{1,1} \frac{3}{4}x(2-x)dx = \frac{3}{4} \left( x^2 - \frac{x^3}{3} \right) \Big|_{0,9}^{1,1} = 0,1495$$

**Các tính chất của phương sai**

**Tính chất 2.2.10.**

$$(a) D(C) = 0 \quad (C = \text{Const})$$

$$(b) D(CX) = C^2D(X), \quad (C = \text{const})$$

$$(c) D(X + Y) = D(X) + D(Y), \quad \text{với } X \text{ và } Y \text{ độc lập.}$$

**Hệ quả 2.2.11.**  $D(X - Y) = D(X) + D(Y)$ , với X và Y độc lập.

**Bản chất và ý nghĩa của phương sai.**

Phương sai chính là trung bình số học của bình phương các sai lệch giữa các giá trị có thể có của đại lượng ngẫu nhiên so với giá trị trung bình của các giá trị đó. Nó phản ánh mức độ phân tán của các giá trị của đại lượng ngẫu nhiên xung quanh giá trị trung tâm của nó là kỳ vọng toán.

**c. Độ lệch tiêu chuẩn**

**Định nghĩa 2.2.12.** Căn bậc hai dương của phương sai được gọi là *độ lệch tiêu chuẩn*.

$$\text{Ký hiệu: } \sigma_X = \sqrt{D(X)} \Rightarrow D(X) = \sigma_X^2$$

**Chú ý 2.2.13.** Khi cần đánh giá mức độ phân tán của đại lượng ngẫu nhiên theo đơn vị đo của nó người ta thường tính độ lệch tiêu chuẩn chứ không phải là phương sai vì độ lệch tiêu chuẩn có cùng đơn vị đo với đại lượng ngẫu nhiên cần nghiên cứu.

## 2.3. Một số quy luật phân phối xác suất thông dụng

### 2.3.1. Quy luật không - một

**Định nghĩa 2.3.1.** Đại lượng ngẫu nhiên rời rạc X nhận một trong hai giá trị có thể có X = 0; 1 với các xác suất tương ứng được tính bằng công thức  $p_x = p^x q^{1-x}$  với x = 0; 1 gọi là *phân phối theo qui luật không - một* với tham số là p.

Ký hiệu: A(p)

#### Các tham số đặc trưng của qui luật không - một

Nếu X là đại lượng ngẫu nhiên phân phối theo qui luật không - một thì:

$$E(X) = p; \quad D(X) = pq$$

### 2.3.2. Phân phối nhị thức

**Định nghĩa 2.3.2.** Đại lượng ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có 0, 1, 2, ..., n với các xác suất tương ứng được tính theo công thức Bernoulli gọi là *phân phối theo qui luật nhị thức* với các tham số là n và p.

Kí hiệu: B (n,p)

Ta có bảng phân phối xác suất của đại lượng ngẫu nhiên có phân phối nhị thức là

X	0	1	...	i	...	n
P	$C_n^0 p^0 q^n$	$C_n^1 p q^{n-1}$		$C_n^i p^i q^{n-i}$		$C_n^n p^n q^0$

#### Các tham số đặc trưng của qui luật nhị thức

Nếu X là đại lượng ngẫu nhiên tuân theo qui luật phân phối nhị thức thì:

$$E(X) = np, \quad D(X) = npq$$

**Ví dụ 2.3.3.** Xác suất để một người bắn trúng bia là 0,8 . Tìm số viên đạn trúng bia trung bình khi người ấy bắn 6 viên đạn.

*Giải:* Bài toán thoả mãn dãy phép thử Bernoulli. Gọi X là số viên đạn trúng bia thì X phân phối theo qui luật nhị thức với các tham số n = 6; p = 0,8.

$$E(X) = 6 \cdot 0,8 = 4,8.$$

Vậy số viên đạn trung bình bắn trúng bia là 5.

### 2.3.3. Qui luật poisson

**Định nghĩa 2.3.4.** Đại lượng ngẫu nhiên rời rạc X nhận một trong các giá trị có thể có 0,1,2,... với xác suất tương ứng được tính bởi  $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ ,  $k = 0, 1, 2, \dots$

được gọi là phân phối theo qui luật Poisson với tham số là  $\lambda$  ( $\lambda = np$ ).

Ký hiệu:  $P(\lambda)$

Phân phối Poisson xuất hiện trong dãy phép thử Bernoulli khi số phép thử khá lớn và xác suất p khá bé.

**Các tham số đặc trưng của qui luật Poisson**

Nếu X là đại lượng ngẫu nhiên tuân theo qui luật Poisson thì:

$$E(X) = \lambda \qquad D(X) = \lambda$$

**Ví dụ 2.3.5.** Xác suất để trong khi vận chuyển mỗi chai rượu bị vỡ là 0,001. Người ta tiến hành vận chuyển 2000 chai rượu đến cửa hàng. Tìm số chai vỡ trung bình khi vận chuyển.

*Giải:* Bài toán thỏa mãn dãy phép thử Bernoulli với  $n = 2000$ ,  $P = 0,001$  (khá nhỏ). Ta có  $\lambda = 2000 \cdot 0,001 = 2$  (không đổi).

Gọi X là số chai rượu bị vỡ khi vận chuyển thì X là đại lượng ngẫu nhiên phân phối theo qui luật Poisson.

Vậy số chai bị vỡ trung bình là:  $E(X) = 2$  (chai).

**2.3.4. Phân phối chuẩn  $N(a, \sigma^2)$**

**Định nghĩa 2.3.6.** Đại lượng ngẫu nhiên liên tục X nhận các giá trị trong khoảng  $(-\infty; +\infty)$  gọi là phân phối theo qui luật chuẩn (hay phân phối chính qui) với các tham số là a và  $\sigma^2$  nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \qquad (2.1)$$

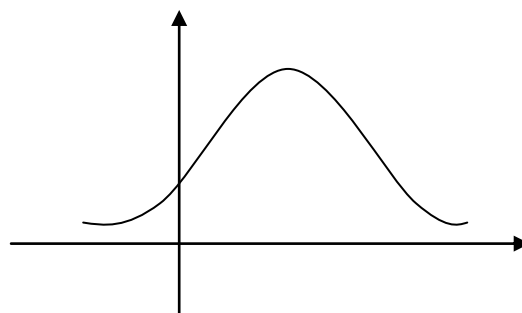
Ký hiệu :  $N(a, \sigma^2)$ .

**Chú ý 2.3.7.**

(a) a và  $\sigma^2$  là hai tham số đặc trưng của phân phối chuẩn.

(b) Đồ thị của hàm (2.1) có dạng hình chuông (h1) nhận trục hoành làm đường tiệm cận,  $f(x) > 0$  với mọi x,  $f_{max} = \frac{1}{\sigma\sqrt{2\pi}}$  đạt tại  $x = a$ .

Đồ thị hàm số có 2 điểm uốn với hoành độ  $x = a \pm \sigma$  và tung độ  $f = \frac{1}{\sigma\sqrt{2\pi e}}$



Khi  $a$  thay đổi thì dạng của đường cong  $f(x)$  không thay đổi. Nó dịch chuyển sang phải nếu  $a$  giảm. Khi  $\sigma$  thay đổi thì dạng của đồ thị thay đổi theo. Nếu  $\sigma$  tăng lên thì đồ thị sẽ thấp xuống và phình ra còn khi  $\sigma$  giảm thì đồ thị sẽ cao lên và nhọn thêm.

### Các tham số đặc trưng của qui luật chuẩn

$$E(X) = a; D(X) = \sigma^2$$

Từ đó ta thấy  $\sigma$  là thước đo độ tản mát các giá trị của biến ngẫu nhiên qua tâm phân phối là  $a$ .

### Tính xác suất trong phân phối chuẩn

Giả sử  $X : N(a, \sigma^2)$ . Hãy tính  $P(\alpha < X < \beta)$ , trong đó  $(\alpha, \beta)$  là khoảng cho trước tùy ý. Theo tính chất của hàm mật độ ta có:

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

Đặt  $\frac{x-a}{\sigma} = t \Rightarrow x = \sigma t + a \Rightarrow dx = \sigma dt$ , do đó

$$P(\alpha < X < \beta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\frac{\alpha-a}{\sigma}}^{\frac{\beta-a}{\sigma}} e^{-\frac{t^2}{2}} dt = \Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right), \quad (2.2)$$

với  $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$  là hàm số Laplatxơ, (với các giá trị được tính sẵn trong bảng phụ lục 2).

Xét trường hợp đặc biệt khi  $(\alpha, \beta)$  là khoảng đối xứng đối với  $a$  tức là khoảng  $(a - \alpha; a + \alpha)$ , ta có

$$P(|X - a| < \alpha) = 2\Phi\left(\frac{\alpha}{\sigma}\right), \quad (2.3)$$

**Ví dụ 2.3.8.** Biết độ dài chi tiết  $X$  do một máy tự động sản xuất ra tuân theo qui luật chuẩn  $N(20; 0,04)$ . Tính  $P(|X - 20| < 0,3)$ .

*Giải:* Theo công thức (2.3) ta có:

$$\begin{aligned} P(|X - 20| < 0,3) &= 2\Phi\left(\frac{0,3}{0,2}\right) = 2\Phi(1,5) \\ &= 2 \cdot 0,4332 = 0,8664 \end{aligned}$$

với 0,4332 là giá trị của hàm  $\Phi(1,5)$  tra ở bảng hàm số Laplatxơ. Vậy có khoảng 87% chi tiết sản xuất ra có kích thước nằm trong khoảng  $(19,7; 20,3)$ .

**Ví dụ 2.3.9.** Chiều cao của nam giới khi trưởng thành ở một vùng dân cư là đại lượng ngẫu nhiên phân phối chuẩn với  $a = 160\text{cm}$  với  $\sigma = 6\text{cm}$ . Một thanh niên bị coi là lùn nếu có chiều cao nhỏ hơn 155 cm.

a/ Tìm tỷ lệ thanh niên bị lùn ở vùng đó.

b/ Tìm xác suất để lấy ngẫu nhiên 4 người thì có ít nhất 1 người không bị lùn.

*Giải:*

a/ Gọi  $X$  là chiều cao của nam thanh niên khi trưởng thành ở vùng đó. Theo giả thiết  $X : N(160; 36)$ . Tỷ lệ thanh niên bị lùn ở vùng đó là xác suất để lấy ngẫu nhiên 1 người thì người đó có kích thước nhỏ hơn 155 cm. Ta có:

$$\begin{aligned} P(0 < X < 155) &= \Phi\left(\frac{155-160}{6}\right) - \Phi\left(\frac{0-160}{6}\right) \\ &= \Phi\left(\frac{160}{6}\right) - \Phi\left(\frac{5}{6}\right) \\ &= 0,4999 - 0,2967 = 0,2033 \end{aligned}$$

b/ Sử dụng công thức Bernoulli ta có:

$$P_4(1; 4) = 1 - P_4(0) = 1 - C_4^0(1 - 0,2033)^0 \cdot (0,2033)^4 = 1 - 0,0017 = 0,9983$$

### Phân phối chuẩn hoá

Trong trường hợp  $a=0$  và  $\sigma=1$ , ta có phân phối chuẩn hoá, khi đó hàm mật độ xác suất của đại lượng ngẫu nhiên  $X$  có dạng:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

đây là hàm số Gauss với các giá trị của nó được tính sẵn trong bảng phụ lục 1.

#### Chú ý 2.3.10.

+ Đồ thị của hàm này nhận trục tung làm trục đối xứng;

+ Hàm phân phối có dạng  $F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{u^2}{2}} du$  (Các giá trị của nó được tính

ở bảng phụ lục 3).

Ký hiệu của phân phối chuẩn hoá là  $N(0,1)$ .

#### Quy tắc Ba xích ma

Trong công thức  $P(|X - a| < \alpha) = 2\Phi\left(\frac{\alpha}{\sigma}\right)$  nếu ta đặt  $\alpha = 3\sigma$  thì:

$$P(|X - a| \leq 3\sigma) = 2\Phi(3) = 2 \cdot 0,49865 = 0,9973$$

Xác suất này rất gần 1 nên có thể coi hầu như chắc chắn biến cố  $|X - a| \leq 3\sigma$  sẽ xảy ra, từ đó ta có quy tắc: Nếu đại lượng ngẫu nhiên  $X$  có phân phối chuẩn  $X : N(a, \sigma^2)$  thì hầu như chắc chắn rằng  $X$  sẽ nhận giá trị trong khoảng  $(a - 3\sigma; a + 3\sigma)$ .

### 2.3.5. Quy luật Student - T(n)

Giả sử  $U$  là đại lượng ngẫu nhiên phân phối chuẩn hóa,  $V$  là đại lượng ngẫu nhiên độc lập với  $U$  phân phối theo quy luật khi bình phương với  $n$  bậc tự do. Xét



đại lượng ngẫu nhiên  $T = \frac{U}{\sqrt{\frac{V}{n}}}$ , đại lượng ngẫu nhiên T sẽ phân phối theo một quy

luật phân phối xác suất gọi là *quy luật Student* với n bậc tự do.

Ký hiệu: T(n).

Cũng giống như phân phối chuẩn hóa, phân phối T đối xứng qua gốc O (tức là có trung bình bằng 0). Khi n bé thì phân phối T có đường cong mật độ "mập" hơn đường mật độ N(0, 1), nhưng khi n khá lớn nó rất gần với chuẩn hóa. Trong thực tế nếu  $n \geq 30$  thì đã có thể coi phân phối T và chuẩn hóa là như nhau.

## 2.4. Các định lý về giới hạn

**2.4.1. Định lý 2.4.1. (Định lý Moavolaplat)** Nếu trong mỗi phép thử độc lập biến cố A xuất hiện với xác suất  $P(A)=P, (0 < P < 1)$  thì khi  $n \rightarrow \infty$  ta có:

$$\lim_{n \rightarrow \infty} \left[ P_n(k) - \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}} \right] = 0$$

Vậy với n khá lớn ta có:  $P_n(k) \approx \frac{1}{\sqrt{npq}} \varphi(x)$ , với  $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$  (hàm

Gauxơ),  $x = \frac{k - np}{\sqrt{npq}}$ .

**Ví dụ 2.4.1.** Xác suất để một cây bị chết khi trồng là 0,2 Tính xác suất để khi trồng 400 cây có 80 cây bị chết.

*Giải:* Ta có  $n=400$ ;  $k=80$ ;  $P(A)=0,2$  nên  $q = 0,8$ , nên  $x = \frac{80 - 400 \cdot 0,2}{\sqrt{400 \cdot 0,8 \cdot 0,2}} = 0$ . Tra

bảng Gauxơ ta được  $\varphi(0) = 0,3989$ , vậy

$$P_{400}(80) = \frac{1}{\sqrt{400 \cdot 0,2 \cdot 0,8}} \cdot 0,3989 \approx 0,05$$

**2.4.2. Định lý 2.4.2. (Định lý giới hạn tích phân)** Với các điều kiện như định lý trên ta có:

$$\lim_{n \rightarrow \infty} \left[ P_n(k_1, k_2) - \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{t^2}{2}} dt \right] = 0, \text{ trong đó } x_i = \frac{k_i - np}{\sqrt{npq}}, i=1,2.$$

Với  $n$  khá lớn ta có:  $P_n(k_1, k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{t^2}{2}} dt = \Phi(x_2) - \Phi(x_1)$ , trong đó

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt \text{ là hàm Laplace.}$$

**2.4.3. Định lý 2.4.3. (Định lý Poisson)** Giả sử tiến hành  $n$  phép thử độc lập. Mỗi phép thử sự kiện  $A$  xuất hiện với xác suất  $P(A)=p$ . Nếu  $n \rightarrow \infty$  mà  $p \rightarrow 0$  sao cho

$$np = \lambda = \text{const} \text{ thì ta có: } \lim_{n \rightarrow \infty} P_n(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

$$\text{Vậy khi } n \text{ khá lớn ta có: } P_n(k) \approx \frac{\lambda^k \cdot e^{-\lambda}}{k!}.$$

**Chú ý 2.4.4.** Trong trường hợp  $p$  rất gần 1 thì  $P(\bar{A}) = 1 - p$  rất gần 0. Do đó để tính  $P_n(k)$  ta chuyển sang tính xác suất cho  $n$  phép thử  $\bar{A}$  xuất hiện  $n-k$  lần.

**Ví dụ 2.4.5.** Một công nhân đứng máy xe xơi gồm 800 ống xác suất để mỗi ống xơi bị đứt xơi trong 1 giờ là 0,005. Tính xác suất

a/ Trong 1 giờ có 3 ống xơi bị đứt.

b/ Trong 1 giờ có không quá 10 ống bị đứt xơi.

Giải:

$$a/ n = 800; \quad p = 0,005; \quad \lambda = n.p = 4$$

$$\text{Vậy } P_{800}(3) = \frac{4^3 \cdot e^{-4}}{3!} \approx 0,1954$$

$$b/ P_{800}(0,10) = \sum_{k=0}^{10} P_{800}(k) = 0,99716$$

## 2.5. Đại lượng ngẫu nhiên hai chiều

### 2.5.1. Khái niệm về đại lượng ngẫu nhiên hai chiều

Ta ký hiệu đại lượng ngẫu nhiên hai chiều là  $(X, Y)$  trong đó  $X$  và  $Y$  được gọi là các thành phần của đại lượng ngẫu nhiên hai chiều thực chất mỗi thành phần lại là một đại lượng ngẫu nhiên một chiều. Vậy đại lượng ngẫu nhiên hai chiều thực chất là hệ hai đại lượng ngẫu nhiên  $X$  và  $Y$  được xét một cách đồng thời. Có hai loại đại lượng ngẫu nhiên:

+ Đại lượng ngẫu nhiên hai chiều là rời rạc nếu các thành phần của nó là rời rạc.

+ Đại lượng ngẫu nhiên hai chiều là liên tục nếu các thành phần của nó là liên tục.

### 2.5.2. Bảng phân phối xác suất của đại lượng ngẫu nhiên hai chiều.

Đối với đại lượng ngẫu nhiên hai chiều người ta cũng dùng bảng phân phối xác suất, hàm phân phối xác suất và hàm mật độ xác suất để thiết lập qui luật phân phối xác suất của chúng.

Bảng phân phối xác suất của đại lượng ngẫu nhiên hai chiều rời rạc liệt kê các giá trị có thể có của nó và các xác suất tương ứng. Nó có dạng sau đây:

X Y	x <sub>1</sub>	x <sub>2</sub>	...	x <sub>i</sub>	...	x <sub>n</sub>
y <sub>1</sub>	p(x <sub>1</sub> ,y <sub>1</sub> )	p(x <sub>2</sub> ,y <sub>1</sub> )	...	p(x <sub>i</sub> ,y <sub>1</sub> )	...	p(x <sub>n</sub> ,y <sub>1</sub> )
y <sub>2</sub>	p(x <sub>1</sub> ,y <sub>2</sub> )	p(x <sub>2</sub> ,y <sub>2</sub> )	...	p(x <sub>i</sub> ,y <sub>2</sub> )	...	p(x <sub>n</sub> ,y <sub>2</sub> )
...	...	...	...	...	...	...
y <sub>i</sub>	p(x <sub>1</sub> ,y <sub>i</sub> )	p(x <sub>2</sub> ,y <sub>i</sub> )	...	p(x <sub>i</sub> ,y <sub>i</sub> )	...	p(x <sub>n</sub> ,y <sub>i</sub> )
...	...	...	...	...	...	...
y <sub>m</sub>	p(x <sub>1</sub> ,y <sub>m</sub> )	p(x <sub>2</sub> ,y <sub>m</sub> )	...	p(x <sub>i</sub> ,y <sub>m</sub> )	...	p(x <sub>n</sub> ,y <sub>m</sub> )

trong đó x<sub>i</sub> (i = 1..n) là các giá trị có thể có của thành phần X, y<sub>j</sub> (j = 1..m) là các giá trị có thể có của thành phần Y; còn p(x<sub>i</sub>,y<sub>j</sub>) là xác suất để đại lượng ngẫu nhiên hai chiều (X, Y) nhận giá trị (x<sub>i</sub>,y<sub>j</sub>).

**Chú ý 2.5.1.** Các xác suất p(x<sub>i</sub>,y<sub>j</sub>) phải thỏa mãn điều kiện

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

Biết bảng phân phối xác suất của đại lượng ngẫu nhiên hai chiều bao giờ cũng có thể tìm được bảng phân phối xác suất của mỗi thành phần.

+ Bảng phân phối xác suất của thành phần X có dạng:

X	x <sub>1</sub>	x <sub>2</sub>	...	x <sub>i</sub>	...	x <sub>n</sub>
p	p(x <sub>1</sub> )	p(x <sub>2</sub> )	...	p(x <sub>i</sub> )	...	p(x <sub>n</sub> )

Trong đó:  $p(x_i) = \sum_{j=1}^m p(x_i, y_j)$  i = 1..n và  $\sum_{i=1}^n p(x_i) = 1$ .

+ Bảng phân phối xác suất của thành phần Y có dạng:

Y	y <sub>1</sub>	y <sub>2</sub>	...	y <sub>i</sub>	...	y <sub>m</sub>
p	p(y <sub>1</sub> )	p(y <sub>2</sub> )	...	p(y <sub>i</sub> )	...	p(y <sub>m</sub> )

Trong đó:  $p(y_j) = \sum_{i=1}^n p(x_i, y_j)$  j = 1..m và  $\sum_{j=1}^m p(y_j) = 1$ .

**Ví dụ 2.5.2.** Tìm bảng phân phối xác suất của các thành phần của đại lượng ngẫu nhiên hai chiều có bảng phân phối xác suất như sau:

X \ Y			
	$x_1$	$x_2$	$x_3$
$y_1$	0,10	0,30	0,20
$y_2$	0,06	0,18	0,16

*Giải:* Cộng các xác suất theo cột ta thu được các xác suất tương ứng với các giá trị của thành phần X:  $p(x_1) = 0,16$ ;  $p(x_2) = 0,48$ ;  $p(x_3) = 0,36$ .

Ta có bảng phân phối xác suất của thành phần X như sau:

X	$x_1$	$x_2$	$x_3$
p	0,16	0,48	0,36

Cộng các xác suất theo dòng ta có các xác suất tương ứng với các giá trị của thành phần Y:  $p(y_1) = 0,6$ ;  $p(y_2) = 0,4$ .

Ta có bảng phân phối xác suất của thành phần Y như sau:

Y	$y_1$	$y_2$
p	0,6	0,4

## BÀI TẬP CHƯƠNG 2

**1.** Một lô hàng gồm 7 sản phẩm trong đó có 3 phế phẩm. Chọn ngẫu nhiên ra 4 sản phẩm để kiểm tra. Gọi X là số sản phẩm tốt trong 4 sản phẩm lấy ra.

a/ Tìm quy luật phân phối xác suất của X .

b/ Tìm hàm phân phối xác suất.

c/ Tính  $E(X)$ ;  $D(X)$ .

**2.** Kiểm tra vấn đáp hết môn cho 4 học sinh, mỗi học sinh chỉ được vào kiểm tra nếu người được kiểm tra trước đó đạt yêu cầu. Xác suất đạt yêu cầu khi kiểm tra của mỗi học sinh là 0,6. Lập bảng phân phối xác suất, tìm hàm phân phối xác suất, tính kỳ vọng và phương sai của số học sinh được vào kiểm tra.

**3.** Trong một chiếc hòm có 5 bóng đèn trong đó có 2 bóng tốt và 3 bóng hỏng. Lấy ngẫu nhiên ra 2 bóng để kiểm tra. Gọi X là số bóng tốt trong số 2 bóng được kiểm tra.

a/ Hãy lập dãy phân phối xác suất của X.

b/ Tìm hàm phân phối  $F(x)$ ?

c/ Tìm  $E(X)$  và  $D(X)$ ?

**4.** Một túi chứa 10 tấm thẻ đỏ và 6 tấm thẻ xanh. Chọn ra 3 tấm thẻ. Gọi X là số thẻ đỏ được lấy ra.

a/ Lập bảng phân phối xác suất của X?

b/ Tìm hàm phân phối xác suất  $F(x)$ ?

c/ Tìm  $E(X)$  và  $D(X)$ ?

**5.** Một thiết bị gồm 3 bộ phận hoạt động độc lập với nhau. Xác suất trong thời gian  $t$  các bộ phận bị hỏng tương ứng là 0,4; 0,2 và 0,3. Gọi X là số bộ phận bị hỏng.

- a/ Tìm quy luật phân phối xác suất  $X$ .
- b/ Tìm hàm phân phối  $F(x)$ .
- c/ Tính  $E(X)$ ;  $D(X)$ .

**6.** Một xí nghiệp có hai ô tô vận tải hoạt động. Xác suất trong ngày làm việc các ô tô bị hỏng tương ứng là 0,1 và 0,2. Gọi  $X$  là số ô tô bị hỏng trong thời gian làm việc.

- a/ Tìm quy luật phân phối xác suất của  $X$ .
- b/ Tìm hàm phân phối xác suất.
- c/ Tính  $E(X)$ ;  $D(X)$ .

**7.** Một người đi từ nhà đến cơ quan phải qua 3 ngã tư, xác suất để người đó gặp đèn đỏ ở các ngã tư tương ứng là: 0,2; 0,4 và 0,5. Hỏi thời gian trung bình phải ngừng trên đường là bao nhiêu. Biết rằng mỗi khi gặp đèn đỏ người đó phải dừng mất 30 giây.

**8.** Trong phòng thí nghiệm có 3 nghiên cứu viên tiến hành 3 thí nghiệm độc lập về tế bào ung thư trong cùng một khoảng thời gian. Xác suất thực hiện thành công thí nghiệm của nghiên cứu viên thứ nhất là 0,75, nghiên cứu viên thứ hai là 0,8 và nghiên cứu viên thứ ba là 0,6. Gọi  $X$  là số thí nghiệm thành công trong ba thí nghiệm.

- a/ Lập bảng phân phối xác suất của  $X$ .
- b/ Tính kỳ vọng và phương sai.

**9.** Có 3 xạ thủ bắn độc lập vào cùng một bia, mỗi xạ thủ bắn 1 viên đạn. Xác suất bắn trúng đích của mỗi xạ thủ là 0,6; 0,5 và 0,4. Gọi  $X$  là biến ngẫu nhiên chỉ số viên đạn bắn trúng bia.

- a/ Lập bảng phân phối xác suất của biến ngẫu nhiên  $X$ .
- b/ Tính kỳ vọng, phương sai của biến ngẫu nhiên  $X$ .

**10.** Một xạ thủ có 4 viên đạn. Xạ thủ đó bắn lần lượt từng viên cho đến khi trúng mục tiêu hoặc hết cả 4 viên thì thôi. Xác suất bắn trúng mục tiêu của mỗi viên đạn là 0,6. Gọi  $X$  là đại lượng ngẫu nhiên chỉ số viên đạn đã bắn.

- a/ Lập bảng phân phối xác suất của  $X$ .
- b/ Tính kỳ vọng, phương sai của  $X$ .

**11.** Cho hàm số: 
$$f(x) = \begin{cases} \frac{x^3}{4}, & x \in (0; 2) \\ 0, & x \notin (0; 2) \end{cases}$$

- a/ Chứng minh hàm  $f(x)$  là hàm mật độ xác suất của một đại lượng ngẫu nhiên liên tục  $X$ .
- b/ Tính kỳ vọng, phương sai của đại lượng ngẫu nhiên có hàm mật độ xác suất nói trên.
- c/ Tính xác suất để trong 3 phép thử độc lập có 1 lần  $X$  nhận giá trị trong  $(1; 3/2)$ .

**12.** Biến ngẫu nhiên liên tục  $X$  có hàm mật độ xác suất:

$$f(x) = \begin{cases} \frac{6}{5}x(x-1) & \text{Khi } x \in [1; 2] \\ 0 & \text{Khi } x \notin [1; 2] \end{cases}$$

a/ Hãy tìm hàm phân phối  $F(x)$ .

b/ Tính  $E(X)$ .

c/ Tính xác suất  $P(0 < X < 1.5)$ .

**13.** Cho đại lượng ngẫu nhiên liên tục  $X$  có hàm mật độ xác suất:

$$f(x) = \begin{cases} a(x^3 + 2x + 1), & \text{khi } x \in [0; 4] \\ 0 & , \text{ khi } x \notin [0; 4] \end{cases}$$

a/ Tìm hệ số  $a$ ?

b/ Tìm hàm phân phối  $F(x)$ ?

c/ Tính  $P(1 < X < 3)$ ?

**14.** Cho đại lượng ngẫu nhiên liên tục  $X$  có hàm mật độ xác suất:

$$f(x) = \begin{cases} kx(4 - x^2), & \text{khi } x \in (0; 2) \\ 0 & , \text{ khi } x \notin (0; 2) \end{cases}$$

a/ Tìm hệ số  $k$ ?

b/ Tìm hàm phân phối  $F(x)$ ?

c/ Tính  $E(X)$ ?

**15.** Cho  $X$  là ĐLNN liên tục với hàm mật độ xác suất:

$$f(x) = \begin{cases} kx^2(1 - x) & , & \text{khi } x \in [0; 1] \\ 0 & , & \text{khi } x \notin [0; 1] \end{cases}$$

a/ Tìm hệ số  $k$ ?

b/ Tìm hàm phân phối  $F(X)$ ?

c/ Tìm  $P(-0,5 < X < 0,5)$

**16.** Cho đại lượng ngẫu nhiên liên tục  $X$  có hàm mật độ xác suất:

$$f(x) = \begin{cases} kx^2(4 - x), & \text{khi } x \in [0; 4] \\ 0 & , \text{ khi } x \notin [0; 4] \end{cases}$$

a/ Tìm hệ số  $k$ ?

b/ Tìm hàm phân phối  $F(x)$ ?

c/ Tính  $E(X)$ ?

**17.** Cho đại lượng ngẫu nhiên liên tục  $X$  có hàm mật độ xác suất:

$$f(x) = \begin{cases} k(1 - x^2), & \text{khi } x \in (-1; 1) \\ 0 & , \text{ khi } x \notin (-1; 1) \end{cases}$$

a/ Tìm hệ số  $k$ ?

b/ Tìm hàm phân phối  $F(x)$ ?

c/ Tính  $E(X)$ ?

# Phần 3. Thống kê

## Chương 1

### Cơ sở lý thuyết mẫu

#### 1.1. Tổng thể và mẫu

##### 1.1.1. Định nghĩa

###### a. Tổng thể

**Định nghĩa 1.1.1.** Toàn bộ tập hợp các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hoặc định lượng nào đó được gọi là tổng thể nghiên cứu hay tổng thể (hay tập chính). Số lượng các phần tử của tổng thể được gọi là *kích thước của tổng thể* (thường được ký hiệu là  $N$ ).

Với mỗi tổng thể ta không nghiên cứu trực tiếp tổng thể đó mà thông qua một hay nhiều dấu hiệu đặc trưng cho tổng thể đó, chúng được gọi là dấu hiệu nghiên cứu, các dấu hiệu này có thể là định tính hoặc định lượng.

###### b. Mẫu

**Định nghĩa 1.1.2.** Nếu từ tổng thể ta chọn ngẫu nhiên ra  $n$  phần tử thì tập hợp  $n$  phần tử này được gọi là mẫu kích thước  $n$ , khi đó ta sẽ tìm cách xem xét đánh giá mẫu đó rồi suy ra kết luận cho tổng thể.

###### c. Mẫu ngẫu nhiên:

Tiến hành  $n$  quan sát độc lập về biến ngẫu nhiên  $X$  nào đó gọi  $X_i$  là việc quan sát lần thứ  $i$  về biến ngẫu nhiên  $X$ . Khi đó  $(X_1, X_2, \dots, X_n)$  được gọi là mẫu ngẫu nhiên,  $n$  gọi là cỡ mẫu hay số lần quan sát. (mẫu ngẫu nhiên cỡ  $n$  thực chất là  $n$  biến ngẫu nhiên độc lập, cùng phân phối như biến ngẫu nhiên  $X$ )

**Định nghĩa 1.1.3.** Mẫu ngẫu nhiên kích thước  $n$  là tập hợp của  $n$  biến ngẫu nhiên độc lập  $X_1, X_2, \dots, X_n$  được thành lập từ biến ngẫu nhiên  $X$  và có cùng phân phối xác suất với  $X$ .

Mẫu ngẫu nhiên được ký hiệu  $W = (X_1, X_2, \dots, X_n)$ . Lúc đó việc thực hiện một phép thử đối với mẫu ngẫu nhiên  $W$  chính là thực hiện một phép thử đối với mỗi thành phần của mẫu. Giả sử  $X_1$  nhận giá trị  $x_1$ ,  $X_2$  nhận giá trị  $x_2, \dots, X_n$  nhận giá trị  $x_n$ . Tập hợp  $n$  giá trị  $x_1, x_2, \dots, x_n$  tạo thành một giá trị của mẫu ngẫu nhiên, hay còn gọi là một mẫu cụ thể, ký hiệu  $w = (x_1, x_2, \dots, x_n)$

**Ví dụ 1.1.4.** Gọi  $X$  là số chấm xuất hiện khi tung một con xúc xắc,  $X$  là ĐLNN với bảng phân phối xác suất như sau:

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

Nếu tung con xúc xắc 3 lần và gọi  $X_i$  là số chấm xuất hiện trong lần tung thứ  $i$  ( $i = 1, 2, 3$ ) thì ta có 3 ĐLNN độc lập có cùng qui luật phân phối xác suất với  $X$ . Vậy ta có mẫu ngẫu nhiên kích thước  $n = 3$ ,  $W = (X_1, X_2, X_3)$  được xây dựng từ ĐLNN gốc  $X$ . Thực hiện một phép thử đối với mẫu ngẫu nhiên này tức là tung cụ thể 3 lần: Giả sử lần thứ nhất được 3 chấm, lần thứ hai được 4 chấm, lần thứ ba được 2 chấm thì ta thu được một mẫu cụ thể  $w = (3, 4, 2)$ .

## 1.1.2. Phương pháp xây dựng mẫu

### a. Nhận xét

Từ kết quả tập mẫu có được ta có thể suy ra các kết quả cho tổng thể bởi vậy bao giờ cũng có thể mắc phải sai lầm nhất định. Độ sai lệch lớn hay bé phụ thuộc vào phương pháp xây dựng mẫu và kích thước mẫu. Độ chính xác trong thống kê thường được gọi là độ tin cậy của kết luận: ký hiệu là  $\gamma$ . Gọi  $\alpha$  là tỷ lệ sai sót (hay mức ý nghĩa) thì  $\alpha = 1 - \gamma$ .

Để có căn cứ vào thông tin của mẫu đưa ra những kết luận đủ chính xác về dấu hiệu nghiên cứu trong tổng thể, tức là phản ánh đúng đặc điểm của tổng thể theo dấu hiệu nghiên cứu đó. Để đảm bảo tính đại diện của mẫu và tiện cho việc mô hình hoá, mẫu được tạo lập với những giả thiết sau:

+ Lấy lần lượt từng phần tử vào mẫu, phương pháp này gọi là phương pháp đơn giản để phân biệt với cách lấy cùng một lúc nhiều phần tử vào mẫu.

+ Mỗi phần tử được lấy vào mẫu một cách hoàn toàn ngẫu nhiên, tức là mọi phần tử của tổng thể đều có thể được lấy vào mẫu với khả năng như nhau.

+ Các phần tử được lấy vào mẫu theo phương thức hoàn lại, tức là trước khi lấy phần tử thứ  $k$  thì trả lại tổng thể phần tử thứ  $(k - 1)$  mà ta đã nghiên cứu xong ( $k = 2..n$ ).

### Chú ý 1.1.5.

+ Trong việc lấy mẫu, do nhiều nguyên nhân khác nhau, sẽ không tránh khỏi các sai số trong số liệu mẫu. Vì vậy trước khi dùng các phương pháp thống kê để phân tích, xử lý ta cần loại bỏ các sai số không đáng có ở trong mẫu đã cho. Có 3 loại sai số

**1/ Sai số thô:** Là sai số sinh ra do vi phạm các điều kiện cơ bản của việc lấy mẫu, hoặc do sơ suất của người thực hiện. Chẳng hạn người kiểm tra cố ý chọn ra các sản phẩm tốt để kiểm tra khi đánh giá chất lượng, hoặc người kỹ thuật viên ghi nhầm kết quả thu được.

**2/ Sai số hệ thống:** Là sai số do không điều chỉnh chính xác dụng cụ hoặc không thống nhất giữa các kỹ thuật viên về cách xác định một đại lượng nào đó..., dẫn đến một loạt kết quả quan sát được bị lệch đi một tỷ lệ nhất định nào đó.



**3/ Sai số ngẫu nhiên:** Là sai số sinh ra do một số lớn các nguyên nhân mà tác dụng của chúng bé đến mức không tách riêng và tính riêng biệt cho từng nguyên nhân được.

Trong ba loại sai số trên, **sai số thô** và **sai số hệ thống** cần phát hiện sớm và khử bỏ ngay, còn sai số ngẫu nhiên thì không thể khử bỏ được trong mỗi lần quan sát. Do đó các kết quả quan sát được đưa vào xử lý bằng các phương pháp toán học ta sẽ giải quyết chúng chỉ chứa các sai số ngẫu nhiên.

+ Trong thực tế nếu kích thước của tổng thể khá lớn còn mẫu chỉ chiếm một phần rất nhỏ của tổng thể thì phương pháp lấy mẫu hoàn lại và không hoàn lại cho ta các kết quả sai lệch không đáng kể.

### **b. Một số phương pháp chọn mẫu**

\* **Mẫu ngẫu nhiên đơn:** là loại mẫu được chọn trực tiếp từ danh sách đã được đánh số của tổng thể. Từ tổng thể kích thước  $N$  người ta dùng cách rút thăm đơn giản ra  $n$  phần tử của mẫu theo một bảng số ngẫu nhiên nào đó.

+ Ưu điểm: Phương pháp này có ưu điểm là cho phép thu được một mẫu có tính đại diện cao cho phép suy rộng các kết quả của mẫu cho tổng thể với một sai số xác định.

+ Nhược điểm: Phải có được toàn bộ danh sách của tổng thể nghiên cứu, mặt khác chi phí chọn mẫu sẽ khá lớn.

\* **Mẫu ngẫu nhiên hệ thống:** Là loại mẫu ngẫu nhiên đã được đơn giản hoá trong cách chọn, trong đó chỉ có phần tử đầu tiên được chọn một cách ngẫu nhiên, sau đó dựa trên danh sách đã được đánh số của tổng thể để chọn ra các phần tử tiếp theo vào mẫu theo một thủ tục nào đó.

+ Nhược điểm của phương pháp này là dễ mắc sai số hệ thống khi danh sách của tổng thể không được sắp xếp một cách ngẫu nhiên mà lại theo một trật tự chủ quan nào đó.

\* **Mẫu chùm:** Tổng thể điều tra được phân chia ra thành nhiều chùm theo nguyên tắc:

+ Mỗi phần tử của tổng thể chỉ được phân vào một chùm.

+ Mỗi chùm cố gắng chứa nhiều phần tử khác nhau về dấu hiệu cần nghiên cứu sao cho nó có độ phân tán cao như của tổng thể.

+ Phân chia sao cho các chùm tương đối đồng đều nhau về qui mô.

+ Các chùm được chọn một cách ngẫu nhiên và tất cả các phần tử của chùm đó đều được chọn vào mẫu.

\* **Mẫu phân tổ:** Trong chọn mẫu phân tổ trước hết người ta phân chia tổng thể ra thành các tổ có độ thuần nhất cao để chọn ra các phần tử đại diện cho từng tổ. Việc phân tổ có hiệu quả khi tổng thể nghiên cứu không thuần nhất theo dấu hiệu nghiên cứu. Sau khi đã phân tổ thì kích thước mẫu được phân bố cho mỗi tổ theo một qui tắc nào đó.

\* **Mẫu nhiều cấp:** Nếu các phần tử của tổng thể phân tán quá rộng và thiếu thông tin về chúng, người ta thường chọn mẫu theo nhiều cấp, khi chọn nhiều cấp ta có nhiều loại đơn vị chọn mẫu ở mỗi cấp, thường được gọi là đơn vị chọn mẫu cấp 1, cấp 2,... Để chọn mẫu ở mỗi cấp chỉ cần có thông tin về phân bố của dấu hiệu ở cấp

ấy là đủ. Việc chọn mẫu ở mỗi cấp có thể tiến hành theo phương pháp mẫu ngẫu nhiên đơn, mẫu ngẫu nhiên hệ thống, mẫu chùm hay mẫu phân tổ.

### c. Các phương pháp sắp xếp mẫu ngẫu nhiên

Giả sử từ tổng thể của ĐLNN gốc  $X$  rút ra một mẫu ngẫu nhiên kích thước  $n$ , trong đó giá trị  $x_1$  xuất hiện với tần số  $n_1$ , giá trị  $x_2$  xuất hiện với tần số  $n_2, \dots$ , giá trị  $x_k$  xuất hiện với tần số  $n_k$ , lúc đó sau khi các  $x_i$  đã được sắp xếp theo trình tự tăng dần giá trị cụ thể của mẫu  $w$  có thể mô tả bằng bảng phân phối tần số thực nghiệm sau:

$x_i$	$x_1$	$x_2$	$\dots$	$x_i$	$\dots$	$x_k$
$n_i$	$n_1$	$n_2$	$\dots$	$n_i$	$\dots$	$n_k$

với  $n_1 + n_2 + \dots + n_k = n$ . Gọi  $f_i = \frac{n_i}{n}$ , ( $i = 1 \dots k$ )

Ta có bảng phân phối tần suất thực nghiệm:

$x_i$	$x_1$	$x_2$	$\dots$	$x_i$	$\dots$	$x_k$
$f_i$	$f_1$	$f_2$	$\dots$	$f_i$	$\dots$	$f_k$

với  $f_1 + f_2 + \dots + f_k = 1$

**Ví dụ 1.1.6.** Gặt ngẫu nhiên 100 điểm trồng lúa của một vùng thu được các số liệu được sắp xếp thành bảng sau (gọi là bảng phân phối tần số thực nghiệm).

Năng suất(tạ/ha)	21	24	25	26	28	32	34
Số điểm gặt tương ứng	10	20	30	15	10	10	5

Bảng phân phối tần suất thực nghiệm có dạng:

$x_i$	21	24	25	26	28	32	34
$f_i$	0,1	0,2	0,3	0,15	0,1	0,1	0,05

**Chú ý 1.1.7.** Nếu các phần tử của mẫu khá gần nhau khi đó ta có thể sắp xếp mẫu thành một dãy phân phối thực nghiệm ghép lớp.

## 1.2. Các đặc trưng của mẫu ngẫu nhiên

Để nghiên cứu đại lượng ngẫu nhiên gốc  $X$  trong tổng thể, nếu chỉ rút ra một mẫu ngẫu nhiên  $W = (X_1, X_2, \dots, X_n)$  thì mới chỉ có được một vài kết luận sơ bộ và rời rạc về  $X$ , vì các giá trị  $X_i$  của mẫu có cùng qui luật phân phối xác suất với  $X$  song qui luật này lại thường chưa được xác định hoàn toàn. Song nếu tổng hợp các

đại lượng ngẫu nhiên  $X_1, X_2, \dots, X_n$  này lại chúng sẽ bộc lộ những tính qui luật mới làm cơ sở để nhận định về đại lượng ngẫu nhiên gốc X trong tổng thể.

Việc tổng hợp mẫu  $W = (X_1, X_2, \dots, X_n)$  được thực hiện dưới dạng một hàm nào đó của các giá trị  $X_1, X_2, \dots, X_n$  của mẫu. Nó được gọi là **thống kê**.

Ký hiệu là G. Vậy  $G = f(X_1, X_2, \dots, X_n)$

Thực chất thống kê là một hàm của các đại lượng ngẫu nhiên, do đó bản thân nó cũng sẽ là một đại lượng ngẫu nhiên tuân theo một qui luật phân phối xác suất nhất định và có các tham số đặc trưng tương ứng. Mặt khác, khi mẫu ngẫu nhiên nhận một giá trị cụ thể  $w = (x_1, x_2, \dots, x_n)$  thì G cũng nhận một giá trị cụ thể là

$$g = f(x_1, x_2, \dots, x_n)$$

Sau đây ta sẽ nghiên cứu một số thống kê thông dụng nhất

### 1.2.1. Trung bình mẫu

Giả sử từ tổng thể của ĐLNN gốc X lập một mẫu ngẫu nhiên kích thước n.

$$W = (X_1, X_2, \dots, X_n).$$

Trung bình mẫu là một thống kê ký hiệu là  $\bar{X}$ :  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Vì tập hợp mẫu  $W = (X_1, X_2, \dots, X_n)$  được thực hiện dưới dạng một hàm nào đó của các giá trị  $X_1, X_2, \dots, X_n$  của mẫu nên nó là một thống kê do đó trung bình mẫu  $\bar{X}$  cũng là một thống kê. Vì vậy nó là một ĐLNN tuân theo một qui luật phân phối xác suất nào đó với các tham số đặc trưng tương ứng. Khi mẫu nhận một giá trị cụ thể  $w = (x_1, x_2, \dots, x_n)$  thì trung bình mẫu cũng nhận một giá trị cụ thể ký hiệu là  $\bar{x}$ .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ hoặc } \bar{x} = \frac{1}{n} \sum_{i=1}^k m_i x_i.$$

#### Tính chất 1.2.1.

Nếu  $E(X) = a$  và  $D(X) = \sigma^2$  thì  $E(\bar{X}) = a$  và  $D(\bar{X}) = \frac{\sigma^2}{n}$

**Chú ý 1.2.2.** Nếu các  $x_i$  cách đều nhau một khoảng là h thì

$$\bar{x} = x_0 + \frac{h}{n} \sum_{i=1}^k m_i \left( \frac{x_i - x_0}{h} \right)$$

trong đó  $x_0$  là giá trị tạo với  $m_i$  tương ứng đạt max

**Ví dụ 1.2.3.** Cho bảng phân phối tần số thực nghiệm (ví dụ 1). Hãy tính trung bình mẫu

*Giải:* Ta có

$$\bar{x} = \frac{1}{100} (21.10 + 24.20 + 25.30 + 26.15 + 28.10 + 32.10 + 34.5) = 26 \text{ (tạ/ha)}$$

**Ví dụ 1.2.4.** Cho mẫu

X	28	30	32	34	36	38	40	42
$m_i$	3	7	10	15	10	3	1	1

Hãy tìm kỳ vọng mẫu.

*Giải:* Nhận xét mẫu trên có kích thước  $n = 50$  và các  $x_i$  cách đều nhau một khoảng  $h = 2$ ;  $\max m_i = 15$  do đó chọn  $x_0 = 34$  ta được

$$\bar{x} = 34 + \frac{2}{50} (3 \cdot (-3) + 7 \cdot (-2) + 10 \cdot (-1) + 15 \cdot 0 + 10 \cdot 1 + 3 \cdot 2 + 1 \cdot 3 + 1 \cdot 4) = 33,6$$

Vậy  $E(\bar{x}) = 33,6$ .

### 1.2.2. Phương sai mẫu

Phương sai mẫu ký hiệu là  $S^2(X)$ .

Hoàn toàn tương tự như trung bình mẫu khi cho một mẫu cụ thể thì phương sai mẫu sẽ nhận một giá trị cụ thể. ký hiệu là  $s^2(x) = s^2$

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

Trong thực tế để tiện cho việc tính toán người ta thường hay sử dụng công thức

$$s^2 = \frac{1}{n} \sum_{i=1}^k m_i x_i^2 - (\bar{x})^2 \Leftrightarrow s^2 = \overline{x^2} - (\bar{x})^2$$

**Chú ý 1.2.5.** Nếu các  $x_i$  cách đều nhau một khoảng là  $h$  thì

$$s^2 = \frac{h^2}{n} \sum_{i=1}^k m_i \left( \frac{x_i - x_0}{h} \right)^2 - (x_0 - \bar{x})^2$$

**Tính chất 1.2.6.** Nếu  $E(X) = a$  và  $D(X) = \sigma^2$  thì  $E(S^2) = \frac{n-1}{n} \sigma^2$

### 1.2.3. Phương sai điều chỉnh mẫu

Phương sai điều chỉnh mẫu ký hiệu  $S'^2$

$$S'^2 = \frac{1}{n-1} \sum_{i=1}^k m_i (X_i - \bar{X})^2 = \frac{n}{n-1} S^2$$

**Chú ý 1.2.7.** Phương sai điều chỉnh mẫu  $S'^2$  là một thống kê còn khi có mẫu cụ thể thì phương sai mẫu điều chỉnh  $s'$  cũng là một số xác định.

### 1.2.4. Độ lệch tiêu chuẩn mẫu và độ lệch tiêu chuẩn điều chỉnh mẫu

+ Độ lệch tiêu chuẩn mẫu  $S = \sqrt{S^2}$

+ Độ lệch tiêu chuẩn điều chỉnh mẫu  $S' = \sqrt{S'^2}$

### 1.2.5. Tần suất mẫu

Giả sử từ tổng thể kích thước  $N$  trong đó có  $M$  phần tử mang dấu hiệu nghiên cứu, lấy ra một mẫu ngẫu nhiên kích thước  $n$  và trong đó thấy có  $X$  phần tử mang dấu hiệu nghiên cứu. Lúc đó tần suất mẫu là một thống kê ký hiệu là  $f$ ,

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

**Ví dụ 1.2.8.** Cho mẫu thực nghiệm

X	65	70	75	80	85	90	95	100	105	110	115	120	125
	70	75	80	85	90	95	100	105	110	115	120	125	130
$m_i$	1	0	2	5	8	16	18	17	16	9	5	2	1

Tìm  $\bar{x}$ ,  $s^2$ ,  $s$ .

*Giải:* Đặt  $x_i = \frac{x_i + x_{i+1}}{2}$  ( $i = 1, \dots, 13$ ), các  $x_i$  cách đều nhau một khoảng  $h = 5$  chọn

$x_0 = 97,5$ , đặt  $t_i = \frac{x_i - x_0}{5}$  với  $i = 1, \dots, 13$ .

Ta lập lại bảng

STT	$x_i$	$m_i$	$t_i$	$m_i t_i$	$t_i^2$	$m_i t_i^2$
1	67,5	1	- 6	- 6	36	36
2	72,5	0	- 5	0	25	0
3	77,5	2	- 4	- 8	16	32
4	82,5	5	- 3	- 15	9	45
5	87,5	8	- 2	- 16	4	32
6	92,5	16	- 1	- 16	1	16
7	97,5	18	0	0	0	0
8	102,5	17	1	17	1	17
9	107,5	16	2	32	4	64
10	112,5	9	3	27	9	81
11	117,5	5	4	20	16	80
12	122,5	2	5	10	25	50
13	127,5	1	6	6	36	36
$\Sigma$		100		51		489

Nhìn vào bảng trên ta thấy:

$$\bar{x} = 97,5 + \frac{5}{100} \cdot 51 = 100,05$$

$$s^2 = \frac{25}{100} \cdot 489 - (97,5 - 100,05)^2 \approx 115,75$$

$$s = \sqrt{s^2} = \sqrt{115,75} \approx 10,76$$

## Chương 2

### Ước lượng tham số

Giả sử khi xét một ĐLNN  $X$  ( $X$  là một dấu hiệu định lượng cần nghiên cứu của một tổng thể) ta cần biết qui luật phân phối xác suất của nó. Bằng phương pháp phân tích lý thuyết ta giả sử biết được dạng phân phối xác suất của nó. Tuy nhiên các tham số đặc trưng của nó như kỳ vọng, phương sai,.. mà ta gọi chung là tham số lý thuyết  $\theta$  lại chưa biết nên ta cần phải xác định  $\theta$ . Việc tính chính xác  $\theta$  là khó có thể thực hiện được mà ta chỉ có thể tính gần đúng. Việc tính gần đúng đó ta gọi là *ước lượng tham số  $\theta$*  và dựa vào mẫu thực nghiệm đã có.

Phương pháp tiến hành: Từ tổng thể cần nghiên cứu rút ra một mẫu ngẫu nhiên kích thước  $n$  và dựa vào mẫu đó mà xây dựng một thống kê  $G$  dùng để ước lượng  $\theta$  bằng cách này hay cách khác, có hai phương pháp sử dụng  $G$  để ước lượng  $\theta$  là phương pháp ước lượng điểm và phương pháp ước lượng bằng khoảng tin cậy.

#### 2.1. Phương pháp ước lượng điểm

Phương pháp ước lượng điểm chủ trương dùng một giá trị để thay thế cho tham số  $\theta$  chưa biết của tổng thể, vì bản thân  $\theta$  là một số xác định. Thông thường giá trị được chọn là một thống kê  $G$  nào đó của mẫu ngẫu nhiên. Có nhiều cách chọn thống kê  $G$  khác nhau tạo nên những phương pháp ước lượng điểm khác nhau.

##### 2.1.1. Phương pháp hàm ước lượng

###### a. Khái niệm

Giả sử cần ước lượng tham số  $\theta$  của đại lượng ngẫu nhiên gốc  $X$ . Từ tổng thể lập mẫu ngẫu nhiên kích thước  $n$ :  $W = (X_1, X_2, \dots, X_n)$ . Chọn lập thống kê  $G = f(X_1, X_2, \dots, X_n)$  mà thực chất là một thống kê đặc trưng mẫu tương ứng với tham số  $\theta$  cần ước lượng. Chẳng hạn, để ước lượng kỳ vọng toán  $a$  của ĐLNN gốc thì chọn thống kê trung bình mẫu  $\bar{X}$ , để ước lượng phương sai  $\sigma^2$  của ĐLNN gốc thì chọn thống kê phương sai mẫu  $S^2$ ... Nếu lập một mẫu cụ thể và tính được giá trị  $g = (x_1, x_2, \dots, x_n)$  của thống kê  $G$  trên mẫu cụ thể đó thì ước lượng của  $\theta$  là giá trị  $g$  vừa tính được.

Vì thống kê  $G = f(X_1, X_2, \dots, X_n)$  thực chất là hàm của các ĐLNN nên nó được gọi là hàm ước lượng của  $\theta$ . Chất lượng của ước lượng không thể đánh giá qua một giá trị cụ thể của  $G$ . Vì như vậy chỉ có cách so sánh trực tiếp  $g$  và  $\theta$  mà  $\theta$  lại chưa biết. Do đó chỉ có thể đánh giá chất lượng của ước lượng thông qua bản thân thống kê  $G = f(x_1, x_2, \dots, x_n)$ . Rõ ràng là có vô số cách chọn hàm  $f$ , tức là có vô số thống kê  $G$  có thể dùng làm hàm ước lượng của  $\theta$ . Vì vậy cần đưa ra một tiêu chuẩn để đánh giá chất lượng của thống kê  $G$ , từ đó lựa chọn được thống kê "xấp xỉ một cách tốt nhất" tham số cần ước lượng.

### b. Các tiêu chuẩn lựa chọn hàm ước lượng

#### \* Ước lượng không chệch

Giả sử thống kê  $G$  là ước lượng của tham số  $\theta$  của ĐLNN gốc. Với  $k$  mẫu cụ thể rút ra từ tổng thể, thống kê  $G$  sẽ nhận  $k$  giá trị cụ thể tương ứng là  $g_1, g_2, \dots, g_k$ . Nếu thống kê  $G$  là một ước lượng có dư của  $\theta$  thì các giá trị  $g_1, g_2, \dots, g_k$  cũng đều sẽ lớn hơn  $\theta$  và giá trị trung bình của chúng (tức là kỳ vọng toán của  $G$ ) cũng lớn hơn  $\theta$ :  $E(G) > \theta$ . Ngược lại, nếu thống kê  $G$  là một ước lượng thiếu của  $\theta$  thì mọi giá trị  $g_1, g_2, \dots, g_k$  cũng đều sẽ nhỏ hơn  $\theta$  nên  $E(G) < \theta$ .

Như vậy, việc sử dụng một thống kê mà kỳ vọng toán của nó khác với tham số cần ước lượng có thể dẫn đến sai số hệ thống (tất cả các giá trị của  $G$  đều lớn hơn hoặc nhỏ hơn  $\theta$ ). Để loại trừ sai số này hiển nhiên là cần yêu cầu  $E(G) = \theta$ . Dĩ nhiên yêu cầu trên không loại trừ được hoàn toàn các sai số, song như vậy các sai số khác dấu sẽ xuất hiện tương đối đều nhau, do đó các giá trị của  $G$  sẽ không bị lệch hẳn về một phía so với  $\theta$ .

**Định nghĩa 2.1.1.** Thống kê  $G$  của mẫu được gọi là ước lượng không chệch của tham số  $\theta$  của ĐLNN gốc  $X$  nếu  $E(G) = \theta$ .

Ngược lại  $E(G) \neq \theta$ . thì  $G$  được gọi là ước lượng chệch của  $\theta$ .

#### Chú ý 2.1.2.

(a)  $G$  là ước lượng không chệch của tham số  $\theta$  không có nghĩa là mọi giá trị của  $G$  đều trùng khít với  $\theta$  mà chỉ có nghĩa: trung bình các giá trị của thống kê  $G$  bằng  $\theta$ . Từng giá trị của  $G$  có thể sai lệch rất lớn so với  $\theta$ .

(b) Trung bình mẫu  $\bar{X}$  là ước lượng không chệch của kỳ vọng toán  $a$  của ĐLNN gốc [ $E(\bar{X}) = a$ ].

(c) Tần suất mẫu  $f$  là ước lượng không chệch của xác suất  $p$  của ĐLNN gốc [ $E(f) = p$ ].

(d) Phương sai điều chỉnh mẫu  $S'^2$  là ước lượng không chệch của phương sai  $\sigma^2$  của ĐLNN gốc [ $E(S'^2) = \sigma^2$ ].

**Ví dụ 2.1.3.** Giả sử trái cây của nông trường đã đóng thùng, mỗi thùng 10 trái. kiểm tra ngẫu nhiên 50 thùng ta thu được kết quả sau:

Số trái cây hỏng trong thùng $k$	0	1	2	3	4	5	6	7	8	9	10
Số thùng có $k$ trái hỏng	0	2	3	7	20	6	4	7	0	0	1

a/ Tìm ước lượng cho tỷ lệ trái cây hỏng trong nông trường.

b/ Tìm ước lượng cho tỷ lệ trái cây hỏng trung bình trong mỗi thùng.

c/ Tìm ước lượng không chệch cho độ biến động tỷ lệ trái cây hỏng ở mỗi thùng,

*Giải:*

a/ Đây là bài toán ước lượng điểm cho tỷ lệ đám đông.

Tổng số trái điều tra là:  $n = 10 \times 50 = 500$  ( trái)

Số trái hỏng phát hiện là:  $m = 1.2 + 2.3 + 3.7 + 4.20 + 5.6 + 6.4 + 7.7 + 10.1 = 22$ .

Tỷ lệ hỏng trong mẫu là:  $f = \frac{22}{500} = 0,444$ .

Vậy tỷ lệ trái hỏng trong nông trường là vào khoảng 44,4%.

b/ Ta quan điểm đây là bài toán về số lượng chứ không phải là chất lượng. Mỗi thùng có tỷ lệ hỏng  $X_i$  và ta cần ước lượng điểm tỷ lệ hỏng trung bình ở mỗi thùng.

Ta có bảng sau: (  $h = 10$ ;  $x_0 = 40$ ;  $t_i = \frac{x_i - 40}{10}$  )

$X_i$	$m_i$	$t_i$	$m_i t_i$	$m_i t_i^2$
10	2	- 3	- 6	18
20	3	- 2	- 6	12
30	7	- 1	- 7	7
40	20	0	0	0
50	6	1	6	6
60	4	2	8	16
70	7	3	21	63
100	1	6	6	36
$\Sigma$	50		22	158

Nhìn vào bảng trên ta thấy:

$$\bar{x} = 40 + \frac{10}{50} \cdot 22 = 44,4.$$

Vậy tỷ lệ hỏng trung bình ở mỗi thùng là khoảng 44,4%.

c/ Ước lượng không chệch cho độ biến động tỷ lệ trái cây hỏng ở mỗi thùng chính là phương sai điều chỉnh mẫu  $s^2$

$$s^2 = \frac{100}{50} \cdot 158 - (40 - 44,4)^2 = 296,64$$

$$s'^2 = \frac{50}{49} \cdot 296,64 = 302,69$$

Ta dự đoán độ biến động của tỷ lệ hỏng giữa các thùng là vào khoảng 302

**\* Ước lượng hiệu quả:**

Như đã phân tích ở trên, dù  $G$  là ước lượng không chệch của  $\theta$  thì từng giá trị cụ thể của  $G$  vẫn có thể sai lệch rất lớn so với  $\theta$ , tức là phương sai  $D(G)$  vẫn có thể rất lớn. Lúc đó, nếu lấy một giá trị của  $G$  tìm được trên một mẫu cụ thể, chẳng hạn  $g_1$  để ước lượng  $\theta$  thì nó có thể sai lệch rất nhiều so với giá trị trung bình  $\bar{G}$  tức là bản thân tham số  $\theta$  cần ước lượng. Như vậy, nếu lấy một giá trị của  $G$ , chẳng hạn  $g_1$  để ước lượng  $\theta$ , thì có thể mắc sai số rất lớn. Còn nếu như đòi hỏi phương sai của  $G$  phải nhỏ thì có thể loại trừ sai số này trong ước lượng.



**Định nghĩa 2.1.4.** Thống kê  $G$  của mẫu được gọi là ước lượng hiệu quả của tham số  $\theta$  của ĐLNN gốc  $X$  nếu nó có phương sai nhỏ nhất so với mọi thống kê khác được xây dựng trên cùng mẫu đó.

Như vậy, để xét xem  $G$  có phải là ước lượng hiệu quả của  $\theta$  hay không, cần phải tìm được giá trị nhỏ nhất có thể có của phương sai các hàm ước lượng.

**\* Ước lượng vững:**

Khi xét những mẫu có kích thước lớn thì nảy sinh vấn đề là mẫu càng lớn thì thống kê  $G$  của mẫu phải càng gần tham số  $\theta$  cần ước lượng.

**Định nghĩa 2.1.5.** Thống kê  $G$  của mẫu được gọi là ước lượng vững của tham số  $\theta$  của ĐLNN gốc  $X$  nếu  $G$  hội tụ theo xác suất đến  $\theta$  khi  $n \rightarrow \infty$ . Tức là với mọi  $\varepsilon$  dương bé tùy ý ta luôn có  $\lim_{n \rightarrow \infty} P(|G - \theta| < \varepsilon) = 1$

### 2.1.2. Phương pháp ước lượng hợp lý tối đa

Giả sử biết qui luật phân phối xác suất tổng quát của ĐLNN gốc  $X$  dưới dạng hàm mật độ  $f(x, \theta)$ . Đó cũng có thể là biểu thức xác suất nếu  $X$  là ĐLNN rời rạc, cần phải ước lượng tham số  $\theta$  nào đó của  $X$ . Lập mẫu ngẫu nhiên kích thước  $n$ :

$$W = (X_1, X_2, \dots, X_n).$$

và xây dựng hàm của đối số  $\theta$  tại một giá trị cụ thể của mẫu:

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

Hàm  $L$  được gọi là hợp lý của tham số  $\theta$ , giá trị của hàm hợp lý chính là xác suất hay mật độ xác suất tại điểm  $(x_1, x_2, \dots, x_n)$ , còn giá trị của thống kê  $G$  tại điểm đó  $g = f(x_1, x_2, \dots, x_n)$  được gọi là ước lượng hợp lý tối đa của  $\theta$  nếu ứng với giá trị này của  $\theta$ , hàm hợp lý đạt cực đại.

## 2.2. Phương pháp ước lượng bằng khoảng tin cậy

### 2.2.1. Khái niệm

Để ước lượng tham số  $\theta$  của ĐLNN gốc  $X$  trong tổng thể, phương pháp này chủ trương từ một thống kê  $G$  nào đó của mẫu xây dựng một khoảng giá trị  $(G_1, G_2)$  sao cho với một xác suất cho trước, tham số  $\theta$  sẽ rơi vào khoảng  $(G_1, G_2)$  đó. Chú ý rằng do  $G$  là ĐLNN nên khoảng  $(G_1, G_2)$  cũng là một khoảng ngẫu nhiên, còn  $\theta$  lại là một số xác định nên phải nói chính xác hơn là khoảng  $(G_1, G_2)$  sẽ chứa đựng giá trị  $\theta$  với một xác suất cho trước.

**Định nghĩa 2.2.1.** Khoảng  $(G_1, G_2)$  của thống kê  $G$  được gọi là *khoảng tin cậy* của tham số  $\theta$  nếu với xác suất bằng  $(1 - \alpha)$  cho trước thỏa mãn điều kiện  $P(G_1 < \theta < G_2) = (1 - \alpha)$ ; xác suất  $(1 - \alpha) = \gamma$  được gọi là độ tin cậy của ước lượng, còn  $I = G_2 - G_1$  được gọi là độ dài khoảng tin cậy.

**Phương pháp tiến hành:**

Từ tổng thể lập mẫu ngẫu nhiên kích thước  $n$ :  $W = (X_1, X_2, \dots, X_n)$  và từ đó xây dựng thống kê  $G = f(X_1, X_2, \dots, X_n, \theta)$  sao cho qui luật phân phối xác suất của  $G$  không phụ thuộc vào các đối số của nó và hoàn toàn xác định. Lúc đó với độ tin

cây bằng  $(1 - \alpha)$  cho trước có thể tìm được cặp giá trị  $\alpha_1$  và  $\alpha_2$  sao cho  $\alpha_1 + \alpha_2 = \alpha$  và tương đương với chúng tìm được cặp giá trị  $g_{\alpha_1}$  và  $g_{\alpha_2}$  (thường là các phân vị của G) thoả mãn điều kiện  $P(G < g_{\alpha_1}) = \alpha_1$  và  $P(G > g_{\alpha_2}) = \alpha_2$ . Từ đó suy ra  $P(g_{\alpha_1} < G < g_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = (1 - \alpha)$

Như vậy, với độ tin cậy  $(1 - \alpha)$  ta đã xây dựng được khoảng tin cậy  $(g_{\alpha_1}, g_{\alpha_2})$  cho G. Bằng các phép biến đổi tương đương bao giờ cũng có thể đưa công thức trên về dạng biểu thức tương đương  $P(G_1 < \theta < G_2) = 1 - \alpha$  đó chính là khoảng tin cậy cần tìm.

Khi tiến hành một phép thử với mẫu ngẫu nhiên  $W = (X_1, X_2, \dots, X_n)$  ta thu được một mẫu cụ thể  $w = (x_1, x_2, \dots, x_n)$ , do đó tính được giá trị của  $G_1$  và  $G_2$  ứng với mẫu cụ thể này là  $g_1$  và  $g_2$ , lúc đó có kết luận là: Qua mẫu cụ thể  $w$ , với độ tin cậy  $1 - \alpha$  tham số  $\theta$  của ĐLNN gốc X sẽ nằm trong khoảng  $(g_1, g_2)$  tức là  $(g_1 < \theta < g_2)$ .

**Nhận xét 2.2.2.** Với độ tin cậy  $1 - \alpha$  cho trước ta có thể tìm được vô số cặp  $\alpha_1, \alpha_2$  ( $0 < \alpha_1 < 1; 0 < \alpha_2 < 1$ ) thoả mãn  $\alpha_1 + \alpha_2 = \alpha$  vì thế có vô số khoảng tin cậy tương ứng với độ tin cậy đã cho.

### 2.2.2. Ước lượng kỳ vọng toán của ĐLNN phân phối theo qui luật chuẩn

Giả sử trong tổng thể ĐLNN gốc X phân phối chuẩn  $N(a, \sigma^2)$  nhưng chưa biết tham số  $a$  của nó. Để ước lượng  $a$  từ tổng thể lập một mẫu ngẫu nhiên kích thước  $n$ ,  $W = (X_1, X_2, \dots, X_n)$ . Để chọn thống kê G thích hợp ta xét các trường hợp sau:

#### a. Trường hợp đã biết phương sai $\sigma^2$ của ĐLNN gốc X

Ta chọn thống kê:  $G = U = \frac{(\bar{X} - a)\sqrt{n}}{\sigma}$ , (I)

Trong đó  $\bar{X}$  là trung bình mẫu. Ta thấy thống kê U phân phối chuẩn hoá  $N(0,1)$ . Do đó với độ tin cậy  $1 - \alpha$  cho trước tìm được cặp giá trị  $\alpha_1$  và  $\alpha_2$  sao cho  $\alpha_1 + \alpha_2 = \alpha$ . Từ đó tìm được hai phân vị chuẩn tương ứng là  $u_{\alpha_1}$  và  $u_{1-\alpha_2}$  thoả mãn điều kiện  $P(U < u_{\alpha_1}) = \alpha_1$  và  $P(U > u_{1-\alpha_2}) = \alpha_2$ .

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

Thay U từ biểu thức (I) vào biểu thức trên và giải ra theo ẩn  $a$  ta thu được

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

Vậy với độ tin cậy bằng  $(1 - \alpha)$  tham số  $a$  của ĐLNN gốc X sẽ nằm trong khoảng:  $\left(\bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha_2}; \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha_1}\right)$ , (1)

#### Các trường hợp đặc biệt:

+ **Khoảng tin cậy đối xứng:**  $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$  thì khoảng tin cậy của  $a$  là:

$$\left( \bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}; \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right), \quad (2)$$

Nếu ký hiệu  $\varepsilon = \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}$  thì biểu thức trên có dạng:  $(\bar{X} - \varepsilon; \bar{X} + \varepsilon)$ , (3)

$\varepsilon$  được gọi là độ chính xác (độ sai số) của ước lượng. Nó phản ánh mức độ sai lệch của trung bình mẫu so với trung bình tổng thể với xác suất  $(1 - \alpha)$  cho trước.

+ **Khoảng tin cậy bên phải:** Nếu  $\alpha_1 = 0, \alpha_2 = \alpha$  thì  $u_{1-\alpha_1} = +\infty$  do đó khoảng tin cậy bên phải của  $a$  là

$$\left( \bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}; +\infty \right), \quad (4)$$

Biểu thức trên dùng để ước lượng giá trị tối thiểu của  $a$ .

+ **Khoảng tin cậy bên trái:** Nếu  $\alpha_2 = 0, \alpha_1 = \alpha$  thì  $u_{1-\alpha_2} = +\infty$  do đó khoảng tin cậy của  $a$  là:

$$\left( -\infty; \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} \right), \quad (5)$$

Biểu thức trên dùng để ước lượng giá trị tối đa của  $a$ .

### Chú ý 2.2.3.

(a) Với cùng độ tin cậy  $(1 - \alpha)$  thì độ dài khoảng tin cậy  $I$  sẽ là ngắn nhất khi khoảng tin cậy là đối xứng và  $I = 2\varepsilon$ .

(b) Từ các công thức trên ta thấy ba số  $\varepsilon, \alpha, n$  luôn phụ thuộc vào nhau, nếu cho trước 2 số thì sẽ tìm được số còn lại, chẳng hạn cho trước  $\alpha$  và  $\varepsilon$  thì ta sẽ tìm được kích thước tối thiểu  $n$  của mẫu cần phải điều tra là:

$$n \geq \left[ \frac{\sigma^2 u_{1-\frac{\alpha}{2}}^2}{\varepsilon^2} \right] + 1 \Leftrightarrow n \geq \frac{4\sigma^2}{I^2} u_{1-\frac{\alpha}{2}}^2 + 1, \quad (6)$$

(c) Các giá trị  $u_{1-\frac{\alpha}{2}}$  và  $u_{1-\alpha}$  được tra ở bảng phụ lục 3

(d) Các khoảng tin cậy nói trên vẫn đang còn là khoảng tin cậy ngẫu nhiên đối với một mẫu ngẫu nhiên  $W = (X_1, X_2, \dots, X_n)$ . Thực hiện một phép thử đối với mẫu này thu được mẫu cụ thể  $w = (x_1, x_2, \dots, x_n)$ , từ đó tìm được giá trị cụ thể  $\bar{x}$  của trung bình mẫu. Lúc đó với độ tin cậy  $(1 - \alpha)$  qua một mẫu cụ thể ta sẽ tìm được một khoảng tin cậy cụ thể của  $a$ . Vậy với một mẫu ngẫu nhiên cho phép xác định khoảng tin cậy ngẫu nhiên, còn mẫu cụ thể cho phép tìm được khoảng tin cậy cụ thể (bằng số) của  $a$ .

**Ví dụ 2.2.4.** Để xác định trọng lượng trung bình của các bao gạo trong kho người ta lấy ngẫu nhiên ra 100 bao và tìm được trọng lượng trung bình  $\bar{x} = 36,06$  kg với  $\sigma^2 = (0,28)^2$ . Hãy tìm khoảng tin cậy đối xứng của trọng lượng trung bình của các

bao gạo trong kho với độ tin cậy  $(1 - \alpha) = 0,99$ . Nếu giả thiết trọng lượng các bao gạo tuân theo luật phân phối chuẩn.

*Giải:* Đây là bài toán tìm khoảng tin cậy đối xứng của giá trị trung bình khi đã biết phương sai. Vậy từ mẫu cụ thể khoảng tin cậy đối xứng của  $a$  là

$$\left( \bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} ; \bar{x} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right)$$

$$\text{với } u_{1-\frac{\alpha}{2}} = u_{0,995} = 2,576 \Rightarrow \varepsilon = \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} = \frac{0,28.2,576}{\sqrt{100}} \approx 0,072$$

$$\bar{x} - \varepsilon = 36,06 - 0,072 = 35,988$$

$$\bar{x} + \varepsilon = 36,06 + 0,072 = 36,132$$

Vậy với độ tin cậy 99% thì trọng lượng trung bình các bao gạo trong kho nằm trong khoảng  $(35,988 ; 36,132)$  kg

**Ví dụ 2.2.5.** Trọng lượng của một loại sản phẩm là một ĐLNN có phân phối chuẩn, với độ lệch tiêu chuẩn là 1. Cần phải điều tra một mẫu có kích thước là bao nhiêu để với độ tin cậy của ước lượng là 0,95 thì sai số cho phép không vượt quá 0,1.

*Giải:* Theo giả thiết ta có  $\varepsilon_0 = 0,1$ ;  $\sigma = 1$ ; với  $1 - \alpha = 0,95$  nên  $u_{1-\frac{\alpha}{2}} = 1,96$

$$\text{Theo công thức (3.5) ta được } n \geq \left[ \frac{1}{(0,1)^2} \cdot (1,96)^2 \right] + 1 = [384,16] + 1$$

Vậy để đáp ứng các yêu cầu của đầu bài ta phải điều tra một mẫu có kích thước tối thiểu  $n = 385$

## b. Trường hợp chưa biết phương sai của ĐLNN gốc X

\* **Kích thước mẫu nhỏ  $n \leq 30$ .**

$$\text{Ta chọn thống kê } G = T = \frac{\bar{X} - a}{S'} \sqrt{n}$$

Trong đó  $S'$  là độ lệch tiêu chuẩn điều chỉnh mẫu. Ta thấy thống kê  $T$  tuân theo qui luật Student với  $(n - 1)$  bậc tự do. Với độ tin cậy  $1 - \alpha$  cho trước có thể tìm được cặp giá trị  $\alpha_1$  và  $\alpha_2$  sao cho  $\alpha_1 + \alpha_2 = \alpha$ , từ đó tìm được hai phân vị Student tương ứng là  $t_{\alpha_1}^{(n-1)}$  và  $t_{1-\alpha_2}^{(n-1)}$  thoả mãn điều kiện:

$$P(T < t_{\alpha_1}^{(n-1)}) = \alpha_1 \quad \text{và} \quad P(T > t_{1-\alpha_2}^{(n-1)}) = \alpha_2$$

$$\text{Từ đó } P(t_{\alpha_1}^{(n-1)} < T < t_{1-\alpha_2}^{(n-1)}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha$$

$$\text{Vì } t_{\alpha_1}^{(n-1)} = -t_{1-\alpha_1}^{(n-1)} \text{ nên biểu thức trên có thể viết } P(-t_{1-\alpha_1}^{(n-1)} < T < t_{1-\alpha_2}^{(n-1)}) = 1 - \alpha$$

Hoàn toàn tương tự như trên khoảng tin cậy của  $a$  với độ tin cậy  $1 - \alpha$  là:

$$\left( \bar{X} - \frac{S'}{\sqrt{n}} t_{1-\alpha_2}^{(n-1)} ; \bar{X} + \frac{S'}{\sqrt{n}} t_{1-\alpha_1}^{(n-1)} \right), \quad (7)$$

**Các trường hợp đặc biệt:**

+ **Khoảng tin cậy đối xứng:**  $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$  thì khoảng tin cậy của a là:

$$\left( \bar{X} - \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{(n-1)} ; \bar{X} + \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{(n-1)} \right), \quad (8)$$

+ **Khoảng tin cậy bên phải:** Khi  $\alpha_1 = 0, \alpha_2 = \alpha$  thì khoảng tin cậy bên phải của a là:

$$\left( \bar{X} - \frac{S'}{\sqrt{n}} t_{1-\alpha}^{(n-1)} ; +\infty \right), \quad (9)$$

Biểu thức trên dùng để ước lượng giá trị tối thiểu của a.

+ **Khoảng tin cậy bên trái:** Khi  $\alpha_2 = 0, \alpha_1 = \alpha$  thì khoảng tin cậy bên trái của a là:

$$\left( -\infty ; \bar{X} + \frac{S'}{\sqrt{n}} t_{1-\alpha}^{(n-1)} \right), \quad (10)$$

Biểu thức trên dùng để ước lượng giá trị tối đa của a.

### Chú ý 2.2.6.

(a) Tra bảng phụ lục 5 (phân vị Student  $t_{\alpha}^{(n)}$ ) ta sẽ tìm được các giá trị  $t_{1-\frac{\alpha}{2}}^{(n-1)}$

hoặc  $t_{1-\alpha}^{(n-1)}$

(b) Nếu đặt  $\varepsilon = \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{(n-1)}$  thì (8)  $\Leftrightarrow (\bar{X} - \varepsilon ; \bar{X} + \varepsilon)$ , (8')

và độ dài khoảng tin cậy I sẽ là ngắn nhất khi khoảng tin cậy sẽ là đối xứng và

$$I = 2\varepsilon = 2 \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{(n-1)}$$

(c) Nếu ấn định độ tin cậy và sai số cho phép không vượt quá  $\varepsilon_0$  thì khi đó dung lượng mẫu cần thiết phải thoả mãn

$$n \geq \left[ \frac{S'^2}{\varepsilon_0^2} \left( t_{1-\frac{\alpha}{2}}^{(n-1)} \right)^2 \right] + 1, \quad (11)$$

(d) Với độ tin cậy khá lớn, để có khoảng tin cậy cụ thể của a, người ta lập mẫu cụ thể  $w = (x_1, x_2, \dots, x_n)$ , từ đó tính được các giá trị  $\bar{x}$  và  $s'$  cụ thể rồi thay chúng vào các công thức trên ta sẽ có khoảng tin cậy cụ thể bằng số phải tìm.

**Ví dụ 2.2.7.** Một giống lúa mới được gieo trong 10 miếng đất thí nghiệm có các điều kiện giống nhau, cho sản lượng tính theo cùng đơn vị như sau:

25,4; 28,0; 20,1; 27,4; 25,6; 23,9; 24,8; 26,4; 27,0; 25,4.

Hãy tìm khoảng tin cậy của sản lượng giống lúa trên với độ tin cậy 95%. Giả thiết sản lượng lúa là đại lượng ngẫu nhiên có phân phối chuẩn.

*Giải:* Gọi X là sản lượng của giống lúa trên, theo giả thiết X phân phối chuẩn  $N(a, \sigma^2)$  với a là tham số chưa biết cần ước lượng. Đây là bài toán tìm khoảng ước lượng đối xứng của sản lượng giống lúa khi chưa biết phương sai với mẫu nhỏ. (Sử dụng công thức 8')

Theo bài ra ta có:  $n=9$      $\bar{x} = 25,4$ ;  $s' = 2,238$      $\gamma = 1 - \alpha = 0,95$ . Tra bảng ta được  $t_{0,975}^9 = 2,262$ ,  $\varepsilon = \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{(n-1)} = 1,601$ ;

$$\bar{x} - \varepsilon = 25,4 - 1,601 = 23,799 ; \quad \bar{x} + \varepsilon = 25,4 + 1,601 = 27,001$$

Vậy với độ tin cậy 95% sản lượng trung bình của giống lúa mới trên nằm trong khoảng  $(23,799, 27,001)$  hay  $23,799 < a < 27,001$

\* **Kích thước mẫu lớn  $n \geq 30$** : Ta chọn thống kê  $G = T = \frac{(\bar{X} - a)\sqrt{n}}{S'}$

Người ta đã chứng minh được rằng khi  $n \rightarrow +\infty$  thì thống kê  $T$  sẽ phân phối xấp xỉ  $N(0,1)$  do đó với độ tin cậy cho trước và với  $n$  đủ lớn ( $n \geq 30$ ) ta sẽ xấp xỉ phân phối Student bằng phân phối chuẩn (Vẫn áp dụng các khoảng tin cậy như với mẫu nhỏ nhưng thay việc tra bảng phụ lục 5 bằng bảng phụ lục 3).

**Ví dụ 2.2.8.** Để ước lượng năng suất trung bình của giống lúa mới tại một vùng, người ta gặt ngẫu nhiên 100 thửa ruộng của vùng đó và thu được kết quả sau:

Năng suất X (tạ/ha)	40	42	44	46	48	50
	42	44	46	48	50	52
Số thửa $m_i$	7	13	25	35	15	5

Biết năng suất là ĐLNN có phân phối chuẩn. Hãy ước lượng năng suất trung bình của giống lúa mới ở vùng đó với độ tin cậy 95%.

*Giải:* Đây là bài toán tìm khoảng ước lượng đối xứng của trung bình đám đông khi chưa biết phương sai với mẫu có kích thước lớn. (áp dụng công thức (8))

Theo bài ra ta có  $\gamma = 0,95$  do đó  $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$ . Ta phải tìm  $\bar{x}$  và  $s'$ ,

Đặt  $x_i = \frac{x_i + x_{i+1}}{2}$ , chọn  $x_0 = 47$ ,  $h = 2$  đặt  $t_i = \frac{x_i - x_0}{h}$

STT	x	$m_i$	$t_i$	$m_i \cdot t_i$	$t_i^2$	$m_i \cdot t_i^2$
1	41	7	-3	-21	9	63
2	43	13	-2	-26	4	52
3	45	25	-1	-25	1	25
4	47	35	0	0	0	0
5	49	15	1	15	1	15
6	51	5	2	10	4	20
$\Sigma$		100		-47		175

Nhìn vào bảng trên ta thấy:

$$\bar{x} = x_0 + \frac{h}{n} \sum_{i=1}^6 m_i t_i = 47 + \frac{2}{100} \cdot (-47) = 46,06$$

$$s^2 = \frac{h^2}{n} \sum_{i=1}^6 m_i t_i^2 - (x_0 - \bar{x})^2 = \frac{4}{100} \cdot 175 - (47 - 46,06)^2 = 6,1164$$

$$s^2 = \frac{100}{99} \cdot 6,1164 = 6,178 \Rightarrow s' = 2,486$$

$$\varepsilon = \frac{S'}{\sqrt{n}} u_{1-\frac{\alpha}{2}} = 2,486 \cdot 1,96/10 = 0,487$$

$$\bar{x} - \varepsilon = 45,573; \bar{x} + \varepsilon = 46,547$$

Vậy với độ tin cậy 95% năng suất trung bình của giống lúa mới tại vùng đó nằm trong khoảng ( 45,573; 46,547 ) tạ/ha.

**Ví dụ 2.2.9.** Chiều dài (X) loại sản phẩm A do một máy tự động sản xuất ra là một ĐLNN tuân theo luật phân phối chuẩn với độ lệch tiêu chuẩn  $\sigma = 3$  cm. Để ước lượng chiều dài trung bình của loại sản phẩm nói trên với độ tin cậy 95% người ta tiến hành đo 36 sản phẩm.

a/ Tìm khoảng tin cậy đối xứng của chiều dài trung bình loại sản phẩm đó.

b/ Để ước lượng với độ tin cậy 99%, độ dài khoảng tin cậy đối xứng không vượt quá 0,6 cm thì phải đo bao nhiêu sản phẩm?

*Giải:* Gọi chiều dài trung bình mẫu là  $\bar{x}$ . Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng của trung bình tổng thể của ĐLNN tuân theo luật phân phối chuẩn khi đã biết phương sai. Theo giả thiết  $n = 36$ ;  $\sigma = 3$ ;  $\gamma = 0,95$  do đó

$$\varepsilon = \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} = \frac{3}{6} \cdot 1,96 = 0,98.$$

a/ Chiều dài trung bình của loại sản phẩm trên là

$$(\bar{x} - 0,98 ; \bar{x} + 0,98)$$

b/ Theo giả thiết  $I \leq 0,6$  cm;  $\varepsilon \leq 0,3$  cm;  $\gamma = 0,99 \Rightarrow u_{1-\frac{\alpha}{2}} = 2,576.$

$$\text{Áp dụng công thức 3.5 ta được } n \geq \left[ \frac{(2,576)^2 \cdot 3^2}{(0,3)^2} \right] + 1 = 664.$$

**Kết luận:** Để đáp ứng các yêu cầu của đầu bài thì cần phải điều tra một mẫu có kích thước tối thiểu là 664 sản phẩm.

**Ví dụ 2.2.10.** Năng suất giống ngô A ở một vùng được báo lên qua 25 điểm thu hoạch và có kết quả sau:

Năng suất (tạ/ha) X	7	9	11	13	17
Số điểm thu hoạch $m_i$	2	7	12	3	1

Biết năng suất ngô của vùng đó là ĐLNN có phân phối chuẩn.

a/ Tìm khoảng tin cậy đối xứng của năng suất trung bình của giống ngô A của vùng đó với độ tin cậy 95%.

b/ Hãy tính năng suất trung bình tối thiểu của giống ngô A của vùng đó với độ tin cậy 95%

*Giải:*

a/ Theo giả thiết X là ĐLNN phân phối chuẩn  $N(a, \sigma^2)$ . Vậy năng suất trung bình chính là giá trị a. Đây là bài toán ước lượng bằng khoảng tin cậy đối xứng giá trị

của tham số  $a$  của phân phối  $N(a, \sigma^2)$  khi chưa biết phương sai  $\sigma^2$  của  $X$  với mẫu nhỏ. Ta có khoảng tin cậy đối xứng của  $a$  là:

$$\left( \bar{X} - \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{(n-1)} ; \bar{X} + \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{(n-1)} \right)$$

Qua 25 điểm báo lên tức là ta có một mẫu ngẫu nhiên kích thước  $n = 25$ , gọi  $X_i$  là năng suất của điểm thứ  $i$  ( $i = 1..25$ ) ta có  $W = (X_1, X_2, \dots, X_{25})$  từ đó

$$\bar{X} = \frac{1}{25} \sum_{i=1}^{25} X_i. \text{ Với độ tin cậy 95\% tra bảng phụ lục 5 ta được } t_{0,975}^{24} = 2,064.$$

Từ mẫu cụ thể ta sẽ tìm được  $\bar{x}$  và  $s'$ . Ta thấy các  $x_i$  cách đều nhau một khoảng  $h = 2$ , chọn  $x_0 = 11$ , đặt  $t_i = \frac{x_i - x_0}{2}$

X	$m_i$	$t_i$	$m_i t_i$	$t_i^2$	$m_i t_i^2$
7	2	-2	-4	4	8
9	7	-1	-7	1	7
11	12	0	0	0	0
13	3	1	3	1	3
17	1	3	3	9	9
$\Sigma$	25		-5		27

Nhìn vào bảng trên ta thấy:

$$\bar{x} = 11 - 0,4 = 10,6; s^2 = 4,16 ; s' = 2,08 ; \varepsilon = 0,859.$$

Kết luận: Với độ tin cậy 95% năng suất trung bình của giống ngô A đó là:

$$(9,741 ; 11,459)$$

b/ Đây là bài toán tìm khoảng tin cậy bên phải của ước lượng  $a$  của  $X: N(a, \sigma^2)$  khi chưa biết phương sai với mẫu cụ thể  $n = 25$

$$\text{Khoảng tin cậy bên phải của } a \text{ là } \left( \bar{X} - \frac{S'}{\sqrt{n}} t_{1-\alpha}^{(n-1)} ; +\infty \right)$$

$$\text{Với độ tin cậy 95\% thì } t_{1-\alpha}^{24} = 1,711 \text{ do đó } \bar{x} - \frac{S'}{\sqrt{n}} t_{1-\alpha}^{n-1} = 9,888.$$

Kết luận: Với độ tin cậy 95% năng suất trung bình tối thiểu của giống ngô A của vùng đó là  $a \geq 9,888$  tạ/ha.

### Chú ý 2.2.11.

(a) Ta không được viết  $P(9,743 < a < 11,457) = 0,95$  vì độ tin cậy gắn với độ tin cậy ngẫu nhiên chứ không gắn với một mẫu cụ thể. Hơn nữa do  $a$  là một hằng số nên nó chỉ có thể thuộc hoặc không thuộc khoảng  $(9,743 ; 11,457)$  tức là với một mẫu cụ thể thì biến cố  $(9,743 < a < 11,457)$  không phải là biến cố ngẫu nhiên. Hoặc nó là biến cố chắc chắn, hoặc nó là biến cố không thể có.

(b) Trong bài toán trên ở ý a, nếu ta tăng độ tin cậy từ 95% lên 99% thì giá trị tra bảng  $t_{1-\frac{\alpha}{2}}^{n-1} = t_{0,995}^{24} = 2,797$  do đó  $\varepsilon = \frac{S'}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{n-1} = 1,164$ , khoảng tin cậy của  $a$  sẽ là  $(9,435 ; 11,764)$ . Vậy nếu tăng độ tin cậy lên mà giữ nguyên kích thước mẫu  $n$



thì giá trị của phân vị chuẩn cũng tăng theo do đó  $\varepsilon$  cũng tăng lên làm cho độ chính xác của ước lượng cũng giảm xuống.

Khi tăng kích thước mẫu  $n$  lên và giữ nguyên độ tin cậy cho trước thì  $\varepsilon$  giảm xuống tức là độ chính xác của ước lượng tăng lên.

### 2.2.3. Ước lượng của kỳ vọng toán của ĐLNN X không phân phối theo qui luật chuẩn:

Giả sử ở một tổng thể, dấu hiệu định lượng cần nghiên cứu nào đó được xem như ĐLNN X phân phối theo một qui luật nào đó khác qui luật chuẩn. X có kỳ vọng toán là  $a$  mà ta cần ước lượng. Để ước lượng  $a$  ta chọn thống kê:

$$G = U = \frac{\bar{X} - a}{\sigma} \sqrt{n} \quad (\text{khi đã biết phương sai}) \text{ hoặc}$$

$$G = T = \frac{\bar{X} - a}{S'} \sqrt{n} \quad (\text{khi chưa biết phương sai})$$

Người ta đã chứng minh khi kích thước của mẫu đủ lớn thì thống kê  $G$  được coi là có phân phối chuẩn hoá  $N(0,1)$ . Do vậy để ước lượng  $a$  ta cần phải chọn mẫu có kích thước lớn và khi đó ta sẽ đi ước lượng  $a$  giống như ước lượng  $a$  của X có phân phối chuẩn.

## 2.3. Ước lượng khoảng tin cậy cho tỷ lệ

Giả sử ở một tổng thể của ĐLNN X có  $N$  phần tử, trong đó có  $M$  phần tử mang đặc tính A và  $N - M$  phần tử không mang đặc tính A ( $0 < M < N$ ). Gọi  $P$  là tỷ lệ giữa số phần tử mang đặc tính A với toàn bộ số phần tử của tổng thể  $P = \frac{M}{N}$ ,  $P$  chính là xác suất để một phần tử trong tổng thể mang đặc tính A. Việc tính chính xác  $P$  là gặp nhiều khó khăn, do đó ta đi ước cho tỷ lệ  $P$  với độ tin cậy cho trước ( $P$  không quá lớn hoặc quá bé)

### Phương pháp tiến hành:

Từ tổng thể cần nghiên cứu ta lấy ra một mẫu kích thước  $n$

$W = (X_1, X_2, \dots, X_n)$ ,  $X_i$  là số phần tử mang đặc tính A ở lần thử thứ  $i$ . Các  $X_i$  ( $i = 1 \dots n$ ) là những ĐLNN độc lập với nhau và có phân phối 0 - 1 với tham số  $P$ .

$E(X_i) = P$  và  $D(X_i) = p(1 - p)$ . Gọi  $f = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \Rightarrow f$  là tỷ lệ mang đặc tính A

trong mẫu ngẫu nhiên được lấy ra.

Ta đã biết  $E(f) = P$  và  $D(f) = \frac{p(1-p)}{n}$ . Ta sử dụng thống kê  $U = \frac{f - p}{\sqrt{p(1-p)}} \sqrt{n}$ .

Ta thấy khi  $n$  đủ lớn ( $n > 30$ )  $f$  không quá bé và cũng không quá lớn thì  $U$  phân phối xấp xỉ  $N(0,1)$ . Khi đó với độ tin cậy cho trước ta có thể tìm được hai phân vị chuẩn  $u_{\frac{\alpha}{2}}$  và  $u_{1-\frac{\alpha}{2}}$  thoả mãn điều kiện:

$$P\left(u_{\frac{\alpha}{2}} < u < u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha. \text{ Do } u_{\frac{\alpha}{2}} = -u_{1-\frac{\alpha}{2}} \text{ nên } P\left(-u_{1-\frac{\alpha}{2}} < u < u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

hay  $P\left(-u_{1-\frac{\alpha}{2}} < \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} < u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$ . Giải hệ bất phương trình trên với ẩn  $p$

theo  $u_{1-\frac{\alpha}{2}}$  ta được khoảng tin cậy đối xứng của  $P$  với độ tin cậy  $1 - \alpha$  cho trước. Ta

thấy khi  $n$  đủ lớn thì thống kê  $U$  cũng có phân phối xấp xỉ  $N(0,1)$ . Như vậy với độ tin cậy  $1 - \alpha$  thì:

+ **Khoảng tin cậy đối xứng của  $P$  là**

$$\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}}; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right)$$

+ **Khoảng tin cậy bên phải của  $P$  là**

$$\left(f - \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{1-\alpha}; +\infty\right),$$

+ **Khoảng tin cậy bên trái của  $P$  là**

$$\left(-\infty; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{1-\alpha}\right)$$

**Chú ý 2.3.1.**

(a) Nếu đặt  $\varepsilon = \frac{\sqrt{f(1-f)}}{\sqrt{n}}u_{1-\frac{\alpha}{2}}$  thì độ dài khoảng tin cậy sẽ là ngắn nhất nếu nó

là khoảng tin cậy đối xứng và  $I = 2\varepsilon$ .  $\varepsilon$  được gọi là độ chính xác hay sai số của ước lượng. Từ đó ta có thể suy ra:

+ Kích thước của mẫu cần phải điều tra đảm bảo cho việc ước lượng  $P$  có độ tin cậy  $1 - \alpha$  và sai số cho phép không vượt quá  $\varepsilon_0$  là:

$$N_1 \geq \left[\frac{f(1-f)}{\varepsilon_0^2}u_{1-\frac{\alpha}{2}}^2\right] + 1 \text{ (khi đã có mẫu định hướng)}$$

hoặc 
$$N_2 \geq \left[\frac{1}{4\varepsilon_0^2}u_{1-\frac{\alpha}{2}}^2\right] + 1 \text{ (khi chưa có mẫu định hướng)}$$

(b) Để sử dụng các công thức trên có kết quả chính xác hơn ta chú ý:  $n$  phải lớn,  $f$  không quá bé hoặc quá lớn. Trong thực hành thường áp dụng với  $n \geq 100$ ;  $0,1 \leq f \leq 0,9$ ;  $nf \geq 10$ ;  $n(1-f) \geq 10$

**Ví dụ 2.3.2.** Kiểm tra ngẫu nhiên 400 sản phẩm do một máy sản xuất ra thấy có 20 sản phẩm là phế phẩm. Với độ tin cậy 95% hãy ước lượng tỷ lệ phế phẩm tối đa của máy đó.

*Giải:* Gọi P là tỷ lệ phế phẩm của máy đó. Như vậy P là cơ cấu của tập hợp sản phẩm do máy đó sản xuất theo dấu hiệu " phế phẩm". Đây là bài toán ước lượng tham số P của qui luật phân phối không - một A(P) bằng khoảng tin cậy bên trái.

Vậy khoảng tin cậy của P có dạng:

$$\left( -\infty; f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{1-\alpha} \right).$$

Qua mẫu cụ thể ta có  $f = \frac{20}{400} = 0,05$ . Với  $1 - \alpha = 0,95 \Rightarrow u_{1-\alpha} = 1,645$ .

Vậy với độ tin cậy 0,95, qua mẫu cụ thể này thì khoảng tin cậy của P là

$$\left( -\infty ; 0,05 + \frac{\sqrt{0,05 \cdot 0,95}}{\sqrt{400}} \cdot 1,645 \right) \text{ hay } P < 0,0679.$$

Kết luận: Với độ tin cậy 95% tỷ lệ phế phẩm tối đa của máy đó là 6,79%

**Ví dụ 2.3.3.** Để kiểm tra số cá trong một hồ, cơ quan quản lý đánh bắt 2000 con cá, đánh dấu rồi thả xuống hồ. Lần sau đánh bắt lại 400 con, được 80 con có dấu. Hãy ước lượng số cá trong hồ với độ tin cậy 95%

*Giải:* Gọi N là số cá có trong hồ (N phải nguyên, dương). Tỷ lệ cá bị đánh dấu là  $P = \frac{2000}{N}$ . Ta phải đi ước lượng P bằng khoảng tin cậy đối xứng. Từ mẫu cụ thể ta

có  $f = \frac{80}{400} = 0,2$  ( ta thấy:  $0,1 < 0,2 < 0,9$ ;  $nf > 10$ ;  $n(1 - f) > 10$ ), với độ tin cậy

95% tra bảng ta được  $u_{1-\frac{\alpha}{2}} = 1,96 \Rightarrow \varepsilon = \frac{\sqrt{0,2 \cdot 0,8}}{\sqrt{400}} \cdot 1,96 = 0,0392$ . Vậy khoảng tin

cậy đối xứng của P là:  $0,2 - 0,0392 < P < 0,2 + 0,0392$ .

$$\text{Hay } 0,1608 < \frac{2000}{N} < 0,2392 \Leftrightarrow 8362 < N < 12438$$

Kết luận: với độ tin cậy 95% số cá có trong hồ nằm trong khoảng từ 8362 đến 12438 con.

## BÀI TẬP CHƯƠNG 2

**1.** Trọng lượng của một loại trứng gà được cho bởi bảng số liệu sau:

X-Trọng lượng (g)	25-30	30-35	35-40	40-45	45-50
Số quả	15	17	40	18	10

Bằng khoảng tin cậy đối xứng hãy ước lượng trọng lượng trung bình của loại trứng gà này với độ tin cậy 95%. Cho biết trọng lượng trứng gà là đại lượng ngẫu nhiên có phân phối chuẩn.

**2.** Kích thước của một loại sản phẩm do một máy tự động sản xuất ra là một đại lượng ngẫu nhiên tuân theo qui luật phân phối chuẩn. Sau khi kiểm tra 25 sản phẩm cụ thể ta thu được bảng số liệu sau:

Kích thước (cm)	20-22	22-24	24-26	26-28	30-32
Số sản phẩm	3	7	10	3	2

Hãy ước lượng kích thước trung bình của loại sản phẩm đó bằng khoảng tin cậy đối xứng với độ tin cậy 95%.

**3.** Để ước lượng năng suất trung bình của một giống lúa mới tại một vùng, người ta gặt ngẫu nhiên trên 50 thửa ruộng của vùng đó và thu được kết quả (tạ/ha):

Năng suất	57	58	59	60	61	62	63	64	65	66	67	68	70
Số thửa	2	3	2	6	4	4	8	6	4	3	4	3	1

Biết năng suất lúa là đại lượng ngẫu nhiên có phân phối chuẩn. Hãy ước lượng năng suất trung bình của giống lúa mới ở vùng đó với độ tin cậy 95%.

**4.** Để xác định giá trung bình đối với một loại hàng hoá trên thị trường, người ta điều tra ngẫu nhiên tại 100 cửa hàng thu được số liệu sau:

Giá (đồng)	83	85	87	89	91	93	95	97	99	101
Số cửa hàng	5	8	13	14	30	11	8	6	4	1

Với độ tin cậy 95% hãy ước lượng giá trung bình của loại hàng đó tại thời điểm đang xét. Biết rằng giá hàng hoá là đại lượng ngẫu nhiên tuân theo qui luật phân phối chuẩn.

**5.** Với độ tin cậy 95%, hãy ước lượng lượng xăng hao phí trung bình cho một loại xe ô tô chạy từ A đến B nếu chạy thử 30 lần trên đoạn đường này người ta ghi nhận được lượng xăng hao phí như sau:

Lượng xăng hao phí (lít)	9,6-9,8	9,8-10,0	10,0-10,2	10,2-10,4	10,4-10,6
Số lần tương ứng	3	5	10	8	4

Biết lượng xăng hao phí là ĐLNN tuân theo qui luật chuẩn.

**6.** Cân thử 100 quả trứng ta có kết quả sau:

X (g)	150	160	165	170	180	185
Số quả	4	20	25	30	15	6

Tìm khoảng ước lượng cho khối lượng trung bình của trứng với độ tin cậy 95%. Biết rằng khối lượng trứng là ĐLNN tuân theo quy luật phân phối chuẩn.

**7.** Để định mức thời gian gia công một chi tiết máy, người ta theo dõi ngẫu nhiên quá trình gia công 25 chi tiết và thu được số liệu sau:

Thời gian gia công (phút)	15-17	17-19	19-21	21-23	23-25	25-27
Số chi tiết máy tương ứng	1	3	4	12	3	2

Bằng khoảng tin cậy đối xứng hãy ước lượng thời gian gia công trung bình một chi tiết máy với độ tin cậy 95%. Giả thiết thời gian gia công chi tiết là ĐLNN tuân theo qui luật phân phối chuẩn.

**8.** Đo chỉ số mỡ sữa của 100 con bò lai Hà - Ấn  $F_1$  ta được bảng số liệu sau:

Chỉ số mỡ sữa (X)	3,0-3,6	3,6-4,2	4,2-4,8	4,8-5,4	5,4-6,0	6,0-6,6	6,6-7,2
Số bò lai	2	8	30	35	15	7	3

Hãy ước lượng chỉ số mỡ sữa trung bình của giống bò lai trên với độ tin cậy 95%. Giả thiết chỉ số mỡ sữa là đại lượng ngẫu nhiên có phân phối chuẩn.

**9.** Đo áp lực  $X$  (tính bằng  $\text{kg/cm}^2$ ) của 18 thùng chứa ta được bảng kết quả sau:

X	19,6	19,5	19,9	20,0	19,8	20,5	21,0	18,5	19,7
Số thùng	1	2	2	4	2	3	2	1	1

Với độ tin cậy 99% hãy tìm khoảng ước lượng đối xứng của áp lực trung bình của thùng trên. Biết rằng áp lực là ĐLNN có phân phối chuẩn.

**10.** Để xác định giá trung bình đối với một loại hàng hoá trên thị trường, người ta điều tra ngẫu nhiên tại 100 cửa hàng và thu được số liệu sau:

Giá (đồng) $X$	81	85	87	89	91	93	95	97	99	101
Số cửa hàng ( $m_i$ )	3	10	13	15	30	12	7	6	3	1

Với độ tin cậy 95% hãy ước lượng giá trung bình của loại hàng đó tại thời điểm đang xét bằng khoảng tin cậy đối xứng. Biết rằng giá của hàng hoá là đại lượng ngẫu nhiên tuân theo qui luật phân phối chuẩn.

**11.** Để xác định chiều cao trung bình của các cây bạch đàn, người ta tiến hành đo ngẫu nhiên 35 cây và có bảng số liệu:

Chiều cao ( $X$ -mét)	6,5-7,0	7,0-7,5	7,5-8,0	8,0-8,5	8,5-9,0	9,0-9,5
Số cây	2	4	10	11	5	3

Với độ tin cậy 95% có thể nói chiều cao trung bình của các cây đàn nằm trong khoảng nào. Giả thiết chiều cao của cây bạch đàn là ĐLNN tuân theo qui luật phân phối chuẩn.

**12.** Có số liệu về trọng lượng của loại trứng gà như ở bảng dưới đây. Bằng khoảng tin cậy đối xứng hãy ước lượng trọng lượng trung bình của loại trứng gà này với độ tin cậy 0,95. Giả thiết trọng lượng trứng gà là ĐLNN tuân theo qui luật phân phối chuẩn.

Trọng lượng ( $X$ -gam)	25-30	30-35	35-40	40-45	45-50
Số quả	2	3	10	8	2

**13.** Điều tra doanh số hàng tháng của 100 hộ kinh doanh một loại hàng, có bảng số liệu:

Doanh số ( $X$ -triệu đồng)	11,5	11,6	11,7	11,8	11,9	12,0
Số hộ tương ứng	10	15	20	30	15	10

Bằng khoảng tin cậy đối xứng hãy ước lượng doanh số trung bình hàng tháng của các hộ kinh doanh mặt hàng này với độ tin cậy 95%. Giả thiết doanh số là ĐLNN tuân theo qui luật phân phối chuẩn.

**14.** Đo độ chịu lực ( $\text{kg/cm}^2$ ) của 200 mẫu bê tông người ta thu được kết quả trong bảng sau:

Độ chịu lực ( $X$ )	190-200	200-210	210-220	220-230	230-240	240-250
Số mẫu bê tông	10	26	56	64	30	14

Hãy ước lượng độ chịu lực trung bình của bê tông với độ tin cậy 0,95. Biết rằng độ chịu lực của bê tông là ĐLNN tuân theo quy luật chuẩn.

**15.** Lấy 50 con sợi để xác định độ bền trung bình, ta có số liệu sau:

Độ bền (X-kg/cm <sup>2</sup> )	0,6- 0,8	0,8- 1,0	1,0- 1,2	1,2- 1,4	1,4- 1,6	1,6- 1,8	1,8- 2,0	2,0- 2,2	2,2- 2,4
Số con sợi	1	2	7	10	11	9	6	3	1

Hãy ước lượng độ bền trung bình của loại sợi này bằng khoảng tin cậy đối xứng với hệ số tin cậy 0,95. Giả thiết độ bền của sợi là ĐLNN tuân theo qui luật phân phối chuẩn.

**16.** Điều tra 365 điểm trồng lúa của một huyện có bảng số liệu:

Năng suất (X- ta/ha)	25	30	33	34	35	36	37	39	40
Số điểm trồng lúa	6	13	38	74	106	85	30	10	3

Với độ tin cậy 95% có thể nói năng suất lúa trung bình của huyện nằm trong khoảng nào. Giả thiết năng suất lúa là ĐLNN tuân theo qui luật phân phối chuẩn.

**17.** Đo đường kính của 20 chi tiết do một máy tiện sản xuất, ta có số liệu (tính bằng mm)

X	24	24	24	250	251	252	253	256	257	258	260
	7	8	9								
m <sub>i</sub>	2	2	3	5	1	1	2	1	1	1	1

Giả thiết đường kính là ĐLNN có phân phối chuẩn

a/ Tìm khoảng ước lượng của độ dài trung bình của đường kính chi tiết với độ tin cậy 0,95.

b/ Các chi tiết có đường kính từ 249 đến 251 được coi là sản phẩm loại A. Hãy tìm khoảng ước lượng cho tỷ lệ loại A với độ tin cậy 0,95.

**18.** Tại một khu rừng nguyên sinh người ta đánh dấu 1000 con chim, sau đó thả chúng vào rừng. Một thời gian sau người ta bắt lại 200 con thấy có 40 có được đánh dấu. Với độ tin cậy 99% thử ước lượng số chim có trong khu rừng

## Chương 3

# Kiểm định giả thuyết thống kê

### 3.1. Khái niệm chung

#### 3.1.1. Giả thuyết thống kê

Trong nhiều lĩnh vực đời sống kinh tế - xã hội chúng ta hay nêu ra các nhận xét khác nhau về các đối tượng quan tâm. Những nhận xét như vậy thường được coi là các giả thuyết, chúng cũng có thể đúng và cũng có thể sai. Vấn đề xác định đúng sai của một giả thuyết sẽ được gọi là **kiểm định**.

Trong thống kê chúng ta xuất phát từ một mẫu  $x_1 \dots x_n$  được chọn từ một tổng thể chưa biết phân phối hoặc biết được dạng phân phối nhưng chưa biết được tham số  $\theta$ . Ta có thể phát biểu nhiều nhận xét khác nhau về các yếu tố chưa biết - đó là **giả thuyết thống kê**. Nếu tham số  $\theta$  chưa biết và giả thuyết  $\theta$  bằng giá trị cụ thể  $\theta_0$  được đưa ra, ta nói rằng có một giả thuyết đơn, nếu khác đi ta có giả thuyết phức.

Giả thuyết được đưa ra kiểm định được gọi là giả thuyết gốc và ký hiệu  $H_0$ ; nó thường là giả thuyết đơn (trong các bài toán kiểm định tham số). Các giả thuyết khác với giả thuyết gốc gọi là đối thuyết ký hiệu  $H_1$ . Việc kiểm định một giả thuyết là đúng hay sai dựa trên thông tin mẫu được gọi là **kiểm định thống kê**

**Định nghĩa 3.1.1.** Giả thuyết thống kê là giả thuyết về dạng phân phối xác suất của ĐLNN, về các tham số đặc trưng của ĐLNN hoặc về tính độc lập của các ĐLNN.

#### 3.1.2. Quy tắc kiểm định:

Nguyên tắc chung của kiểm định giả thuyết thống kê là dựa trên nguyên lý xác suất nhỏ: Một sự kiện có xác suất xuất hiện khác bé thì có thể coi rằng nó không xảy ra khi thực hiện một phép thử có liên quan đến sự kiện đó. Tuy nhiên trong thực tế, vấn đề phức tạp và tế nhị hơn nhiều.

##### a. Tiêu chuẩn kiểm định

Từ tổng thể rút ra mẫu, từ thông tin mẫu ta chọn một thống kê  $G = f(X_1, \dots, X_n)$  có thể phụ thuộc vào tham số đã biết trong  $H_0$ . Nếu  $H_0$  đúng thì qui luật của  $G$  phải hoàn toàn xác định. Một thống kê như vậy được gọi là tiêu chuẩn kiểm định.

**Quy tắc kiểm định.**

Nếu ta thành công trong việc chia miền xác định của tiêu chuẩn G thành 2 phần  $S_\alpha$  và  $\overline{S_\alpha}$ , trong đó  $S_\alpha$  là miền bác bỏ  $H_0$  còn  $\overline{S_\alpha}$  là miền chấp nhận  $H_0$ . Nếu G tính trên mẫu có giá trị thuộc miền  $S_\alpha$  ta bác bỏ  $H_0$  nếu ngược lại ta chấp nhận. Miền bác bỏ  $H_0$  được gọi là miền tới hạn của tiêu chuẩn G.

Nếu dùng qui tắc trên có thể mắc 2 sai lầm

**Sai lầm loại 1:** Giả thuyết  $H_0$  đúng mà ta lại bác bỏ

**Sai lầm loại 2:** Giả thuyết  $H_0$  sai mà ta lại chấp nhận

Do giả thiết G có phân phối xác định khi  $H_0$  đúng và nếu gọi  $\alpha$  là xác suất để xảy ra sai lầm loại 1 thì  $\alpha = P(G_{qs} \in S_\alpha / H_0 \text{ đúng})$  (1). Trong đó  $G_{qs}$  chính là giá trị của G trên mẫu cụ thể đang xét. Tương tự nếu gọi  $\beta$  là xác suất phạm sai lầm loại 2 thì  $\beta = P(G_{qs} \in \overline{S_\alpha} / H_0 \text{ sai})$  (2).

### Chú ý 3.1.2.

(a) Bác bỏ một giả thuyết chỉ có nghĩa là chấp nhận một giả thuyết khác chứ không có nghĩa giả thuyết bị bác bỏ là sai.

(b) Chấp nhận một giả thuyết chỉ có nghĩa là không chấp nhận các giả thuyết khác chứ không có nghĩa giả thuyết được chấp nhận là đúng.

(c) Ta mong muốn cả hai xác suất (1) và (2) càng bé càng tốt. Trong thực tế ta không thể đồng thời làm giảm cả 2 xác suất đó bởi vì cứ  $\alpha$  giảm thì  $\beta$  tăng và ngược lại. Thông thường do sai lầm loại 1 dễ kiểm soát và (1) dễ tính hơn nên người ta hay chọn trước  $\alpha$  luôn là ngưỡng để xác suất phạm sai lầm loại 1 luôn nhỏ hơn  $\alpha$  đủ bé đó. Các giá trị của  $\alpha$  có thể là 0,1; 0,05; 0,01.. phụ thuộc vào yêu cầu của thực tế và nhà nghiên cứu. Giá trị  $\alpha$  được gọi là mức ý nghĩa của qui tắc kiểm định

## 3.2. Kiểm định giả thuyết về giá trị trung bình

Giả sử từ tổng thể của ĐLNN X có phân phối chuẩn  $N(a, \sigma^2)$  rút ra mẫu ngẫu nhiên  $(X_1, X_2, \dots, X_n)$ . Căn cứ vào kinh nghiệm và đưa ra:

$$H_0 : a = a_0; \quad H_1 : a \neq a_0$$

$$\text{hoặc } H_0 : a = a_0; \quad H_1 : a > a_0$$

$$\text{hoặc } H_0 : a = a_0; \quad H_1 : a < a_0$$

### 3.2.1. Trường hợp đã biết phương sai

Chọn tiêu chuẩn kiểm định là thống kê  $U = \frac{\bar{X} - a_0}{\sigma} \sqrt{n}$

Nếu  $H_0$  đúng thì ta có  $U = \frac{\bar{X} - a_0}{\sigma} \sqrt{n} = \frac{\bar{X} - a}{\sigma} \sqrt{n}$  và thống kê U phân phối chuẩn

hoá  $N(0,1)$

Từ mẫu cụ thể ta sẽ tìm được trung bình mẫu  $\bar{x}$  và tính được giá trị quan sát của tiêu chuẩn kiểm định là  $U_{qs}$



Với mức ý nghĩa  $\alpha$  cho trước tra bảng ( bảng phụ lục 3 ) ta sẽ tìm được các giá trị  $u_{1-\frac{\alpha}{2}}$  hoặc  $u_{1-\alpha}$

Tìm miền bác bỏ S:

\* **Miền bác bỏ hai phía:** ( $H_1: a \neq a_0$ )

$$S = \left( -\infty; -u_{1-\frac{\alpha}{2}} \right) \cup \left( u_{1-\frac{\alpha}{2}}; +\infty \right)$$

Tức là nếu  $|u_{qs}| \leq u_{1-\frac{\alpha}{2}}$  thì chưa có cơ sở để bác bỏ  $H_0$

Nếu  $|u_{qs}| \geq u_{1-\frac{\alpha}{2}}$  thì bác bỏ  $H_0$

\* **Miền bác bỏ bên phải:**  $H_1: a > a_0$

$$S = (u_{1-\alpha}; +\infty)$$

\* **Miền bác bỏ bên trái:** ( $H_1: a < a_0$ )

$$S = (-\infty; -u_{1-\alpha})$$

**Ví dụ 3.2.1.** Trọng lượng sản phẩm (X) do nhà máy sản xuất ra là một ĐLNN có phân phối chuẩn với độ lệch tiêu chuẩn  $\sigma = 2$  kg và trọng lượng trung bình là 20kg. Nghi ngờ máy hoạt động không bình thường đã làm thay đổi trọng lượng trung bình của sản phẩm. Người ta cân thử 100 sản phẩm và thu được kết quả sau:

Trọng lượng SP(X)	19	20	21	22	23
Số SP tương ứng	10	60	20	5	5

Với mức ý nghĩa  $\alpha = 0,05$  hãy cho kết luận về điều nghi ngờ trên?

*Giải:* Đây là bài toán kiểm định giả thuyết thống kê về giá trị trung bình  $a$  của đám đông có phân phối chuẩn khi biết phương sai.

Chọn giả thuyết:  $H_0: a = 20$

Đối thuyết:  $H_1: a \neq 20$

Chọn tiêu chuẩn kiểm định  $u = \frac{\bar{X} - 20}{2} \sqrt{100}$

Trong đó  $\bar{X}$  là trung bình mẫu ngẫu nhiên kích thước  $n = 100$ . Từ mẫu cụ thể trên ta tìm được  $\bar{x} = 20,35$  do đó giá trị quan sát của tiêu chuẩn kiểm định là  $u_{qs} = 1,75$ . Với mức ý nghĩa  $\alpha = 0,05$  thì  $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$ .

Miền bác bỏ hai phía là:  $S = (-\infty; -1,96) \cup (1,96; +\infty)$ .

Ta thấy  $u_{qs} \notin S$  do đó chưa có cơ sở để bác bỏ  $H_0$  nghĩa là điều nghi ngờ trên là sai.

**Ví dụ 3.2.2.** Nếu máy móc hoạt động bình thường thì trọng lượng sản phẩm (X) là ĐLNN có phân phối chuẩn  $N(a, \sigma^2)$  với kỳ vọng toán  $a = 100$  gam. Độ lệch tiêu chuẩn  $\sigma = 2$  gam. Qua một thời gian sản xuất người ta nghi ngờ trọng lượng sản phẩm có xu hướng tăng lên, cân thử 100 sản phẩm thì trọng lượng trung bình sản

phẩm của chúng là 100,4 gam. với mức ý nghĩa  $\alpha = 0,05$  hãy kết luận về điều nghi ngờ trên?

*Giải:* Đây là bài toán kiểm định giả thuyết thống kê về giá trị trung bình của tham số  $a$  của phân phối  $N(a, \sigma^2)$  khi đã biết phương sai

Chọn giả thuyết:  $H_0: a = 100$

Đối thuyết:  $H_1: a > 100$

Chọn tiêu chuẩn kiểm định  $u = \frac{\bar{X} - 100}{2} \sqrt{100}$  trong đó  $\bar{X}$  là trung bình mẫu ngẫu nhiên. Từ mẫu cụ thể  $\bar{x} = 100,4$  ta được  $u_{qs} = 2$

với mức ý nghĩa  $\alpha = 0,05$  thì  $u_{1-\alpha} = 1,645$

Miền bác bỏ là  $S = (1,645 ; +\infty)$ . Ta thấy  $u_{qs} \in S$  Vậy điều nghi ngờ trên là đúng tức là trọng lượng sản phẩm đã có xu hướng tăng lên.

### 3.2.2. Trường hợp chưa biết phương sai

Chọn tiêu chuẩn kiểm định là thống kê  $T = \frac{\bar{X} - a_0}{S'} \sqrt{n}$

Nếu  $H_0$  đúng thì ta có  $T = \frac{\bar{X} - a_0}{S'} \sqrt{n} = \frac{\bar{X} - a}{S'} \sqrt{n}$

+ Nếu mẫu nhỏ thì thống kê  $T$  tuân theo qui luật Student với  $n-1$  bậc tự do.

+ Nếu mẫu lớn thì phân phối Student tiến tới phân phối  $N(0,1)$  nên ta có thể xấp xỉ phân phối Student bằng phân phối chuẩn.

Từ mẫu cụ thể ta tìm được  $T_{qs}$ .

Với mức ý nghĩa  $\alpha$  cho trước tra bảng phụ lục 5ta sẽ tìm được  $t_{\frac{1-\alpha}{2}}^{(n-1)}$  hoặc  $t_{1-\alpha}^{(n-1)}$

(đối với mẫu nhỏ) hoặc  $u_{\frac{1-\alpha}{2}}$  hay  $u_{1-\alpha}$  (đối với mẫu lớn). Từ đó sẽ tìm được miền

bác bỏ  $S$ , so sánh và rút ra kết luận (tương tự như khi đã biết phương sai)

**Ví dụ 3.2.3.** Mức hao phí xăng ( $X$ ) cho một loại xe ô tô chạy trên đoạn đường AB là ĐLNN có phân phối chuẩn với kỳ vọng toán là 50 lít. Do đường được tu sửa lại, người ta cho rằng mức hao phí xăng trung bình đã giảm xuống. Quan sát 30 chuyến xe chạy trên đoạn đường AB ta thu được bảng số liệu sau:

Mức hao phí X ( lít )	48,5	49,0	49,5	50,0	50,5
	49,0	49,5	50,0	50,5	51,0
Số chuyến	5	10	10	3	2

Với mức ý nghĩa  $\alpha = 0,05$ , hãy kết luận về ý kiến nêu trên?

*Giải:* Đây là bài toán kiểm định giả thiết thống kê về giá trị trung bình của ĐLNN có phân phối chuẩn khi chưa biết phương sai với mẫu nhỏ.

Chọn giả thuyết:  $H_0: a = 50$

Đối thuyết:  $H_1: a < 50$

Chọn tiêu chuẩn kiểm định  $T = \frac{\bar{X} - 50}{S'} \sqrt{30}$

Từ mẫu cụ thể ta tìm được:  $\bar{x} = 49,53$  và  $s' = 0,55$  thay các giá trị vừa tìm được vào thống kê trên ta được  $T_{qs} = -4,68$

Với mức ý nghĩa  $\alpha = 0,05$  tra bảng phụ lục 5 ta được  $t_{0,95}^{(29)} = 1,69$

Miền bác bỏ bên trái là:  $S = (-\infty ; -1,69)$ . Ta thấy  $T_{qs} \in S \Rightarrow$  bác bỏ  $H_0$  nghĩa là điều nghi ngờ trên là đúng (mức hao phí xăng đã giảm xuống).

**Chú ý 3.2.4.** Nếu tổng thể của ĐLNN  $X$  không tuân theo qui luật phân phối chuẩn thì ta có thể tiến hành chọn mẫu có kích thước lớn ( $n > 30$ ) khi đó ta có thể tiến hành kiểm định tương tự như tiến hành kiểm định đối với ĐLNN có phân phối chuẩn.

### 3.3. Kiểm định sự bằng nhau của hai kỳ vọng

Giả sử có hai tổng thể nghiên cứu, trong tổng thể thứ nhất ĐLNN gốc  $X_1$  phân phối chuẩn  $N(a_1, \sigma_1^2)$ , trong tổng thể thứ hai ĐLNN gốc  $X_2$  phân phối  $N(a_2, \sigma_2^2)$ . Nếu  $a_1$  và  $a_2$  chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau, người ta đưa ra giả thuyết thống kê:  $H_0: a_1 = a_2$  và đối thuyết:  $H_1: a_1 \neq a_2$  (hoặc  $H_1: a_1 > a_2$ , hoặc  $H_1: a_1 < a_2$ )

Để kiểm định giả thuyết trên, từ hai tổng thể nói trên rút ra hai mẫu ngẫu nhiên độc lập kích thước tương ứng là  $n_1$  và  $n_2$ :

$$W_1 = (X_{11}, X_{12}, \dots, X_{1n})$$

$$W_2 = (X_{21}, X_{22}, \dots, X_{2n})$$

Ta xét hai trường hợp sau:

**3.3.1. Trường hợp 1.** Biết phương sai  $\sigma_1^2$  và  $\sigma_2^2$  của các ĐLNN gốc trong tổng thể

Chọn tiêu chuẩn kiểm định:  $G = U = \frac{(\bar{X}_1 - \bar{X}_2) - (a_1 - a_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ .

Ta thấy thống kê  $U$  phân phối  $N(0,1)$ .

Nếu giả thuyết  $H_0$  đúng thì thống kê  $U$  có dạng:

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ và cũng phân phối } N(0,1)$$

Từ mẫu cụ thể ta sẽ tìm được các giá trị  $\bar{x}_1$ , và  $\bar{x}_2$  cụ thể thay vào ta sẽ tìm được giá trị quan sát là  $U_{qs}$ .

Với mức ý nghĩa cho trước tra bảng ta sẽ tìm được giá trị  $u_{1-\frac{\alpha}{2}}$  hoặc  $u_{1-\alpha}$  từ đó ta

tìm được miền bác bỏ  $S$

\* **Miền bác bỏ hai phía:**  $S = (-\infty ; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}} ; +\infty)$ .

\* **Miền bác bỏ bên phải:**  $S = (u_{1-\alpha} ; +\infty)$ .

\* **Miền bác bỏ bên trái:**  $S = (-\infty ; -u_{1-\alpha})$ .

So sánh  $U_{qs}$  với miền bác bỏ  $S$  rồi rút ra kết luận.

**3.3.2. Trường hợp 2.** Chưa biết phương sai  $\sigma_1^2$  và  $\sigma_2^2$  của các ĐLNN gốc trong tổng thể: Với kích thước mẫu  $n_1$  và  $n_2$  đủ lớn ta chọn tiêu chuẩn kiểm định là thống kê:

$$G = U = \frac{(\bar{X}_1 - \bar{X}_2) - (a_1 - a_2)}{\sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}}$$

Ta thấy  $U$  phân phối xấp xỉ  $N(0,1)$ . Nếu giả thuyết  $H_0$  đúng, tiêu chuẩn kiểm định

có dạng  $U = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}}$  và cũng phân phối  $N(0,1)$ , Từ mẫu cụ thể ta sẽ tìm được

các giá trị  $\bar{x}_1$ , và  $\bar{x}_2$ ,  $S_1'^2$ ,  $S_2'^2$  cụ thể thay vào thống kê trên ta sẽ tìm được  $U_{qs}$ , vì vậy với mức ý nghĩa cho trước tra bảng ta sẽ tìm được giá trị  $u_{1-\frac{\alpha}{2}}$  hoặc  $u_{1-\alpha}$  từ đó

ta tìm được miền bác bỏ  $S$ :

\* **Miền bác bỏ hai phía:**  $S = (-\infty ; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}} ; +\infty)$

\* **Miền bác bỏ bên phải:**  $S = (u_{1-\alpha} ; +\infty)$

\* **Miền bác bỏ bên trái:**  $S = (-\infty ; -u_{1-\alpha})$

So sánh  $U_{qs}$  với miền bác bỏ  $S$  rồi rút ra kết luận.

**Ví dụ 3.3.1.** Biết trọng lượng sản phẩm do hai máy sản xuất ra là ĐLNN có phân phối chuẩn với  $N(a_1,1)$  và  $N(a_2,1)$  với  $\alpha = 0,05$ . Có thể xem trọng lượng trung bình của sản phẩm do hai máy sản xuất ra là như nhau hay không? nếu kiểm tra ngẫu nhiên  $n_1 = 25$  sản phẩm do máy 1 sản xuất ra thu được  $\bar{x}_1 = 50$  kg và  $n_2 = 20$  sản phẩm do máy 2 sản xuất ra thu được  $\bar{x}_2 = 50,6$  kg.

*Giải:* Đây là bài toán kiểm định sự bằng nhau của 2 kỳ vọng của 2 ĐLNN có phân phối chuẩn khi đã biết phương sai.

Chọn giả thuyết  $H_0: a_1 = a_2$

Đôi thuyết  $H_1: a_1 \neq a_2$

Tiêu chuẩn kiểm định là:  $U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Từ mẫu cụ thể ta tìm được giá trị quan sát  $U_{qs} = \frac{50 - 50,6}{\sqrt{\frac{1}{25} + \frac{1}{20}}} = -2$

Với  $\alpha = 0,05$  tra bảng ta được  $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$

Miền bác bỏ hai phía là:  $S = (-\infty; -1,96) \cup (1,96; +\infty)$

Ta thấy  $U_{qs}$  thuộc miền bác bỏ. Vậy bác bỏ  $H_0$

Kết luận: Không thể coi trọng lượng sản phẩm do 2 máy sản xuất ra là như nhau được.

**Ví dụ 3.3.2.** Nếu áp dụng biện pháp kỹ thuật thứ nhất thì khi điều tra ngẫu nhiên  $n_1 = 100$  thửa ruộng trồng giống lúa A thu được năng suất trung bình  $\bar{x}_1 = 100$  tạ/ha và  $s'_1 = 10$  tạ/ha. Còn nếu áp dụng biện pháp kỹ thuật thứ 2 thì khi điều tra ngẫu nhiên  $n_2 = 50$  thửa ruộng thu được năng suất trung bình  $\bar{x}_2 = 95$  tạ/ha và  $s'_1 = 9$  tạ/ha. Hãy kết luận với mức ý nghĩa  $\alpha = 0,05$  nếu áp dụng biện pháp kỹ thuật thứ nhất thì năng suất giống lúa A cao hơn thực sự so với kết quả áp dụng biện pháp kỹ thuật thứ hai không? giả thiết năng suất lúa tuân theo qui luật chuẩn.

*Giải:* Gọi  $X_1, X_2$  tương ứng là năng suất của giống lúa A khi áp dụng biện pháp kỹ thuật thứ nhất và thứ hai,  $a_1, a_2$  là năng suất trung bình tương ứng khi áp dụng các biện pháp đó.

Đây là bài toán kiểm định sự bằng nhau của hai kỳ vọng của hai ĐLNN có phân phối chuẩn khi chưa biết phương sai với mẫu lớn.

Chọn giả thuyết  $H_0: a_1 = a_2$

Đối thuyết  $H_1: a_1 > a_2$

Nếu  $H_0$  đúng thì tiêu chuẩn kiểm định có dạng  $U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}}$

Từ mẫu cụ thể với các giá trị tương ứng tìm được thay vào thống kê trên ta tìm được giá trị quan sát  $U_{qs} = 3,09$

Với  $\alpha = 0,05$  tra bảng ta được  $u_{0,95} = 1,645$  ta tìm được miền bác bỏ bên phải là:  $S = (1,645, +\infty)$ ;  $U_{qs} \in S \Rightarrow$  bác bỏ  $H_0$ .

## 3.4. Kiểm định giả thuyết của xác suất

### 3.4.1. Trường hợp một tổng thể

Giả sử trong tổng thể của ĐLNN  $X$  có xác suất xuất hiện biến cố A là  $p$ , nếu chưa biết  $p$  song có cơ sở để giả thuyết rằng giá trị của nó bằng  $p_0$

Chọn giả thuyết  $H_0: p = p_0$

Đối thuyết  $H_1: p \neq p_0$  ( hoặc  $p > p_0$ , hoặc  $p < p_0$  )

Từ tổng thể lập mẫu ngẫu nhiên kích thước  $n$ , với  $n$  đủ lớn ta chọn tiêu chuẩn kiểm định là thống kê  $G = U = \frac{f - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n}$

Nếu  $H_0$  đúng thì  $U = \frac{f - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} = \frac{f - p}{\sqrt{p(1-p)}} \sqrt{n}$  phân phối xấp xỉ  $N(0,1)$

Từ mẫu cụ thể ta tính được giá trị quan sát  $U_{qs}$  với mức ý nghĩa cho trước tra bảng ta sẽ tìm được giá trị  $u_{1-\frac{\alpha}{2}}$  hoặc  $u_{1-\alpha}$  từ đó ta tìm được miền bác bỏ  $S$

\* **Miền bác bỏ hai phía:**  $S = (-\infty; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$

\* **Miền bác bỏ bên phải:**  $S = (u_{1-\alpha}; +\infty)$

\* **Miền bác bỏ bên trái:**  $S = (-\infty; -u_{1-\alpha})$

So sánh  $U_{qs}$  với miền bác bỏ  $S$  rồi rút ra kết luận.

### 3.4.2. Trường hợp hai tổng thể

Giả sử trong tổng thể thứ nhất, ĐLNN gốc  $X_1$  phân phối  $A(p_1)$ , ĐLNN gốc  $X_2$  phân phối  $A(p_2)$ .

Nếu  $p_1$  và  $p_2$  chưa biết song có cơ sở để giả thuyết rằng giá trị của chúng bằng nhau ta đưa ra giả thuyết thống kê

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2, \text{ (hoặc } p_1 < p_2; \text{ hoặc } p_1 > p_2)$$

Để kiểm định giả thuyết trên, từ tổng thể nói trên rút ra hai mẫu ngẫu nhiên độc lập kích thước tương ứng là  $n_1$  và  $n_2$ :

$$W_1 = (X_{11}, X_{12}, \dots, X_{1n_1}) \text{ và } W_2 = (X_{21}, X_{22}, \dots, X_{2n_2})$$

Chọn tiêu chuẩn kiểm định là  $G = U = \frac{(f_1 - f_2) - (P_1 - P_2)}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ , trong đó  $f_1$  và  $f_2$  là các

tần suất mẫu tương ứng và  $f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$

Với  $n_1$  và  $n_2$  khá lớn thì thống kê  $U$  nói trên sẽ phân phối xấp xỉ  $N(0,1)$

Nếu giả thuyết  $H_0$  đúng thì tiêu chuẩn kiểm định có dạng:

$$U = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ và vẫn phân phối xấp xỉ } N(0,1).$$

Từ mẫu cụ thể ta tính được các giá trị cụ thể  $f_1$ ,  $f_2$  và  $f$  và giá trị quan sát  $U_{qs}$  với mức ý nghĩa cho trước tra bảng ta sẽ tìm được giá trị  $u_{1-\alpha/2}$  hoặc  $u_{1-\alpha}$  từ đó ta tìm được miền bác bỏ.

\* **Miền bác bỏ hai phía:**  $S = (-\infty; -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}; +\infty)$

\* **Miền bác bỏ bên phải:**  $S = (u_{1-\alpha}; +\infty)$

\* **Miền bác bỏ bên trái:**  $S = (-\infty ; -u_{1-\alpha})$

So sánh  $U_{qs}$  với miền bác bỏ  $S$  rồi rút ra kết luận.

**Ví dụ 3.4.1.** Tỷ lệ phế phẩm do một máy tự động sản xuất ra là 5%. Kiểm tra ngẫu nhiên 300 sản phẩm thấy có 24 sản phẩm là phế phẩm. Từ đó có ý kiến cho rằng tỷ lệ phế phẩm do máy đó sản xuất có chiều hướng tăng lên. Hãy kết luận ý kiến nêu trên với mức ý nghĩa  $\alpha = 0,05$ .

*Giải:* Gọi  $p$  là tỷ lệ sản phẩm là phế phẩm. Đây là bài toán kiểm định giả thuyết thống kê về tỷ lệ

Chọn giả thuyết thống kê  $H_0: p = 0,05$

Đối thuyết  $H_1: p > 0,05$

Chọn thống kê  $U = \frac{f - 0,05}{\sqrt{0,05 \cdot 0,95}} \sqrt{300}$ , trong đó  $f$  là tần suất của mẫu ngẫu

nhiên  $n = 300$

Với mức ý nghĩa  $\alpha = 0,05$ , tra bảng ta được  $u_{0,95} = 1,645$

Vậy miền bác bỏ bên phải là:  $S = (1,645 ; +\infty)$

Từ mẫu cụ thể ta tìm được  $f = 0,08$  và  $U_{qs} = 2,38$

Ta thấy  $U_{qs} \in S \Rightarrow$  bác bỏ  $H_0$  nghĩa là tỷ lệ phế phẩm do máy đó sản xuất ra đã tăng lên thật sự.

**Ví dụ 3.4.2.** Thống kê số tai nạn lao động tại hai xí nghiệp có các số liệu sau:

Xí nghiệp	Số công nhân	Số tai nạn lao động
I	200	20
II	800	120

Với mức ý nghĩa  $\alpha = 0,05$  hãy kết luận xem chất lượng của công tác bảo hộ lao động tại hai xí nghiệp trên có khác nhau không?

*Giải:* Gọi  $p_1, p_2$  tương ứng là tỷ lệ bị tai nạn lao động của hai xí nghiệp I và II

Đây là bài toán kiểm định giả thuyết thống kê về tỷ lệ của hai tổng thể

Chọn giả thuyết  $H_0: p_1 = p_2$

Đối thuyết  $H_1: p_1 \neq p_2$

Chọn tiêu chuẩn kiểm định  $U = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Từ mẫu cụ thể ta tìm được  $f_1 = 0,1; f_2 = 0,15; f = 0,14; U_{qs} = -1,85$

với mức ý nghĩa  $\alpha = 0,05$  tra bảng ta được  $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$

Vậy miền bác bỏ hai phía là  $(-\infty ; -1,96) \cup (1,96 ; +\infty)$

Ta thấy  $U_{qs} \notin S$

Kết luận: chưa có cơ sở để bác bỏ  $H_0$  nghĩa là công tác bảo hộ lao động tại hai xí nghiệp là như nhau.

### BÀI TẬP CHƯƠNG 3

**1.** Hàm lượng đường trung bình của một loại trái cây lúc đầu là 5%. Người ta chăm bón bằng một loại phân N và sau một thời gian kiểm tra một số trái cây được kết quả sau:

Hàm lượng X(%)	1-5	5-9	9-13	13-17	17-21	21-25	25-29	29-33	37-41
Số trái	51	47	39	36	32	8	7	3	2

Hãy cho kết luận về loại phân N trên với mức ý nghĩa 5%. Giả thiết hàm lượng đường của loại trái cây trên là ĐLNN tuân theo quy luật chuẩn.

**2.** Đo chỉ số mỡ sữa của 130 con bò lai Hà - Ấn  $F_1$  ta được bảng số liệu sau:

Chỉ số mỡ sữa (X)	3,0	3,6	4,2	4,8	5,4	6,0	6,6
	3,6	4,2	4,8	5,4	6,0	6,6	7,2
Số bò lai	2	8	35	43	22	15	5

Biết rằng chỉ số mỡ sữa trung bình của giống bò lai thuần chủng là 4,95. Với mức ý nghĩa 1% hãy cho kết luận về hiệu quả của việc lai giống.

**3.** Định mức thời gian hoàn thành một sản phẩm là 14 phút. Có cần thay đổi định mức không, nếu theo dõi thời gian hoàn thành một sản phẩm ở 25 công nhân ta thu được bảng số liệu sau:

Thời gian để SX 1 sản phẩm (phút)	10-12	12-14	14-16	16-18	20-22
Số công nhân tương ứng	3	6	10	4	2

Yêu cầu kết luận với mức ý nghĩa  $\alpha = 0,05$  biết rằng thời gian hoàn thành một sản phẩm là đại lượng ngẫu nhiên có phân phối chuẩn.

**4.** Định mức cũ để sản xuất một sản phẩm là 20 phút. Nay do cải tiến kỹ thuật, người ta sản xuất thử 100 sản phẩm và thu được số liệu:

Thời gian sản xuất 1 sản phẩm (X - phút)	16-17	17-18	18-19	19-20	20-21	21-22
Số sản phẩm tương ứng	6	10	24	30	18	12

Với mức ý nghĩa  $\alpha = 0,05$  có thể nói rằng việc cải tiến kỹ thuật giảm bớt thời gian sản xuất một sản phẩm hay không? Biết rằng thời gian sản xuất một sản phẩm là đại lượng ngẫu nhiên tuân theo qui luật phân phối chuẩn.

**5.** Mức hao phí xăng (X) cho một loại xe ô tô trên đoạn đường AB là đại lượng ngẫu nhiên phân phối chuẩn với kỳ vọng là 50 lít. Do đoạn đường được tu sửa lại, người ta cho rằng mức hao phí xăng trung bình đã giảm xuống. Quan sát 100 chuyến xe chạy trên đoạn đường AB thu được bảng số liệu:

Mức xăng hao phí (lít)	48,5-49,0	49,0-49,5	49,5-50,0	50,0-50,5	50,5-51,0
Số chuyến xe	15	17	40	18	10

Với mức ý nghĩa  $\alpha = 0,05$ , hãy kết luận về ý kiến nêu trên.

**6.** Kiểm tra các gói đường loại 1kg trong một siêu thị ta có kết quả:

Khối lượng (X-kg)	0,95	0,96	0,97	0,99	1,00	1,01	1,03	1,05
Số gói	19	30	32	8	2	3	5	1



Với mức ý nghĩa  $\alpha = 0,05$  có thể kết luận việc đóng gói đảm bảo yêu cầu hay không. Biết rằng khối lượng các gói đường là đại lượng ngẫu nhiên tuân theo qui luật phân phối chuẩn.

**7.** Sản phẩm của một xí nghiệp đúc cho phép số khuyết tật trung bình của một sản phẩm là 3. Sau khi đổi mới thiết bị, kiểm tra ngẫu nhiên 36 sản phẩm kết quả thu được:

Số khuyết tật trên 1 sản phẩm	0	1	2	3	4	5	6
Số sản phẩm	7	4	4	6	8	6	1

a/ Với độ tin cậy 95% hãy ước lượng số khuyết tật trung bình của mỗi sản phẩm sau khi đổi mới thiết bị.

b/ Kết luận về hiệu quả việc đổi mới thiết bị với mức ý nghĩa 5%

**8.** Kiểm tra chất lượng hai lô sản phẩm từ 2 cơ sở chuyên đến ta thấy: Trong 120 sản phẩm ở lô I có 70 sản phẩm loại A. Còn trong 150 sản phẩm ở lô II có 98 sản phẩm loại A. Hỏi với mức ý nghĩa 1% có thể coi hai nguồn hàng có cùng tỉ lệ hàng loại A hay không?

**9.** Điều tra về số người mắc bệnh bứu cổ ở một tỉnh phía Bắc thấy có 107 người bị bệnh trong 380 người đến khám. Trong khi ở một tỉnh miền Trung có 90 người trong số 310 người khám bệnh. Có thể kết luận về tỉ lệ mắc bệnh ở hai tỉnh trên là như nhau không, với mức ý nghĩa 5%.

## Chương 4

# Tương quan và hồi quy

Khi nghiên cứu sự phụ thuộc giữa hai đại lượng mà mỗi đại lượng chịu một sự phân tán ngẫu nhiên ( sự tản mát kiểm tra được ) ta dùng phương pháp phân tích tương quan. Phân tích tương quan không những phát hiện được mối quan hệ phụ thuộc giữa chúng mà còn "lượng hoá" được mối quan hệ này.

### 4.1. Hệ số tương quan

#### 4.1.1. Phân tích hệ số tương quan

**Định nghĩa 4.1.1.** Giả sử  $X$  và  $Y$  là hai đại lượng ngẫu nhiên có  $D(X) > 0$  và  $D(Y) > 0$ . Hệ số tương quan của hai đại lượng ngẫu nhiên  $X$  và  $Y$ , ký hiệu  $\rho(X, Y)$  được xác định như sau: 
$$\rho(X, Y) = \frac{E[X - E(X)][Y - E(Y)]}{\sqrt{D(X)}\sqrt{D(Y)}}$$

#### 4.1.2. Tính chất của hệ số tương quan

- (a)  $\rho(X, Y) = \rho(Y, X)$
- (b)  $\rho(X, Y) = 0$  nếu  $X$  và  $Y$  độc lập với nhau
- (c)  $|\rho(X, Y)| \leq 1$

#### Nhận xét 4.1.2.

- Ta có thể dùng  $\rho(X, Y)$  để đo mức độ phụ thuộc tuyến tính giữa hai ĐLNN.  $|\rho|$  càng lớn thì sự phụ thuộc tuyến tính càng rõ. Đặc biệt nếu  $\rho = \pm 1$  thì sự phụ thuộc tuyến tính xảy ra với xác suất bằng 1

- Nếu  $|\rho|$  càng nhỏ thì mức độ phụ thuộc tuyến tính giữa  $X$  và  $Y$  càng ít. Đặc biệt khi  $\rho = 0$  thì giữa  $X$  và  $Y$  không có quan hệ phụ thuộc tuyến tính. Trong trường hợp này ta nói  $X$  và  $Y$  không tương quan với nhau ( chú ý: hai ĐLNN độc lập với nhau thì không tương quan, nhưng điều ngược lại chưa chắc đã đúng. Riêng đối với những ĐLNN phân phối chuẩn thì tính không tương quan và tính độc lập là tương đương nhau.)

#### 4.1.3. Hệ số tương quan mẫu

Giả sử tiến hành n quan sát độc lập đối với cặp ĐLNN (X,Y) ta có mẫu ngẫu nhiên kích thước n sau đây:  $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$

Hệ số tương quan mẫu của X và Y, ký hiệu  $r(X, Y) = r$  được xác định bằng công thức sau đây:

$$r = \frac{\frac{1}{n} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n} \sum_i (X_i - \bar{X})^2} \sqrt{\frac{1}{n} \sum_i (Y_i - \bar{Y})^2}} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{S_X \cdot S_Y}$$

Nếu trong mẫu cụ thể các  $x_i$  và  $y_i$  cách đều nhau một khoảng  $h_x$  và  $h_y$  thì ta dùng phép đổi biến với  $x_0$  và  $y_0$  được chọn thích hợp tùy theo bảng số liệu

đặt:  $u_i = \frac{x_i - x_0}{h_x}$  và  $v_i = \frac{y_i - y_0}{h_y}$ , khi đó

$$r = \frac{\sum_i u_i v_i - n \bar{u} \bar{v}}{\sqrt{\sum_i u_i^2 - n(\bar{u})^2} \sqrt{\sum_i v_i^2 - n(\bar{v})^2}} = \frac{\bar{uv} - \bar{u} \bar{v}}{\sqrt{u^2 - (\bar{u})^2} \sqrt{v^2 - (\bar{v})^2}}$$

**Chú ý 4.1.3.** Sau khi đổi biến khi muốn trở về biến cũ thì:

$$\bar{x} = h_x \bar{u} + x_0; s_x = h_x s_u$$

$$\bar{y} = h_y \bar{v} + y_0; s_y = h_y s_v$$

**Nhận xét 4.1.4.**

(a)  $|r| \leq 1$

$r > 0$  thì X và Y là tương quan thuận

$r < 0$  thì X và Y là tương quan nghịch

(b)  $0,7 \leq r \leq 1$  thì X và Y là tương quan thuận và mạnh

$-1 \leq r \leq -0,7$  thì X và Y là tương quan nghịch và mạnh

$|r| \leq 0,3$  thì X và Y là tương quan yếu.

## 4.2. Hồi quy tuyến tính đơn giản

### 4.2.1. Phương trình hồi quy tuyến tính đơn giản của tổng thể

Khi có sự phụ thuộc tuyến tính giữa hai biến ngẫu nhiên tương đối chặt chẽ ta có thể hy vọng xấp xỉ biến này bởi một hàm tuyến tính của biến kia. Nghĩa là cần tìm biểu thức  $aX + b$  sao cho xấp xỉ Y tốt nhất theo nghĩa cực tiểu sai số bình phương trung bình  $E(Y - aX - b)^2$ .

Ta có:  $E(Y - aX - b)^2 = E\{(Y - E(Y)) - a(X - E(X)) + E(Y) - aE(X) - b\}^2$

$$= E(Y - E(Y))^2 + a^2 E(X - E(X))^2 + (E(Y) - aE(X) - b)^2 - 2aE\{(Y - E(Y))(X - E(X))\}$$

Vế phải sẽ đạt cực tiểu nếu và chỉ nếu tam thức bậc hai theo a:  $a^2 D(X) - 2ap\sqrt{DX}\sqrt{DY} + DY$  đạt cực tiểu và số hạng  $(EY - aEX - b)^2 = 0$ . Do đó ta chọn  $b = E(Y) - aE(X)$  còn a là tọa độ đỉnh của tam thức bậc hai:

$$a = -\frac{-2\rho\sqrt{D(X)}\sqrt{D(Y)}}{2D(X)} = \rho\sqrt{\frac{D(Y)}{D(X)}}$$

Khi đó giá trị nhỏ nhất của vế phải chính là giá trị của tam thức bậc hai theo a tại đỉnh của nó:

$$\min E(Y - aX - b)^2 = \rho^2 \frac{DY}{DX} DX - 2\rho\sqrt{\frac{DY}{DX}}\rho\sqrt{DX}\sqrt{DY} + DY = DY(1 - \rho^2)$$

Vậy biểu thức aX + b cần tìm chính là  $\rho\sqrt{\frac{DY}{DX}}X + EY - \rho\sqrt{\frac{DY}{DX}}EX$

Phương trình đường hồi qui bình phương trung bình tuyến tính của Y theo X là:

$$Y = \rho\sqrt{\frac{DY}{DX}}X + EY - \rho\sqrt{\frac{DY}{DX}}EX \Leftrightarrow Y = \rho\sqrt{\frac{DY}{DX}}(X - EX) + EY \quad (1)$$

Sai số bình phương trung bình khi dùng đường hồi qui trung bình tuyến tính để xấp xỉ Y là:  $\sigma_{\frac{y}{x}}^2 = DY(1 - \rho^2)$  (2)

Sai số này càng nhỏ khi  $|\rho|$  càng gần 1 tức là mức độ phụ thuộc tuyến tính giữa hai biến càng chặt.

Tương tự phương trình đường hồi qui bình phương trung bình tuyến tính của X theo Y là:

$$X = \rho\sqrt{\frac{DX}{DY}}Y + EX - \rho\sqrt{\frac{DX}{DY}}EY \Leftrightarrow X = \rho\sqrt{\frac{DX}{DY}}(Y - EY) + EX \quad (3)$$

Sai số bình phương trung bình khi dùng đường hồi qui trung bình tuyến tính để xấp xỉ X là:  $\sigma_{\frac{x}{y}}^2 = DX(1 - \rho^2)$  (4)

**Chú ý 4.2.1.** Khi có mẫu ngẫu nhiên  $(x_i, y_i); i = 1..n$  ta xây dựng đường hồi qui trung bình tuyến tính thực nghiệm bằng cách thay trong (1) và (3)  $EY$  bởi  $\bar{Y}$ ;  $EX$  bởi  $\bar{X}$  và  $\rho\sqrt{\frac{DY}{DX}}$  bởi  $r\frac{s_y}{s_x}$ ;  $\rho\sqrt{\frac{DX}{DY}}$  bởi  $r\frac{s_x}{s_y}$ .

#### 4.2.2. Phương trình đường hồi qui bình phương trung bình tuyến tính thực mẫu

- Phương trình đường hồi qui bình phương trung bình tuyến tính thực nghiệm của Y theo X là

$$y = r\frac{s_y}{s_x}(x - \bar{x}) + \bar{y}$$

- Phương trình đường hồi qui bình phương trung bình tuyến tính thực nghiệm của X theo Y là

$$x = r\frac{s_x}{s_y}(y - \bar{y}) + \bar{x}$$

\* Ước lượng sai số bình phương trung bình

$$s_{y/x}^2 = s_y^2(1 - r^2); \quad s_{x/y}^2 = s_x^2(1 - r^2)$$

**Ví dụ 4.2.1.** Tính hệ số tương quan và phương trình đường hồi qui trung bình tuyến tính thực nghiệm của Y đối với X dựa vào số liệu cho trong bảng tương quan sau:

$x_i$	1	2	2	3	3	4
$y_i$	3	4	5	5	6	7
$n_i$	3	2	1	2	1	1

*Giải:*

Ta thiết lập lại bảng số liệu:

$x_i$	$y_i$	$n_i$	$n_i x_i$	$n_i y_i$	$n_i x_i^2$	$n_i y_i^2$	$n_i x_i y_i$
1	3	3	3	9	3	27	9
2	4	2	4	8	8	32	16
2	5	1	2	5	4	25	10
3	5	2	6	10	18	50	30
3	6	1	3	6	9	36	18
4	7	1	4	7	16	49	28
		$\Sigma 10$	$\Sigma 22$	$\Sigma 45$	$\Sigma 58$	$\Sigma 219$	$\Sigma 111$

Nhìn vào bảng số liệu trên ta có:

$$\bar{x} = 2,2; \quad \bar{y} = 4,5; \quad \bar{x^2} = 5,8; \quad \bar{y^2} = 21,9; \quad \bar{xy} = 11,1; \quad \bar{xy} - \bar{x}\bar{y} = 1,2$$

$$s_x^2 = 5,8 - (2,2)^2 = 0,96; \quad s_y^2 = 21,9 - (4,5)^2 = 1,645; \quad s_x s_y = 1,26$$

$$\text{Hệ số tương quan mẫu là: } r = \frac{1,2}{1,26} = 0,952$$

Phương trình hồi qui tuyến tính thực nghiệm của Y theo X là:

$$y = r \frac{s_y}{s_x} (x - \bar{x}) + \bar{y} = 0,952 \frac{1,28}{0,98} (x - 2,2) + 4,5 = 1,24x + 1,76$$

**Ví dụ 4.2.2.**

Ở một vùng có nghề phụ thủ công, quan sát 10 gia đình về 2 tiêu thức: Số trẻ em dưới 16 tuổi (X) và thu nhập thêm bằng nghề phụ (Y đơn vị nghìn đồng) thu được số liệu sau:

Gia đình	A	B	C	D	E	G	H	I	K	L
Số trẻ em dưới 16 tuổi (X)	3	5	2	4	4	4	6	1	3	3
Thu nhập Y (Nghìn đồng)	58	89	72	71	68	64	98	49	59	62

a/ Hãy khảo sát mối tương quan giữa hai tiêu thức trên

b/ Xây dựng đường hồi qui bình phương trung bình tuyến tính thực nghiệm của thu nhập theo số trẻ em

c/ Ước lượng sai số bình phương trung bình

Giải: a/ Ta thấy ở đây  $m_i = 1$  với mọi  $i$ . Đặt  $u_i = x_i - 4$  ta có bảng số liệu sau:

$x_i$	$y_i$	$u_i$	$u_i^2$	$y_i^2$	$u_i y_i$
3	58	-1	1	3364	-58
5	89	1	1	7921	89
2	72	-2	4	5184	-144
4	71	0	0	5041	0
4	68	0	0	4624	0
4	64	0	0	4096	0
6	98	2	4	9604	196
1	49	-3	9	2401	-147
3	59	-1	1	3481	-59
3	62	-1	1	3844	-62
$\Sigma$	690	-5	21	49560	-185

Từ bảng trên ta thấy:

$$\bar{u} = -5/10 = -0,5 \quad \Rightarrow \quad \overline{u^2} = 0,25 \quad \Rightarrow \quad \bar{x} = 3,5$$

$$\bar{y} = 690/10 = 69 \quad \overline{y^2} = 4761$$

$$\overline{u^2} = 21/10 = 2,1 \quad \overline{y^2} = 49560/10 = 4956$$

$$\overline{uy} = -185/10 = -18,5$$

$$s_u = \sqrt{2,1 - 0,25} \approx 1,36 = s_x \quad s_y = \sqrt{4956 - 4761} \approx 13,964$$

$$\Rightarrow r = \frac{-18,5 - (-0,5) \times 69}{1,36 \times 13,964} \approx 0,842$$

Điều đó chứng tỏ có sự phụ thuộc tuyến tính chặt chẽ giữa thu nhập và số trẻ em. Sự phụ thuộc là đồng biến.

b/ Phương trình đường hồi qui bình phương tuyến tính thực nghiệm của thu nhập theo số trẻ là:

$$y = 0,842 \times \frac{13,964}{1,36} (x - 3,5) + 69$$

$$y = 8,6446x + 39,256$$

c/ Ước lượng sai số bình phương trung bình  $s_{y/x}^2 = 195 \times (1 - 0,842^2) = 56,745$

**Ví dụ 4.2.3.** Nghiên cứu mối liên hệ giữa X là số tiền đầu tư cho việc phòng bệnh tính theo đầu người và Y là tỷ lệ người mắc bệnh ở 50 địa phương, ta thu được bảng tương quan thực nghiệm sau đây: (X: đơn vị nghìn đồng; Y: tính theo phần trăm).

X \ Y	2	2,5	3	3,5	4
100				2	3

200			3	6	2
300		4	6	3	
400	1	6	4	1	
500	6	3			

a/ Tìm hệ số tương quan tuyến tính.

b/ Tìm phương trình hồi qui tuyến tính của Y đối với X qua mẫu trên.

c/ Nếu năm sau đầu tư cho phòng bệnh là 600.000 đồng/người thì tỷ lệ mắc bệnh khoảng bao nhiêu phần trăm?

*Giải:* Đặt  $u_i = \frac{x_i - 300}{100}$ ;  $v_j = \frac{y_j - 3}{0,5}$

Ta có bảng số liệu sau:

U \ V	-2	-1	0	1	2	$n_i$	$n_i u_i$	$n_i u_i^2$
-2				2 / -4	3 / -12	5	-10	20
-1			3 0	6 / -6	2 / -4	11	-11	11
0		4 0	6 0	3 0		13	0	0
1	1 / -2	6 / -6	4 0	1 / 1		12	12	12
2	6 / -24	3 / -6				9	18	36
$n_i$	7	13	13	12	5	50	9	79
$n_i v_i$	-14	-13	0	12	10	-5		
$n_i v_i^2$	28	13	0	12	20	73		-63

$$\sum n_i u_i = 9; \sum n_j v_j = -5; \sum n_i u_i^2 = 79; \sum n_j v_j^2 = 73; \sum n_{uv} = -63$$

$$\bar{u} = 0,18; \bar{v} = -0,1; \bar{u}^2 = 1,58; \bar{v}^2 = 1,46; \bar{uv} = -1,26; s_u = 1,244; s_v = 1,204$$

$$a/ r = \frac{\bar{uv} - \bar{u} \cdot \bar{v}}{s_u s_v} = -0,829$$

$$b/ \bar{x} = 318; \bar{y} = 2,95; s_x = 124,4; s_y = 0,602$$

$$\text{Phương trình hồi qui } y = r \frac{s_y}{s_x} (x - \bar{x}) + \bar{y} = -0,004x + 4,222$$

$$c/ y (x=600) = -0,004 \cdot 600 + 4,222 = 1,822 \%$$

#### BÀI TẬP CHƯƠNG 4

1. Cho bảng tương quan thực nghiệm 2 chiều: (Từ bài ý 1 đến bài ý 22)

a/ Hãy tìm hệ số tương quan mẫu?

b/ Viết phương trình đường hồi qui tuyến tính thực nghiệm của Y theo X

1/

Y \ X	100	200	300	400	500
26	8	6			
30	2	10	4		
34		4	26	6	
38			5	10	7
42				4	8

2/

Y \ X	100	200	300	400	500
20	8	6			
30	2	10	4		
40		4	26	6	
50			5	10	7
60				4	8

3/

Y \ X	100	200	300	400	500
26	8	6			
30	2	10	4		
34		4	26	6	
38			5	10	7
42				4	8

4/

Y \ X	50	100	150	200	250
100				4	4
110		2	6	1	1
120	1	4	2		
130	3		1		1

5/

Y \ X	50	100	150	200	250
200				4	4
210		2	6	1	1
220	1	4	2		
230	3		1		1

6/

Y \ X	50	60	70	80	90



100				4	4
110		2	6	1	1
120	1	4	2		
130	3		1		1

7/

Y \ X	50	100	150	200	250
100				4	4
110		2	6	1	1
120	1	4	2		
130	3		1		1

8/

Y \ X	10	20	30	40	50	60
15	5	7				
25		20	23			
35			30	47	2	
45			10	11	20	6
55				9	7	3

9/

Y \ X	10	11	12	13
5				2
4			1	2
3		2	2	
2		1	2	
1	2	1		

10/

Y \ X	10	20	30	40	50	60
25	5	7				
35		20	23			
45			30	47	2	
55			10	11	20	6
65				9	7	3

11/

Y \ X	10	20	30	40
-------	----	----	----	----

5				2
4			1	2
3		2	2	
2		1	2	
1	2	1		

12/

Y \ X	24	27	30	33	36
120	1	3			
125		2	6	1	
130		1	5	5	
135		1	6	7	2
140			1	4	2
145				1	1
150					1

13/

X \ Y	25	28	31	34	37
50	1	3			
55		2	6	1	
60		1	5	5	
65		1	6	7	2
70			1	4	2
75				1	1
80					1

14/

X \ Y	20	30	40	50	60
120	1	3			
130		2	6	1	
140		1	5	5	
150		1	6	7	2
160			1	4	2
170				1	1
180					1

2. Kiểm tra hai môn toán và vật lý một nhóm 10 sinh viên được chọn ngẫu nhiên từ một lớp ta có kết quả sau:

Điểm toán (X)	7	6	7	10	4	5	7	8	8	9
Điểm vật lý (Y)	6	7	7	9	5	3	8	9	6	7

- a/ Hãy tìm hệ số tương quan mẫu?  
 b/ Viết phương trình đường hồi qui tuyến tính thực nghiệm của Y theo X.
3. Số vi khuẩn Y sinh sản sau X giờ được ghi lại trong bảng sau qua một thí nghiệm:

Thời gian (X)	0	1	2	3	4	5	6	7	8
Số vi khuẩn (Y)(triệu)	30	32	35	40	48	52	58	62	69

- a/ Hãy tính hệ số tương quan mẫu.  
 b/ Tìm phương trình đường hồi qui tuyến tính thực nghiệm của Y theo X.
4. Để nghiên cứu về lượng Protein chứa trong hạt lúa mì người ta tiến hành điều tra trên 10 thửa ruộng và được kết quả sau:

Năng suất X	9,0	9,5	10,0	10,5	11,0	11,5	12,0	12,5	13,0	13,5
Tỉ lệ Protein Y	10,0	10,2	11,0	10,5	12,0	12,2	12,5	12,6	12,7	12,8

- a/ Hãy tính hệ số tương quan mẫu.  
 b/ Tìm phương trình đường hồi qui tuyến tính thực nghiệm của Y theo X.

## TÀI LIỆU THAM KHẢO

- [1] Nguyễn Đình Trí, *Toán học cao cấp tập 1*, Nhà xuất bản giáo dục, năm 2004.
- [2] Nguyễn Đình Trí, *Toán học cao cấp tập 2*, Nhà xuất bản giáo dục, năm 2004.
- [3] Nguyễn Đình Trí, *Toán học cao cấp tập 3*, Nhà xuất bản giáo dục, năm 2002.
- [4] Đào Hữu Hồ, *Xác suất thống kê*, Nhà xuất bản Đại học quốc gia Hà Nội, 2007.
- [5] Nguyễn Văn Cao, *Giáo trình Lý thuyết Xác suất & Thống kê toán*, Trường Đại học Kinh tế quốc dân, Nhà xuất bản Thống kê, 2005.

## **Phụ lục 2**

Bảng giá trị của hàm  $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-t^2/2} dt$  (Hàm Láp-la-xơ)

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
1	2	3	4	5	6
0,00	0,0000	0,41	0,1591	0,82	0,2939
0,01	0,0040	0,42	0,1628	0,83	0,2967
0,02	0,0080	0,43	0,1664	0,84	0,3995
0,03	0,0120	0,44	0,1700	0,85	0,3023
0,04	0,0160	0,45	0,1736	0,86	0,3051
0,05	0,0199	0,46	0,1772	0,87	0,3078
0,06	0,0239	0,47	0,1808	0,88	0,3106
0,07	0,0279	0,48	0,1844	0,89	0,3133
0,08	0,0319	0,49	0,1879	0,90	0,3159
0,09	0,0359	0,50	0,1915	0,91	0,3186
0,10	0,0398	0,51	0,1950	0,92	0,3412
0,11	0,0438	0,52	0,1985	0,93	0,3238
0,12	0,0478	0,53	0,2019	0,94	0,3264
0,13	0,0517	0,54	0,2054	0,95	0,3289
0,14	0,0557	0,55	0,2088	0,96	0,3315
0,15	0,0596	0,56	0,2123	0,97	0,3340
0,16	0,0636	0,57	0,2157	0,98	0,3365
0,17	0,0675	0,58	0,2190	0,99	0,3389
0,18	0,0714	0,59	0,2224	1,00	0,3413
0,19	0,0753	0,60	0,2257	1,01	0,3438
0,20	0,0793	0,61	0,2291	1,02	0,3461
0,21	0,0832	0,62	0,2324	1,03	0,3485
0,22	0,0871	0,63	0,2357	1,04	0,3508
0,23	0,0910	0,64	0,2389	1,05	0,3531
0,24	0,0948	0,65	0,2422	1,06	0,3554
0,25	0,0984	0,66	0,2454	1,07	0,3577
0,26	0,1026	0,67	0,2486	1,08	0,3599
0,27	0,1064	0,68	0,2517	1,09	0,3621
0,28	0,1103	0,69	0,2549	1,10	0,3643
0,29	0,1141	0,70	0,2580	1,11	0,3665
0,30	0,1179	0,71	0,2611	1,12	0,3686
0,31	0,1217	0,72	0,2612	1,13	0,3708
0,32	0,1255	0,73	0,2673	1,14	0,3729
0,33	0,1293	0,74	0,2703	1,15	0,3749
0,34	0,1331	0,75	0,2734	1,16	0,3770
0,35	0,1368	0,76	0,2764	1,17	0,3790
0,36	0,1406	0,77	0,2794	1,18	0,3810
0,37	0,1443	0,78	0,2823	1,19	0,3830
0,38	0,1480	0,79	0,2852	1,20	0,3849
0,39	0,1517	0,80	0,2881	1,21	0,3869
0,40	0,1554	0,81	0,2910	1,22	0,3883
u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$

1	2	3	4	5	6
1,23	0,3907	1,68	0,4535	2,26	0,4881
1,24	0,3925	1,69	0,4545	2,28	0,4887
1,25	0,3944	1,70	0,4554	2,30	0,4893
1,26	0,3962	1,71	0,4564	2,32	0,4898
1,27	0,3980	1,72	0,4573	2,34	0,4904
1,28	0,3997	1,73	0,4582	2,36	0,4909
1,29	0,4015	1,74	0,4591	2,38	0,4913
1,30	0,4032	1,75	0,4599	2,40	0,4918
1,31	0,4049	1,76	0,4608	2,42	0,4922
1,32	0,4066	1,77	0,4616	2,44	0,4927
1,33	0,4082	1,78	0,4625	2,46	0,4931
1,34	0,4099	1,79	0,4633	2,48	0,4934
1,35	0,4115	1,80	0,4641	2,50	0,4938
1,36	0,4131	1,81	0,4649	2,52	0,4941
1,37	0,4147	1,82	0,4656	2,54	0,4945
1,38	0,4162	1,83	0,4664	2,56	0,4948
1,39	0,4177	1,84	0,4671	2,58	0,4951
1,40	0,4192	1,85	0,4678	2,60	0,4953
1,41	0,4207	1,86	0,4686	2,62	0,4956
1,42	0,4222	1,87	0,4693	2,64	0,4959
1,43	0,4236	1,88	0,4699	2,66	0,4961
1,44	0,4251	1,89	0,4706	2,68	0,4963
1,45	0,4265	1,90	0,4713	2,70	0,4965
1,46	0,4279	1,91	0,4719	2,72	0,4967
1,47	0,4292	1,92	0,4726	2,74	0,4969
1,48	0,4306	1,93	0,4732	2,76	0,4971
1,49	0,4319	1,94	0,4738	2,78	0,4973
1,50	0,4332	1,95	0,4744	2,80	0,4974
1,51	0,4345	1,96	0,4750	2,82	0,4976
1,52	0,4357	1,97	0,4756	2,84	0,4977
1,53	0,4370	1,98	0,4761	2,86	0,4979
1,54	0,4382	1,99	0,4767	2,88	0,4980
1,55	0,4394	2,00	0,4772	2,90	0,4981
1,56	0,4406	2,02	0,4783	2,92	0,4982
1,57	0,4418	2,04	0,4793	2,94	0,4984
1,58	0,4429	2,06	0,4803	2,96	0,4985
1,59	0,4441	2,08	0,4812	2,98	0,4986
1,60	0,4452	2,10	0,4921	3,00	0,49865
1,61	0,4463	2,12	0,4830	3,20	0,49931
1,62	0,4474	2,14	0,4838	3,40	0,49966
1,63	0,4484	2,16	0,4846	3,60	0,499841
1,64	0,4495	2,18	0,4854	3,80	0,499928
1,65	0,4505	2,20	0,4861	4,00	0,499968
1,66	0,4515	2,22	0,4868	4,50	0,499997
1,67	0,4525	2,24	0,4875	5,00	0,499997

**Phụ lục 3**

Bảng giá trị phân vị  $U_\alpha$

$\alpha$	$U_\alpha$	$\alpha$	$U_\alpha$	$\alpha$	$U_\alpha$	$\alpha$	$U_\alpha$
0,50	0,00	0,75	0,674	0,95	1,645	0,975	1,96
0,51	0,025	0,76	0,706	0,951	1,655	0,976	1,977
0,52	0,030	0,77	0,739	0,952	1,665	0,977	1,995
0,53	0,075	0,78	0,772	0,953	1,675	0,978	2,014
0,54	0,100	0,79	0,806	0,954	1,685	0,979	2,034
0,55	0,126	0,80	0,842	0,955	1,695	0,980	2,054
0,56	0,151	0,81	0,878	0,956	1,706	0,981	2,075
0,57	0,176	0,82	0,915	0,957	1,717	0,982	2,097
0,58	0,202	0,83	0,954	0,958	1,728	0,983	2,120
0,59	0,228	0,84	0,994	0,959	1,739	0,984	2,144
0,60	0,253	0,85	1,036	0,960	1,751	0,985	2,170
0,61	0,279	0,86	1,080	0,961	1,762	0,986	2,197
0,62	0,305	0,87	1,126	0,962	1,774	0,987	2,226
0,63	0,332	0,88	1,175	0,963	1,787	0,988	2,257
0,64	0,358	0,89	1,227	0,964	1,799	0,989	2,290
0,65	0,385	0,90	1,282	0,965	1,812	0,990	2,326
0,66	0,412	0,905	1,311	0,966	1,825	0,991	2,366
0,67	0,440	0,910	1,341	0,967	1,837	0,992	2,409
0,68	0,468	0,915	1,372	0,968	1,852	0,993	2,457
0,69	0,496	0,920	1,405	0,969	1,866	0,994	2,512
0,70	0,524	0,925	1,440	0,970	1,881	0,995	2,576
0,71	0,553	0,930	1,476	0,971	1,896	0,996	2,652
0,72	0,583	0,935	1,514	0,972	1,911	0,997	2,748
0,73	0,613	0,940	1,555	0,973	1,927	0,998	2,878
0,74	0,643	0,945	1,598	0,974	1,943	0,999	3,090

Với  $\alpha < 0,5$  suy ra giá trị  $U_\alpha = -U_{1-\alpha}$

Phân vị student  $t_{\alpha}^{(n)}$

$\alpha \backslash n$	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,183	4,541	5,842
4	1,533	2,132	2,776	3,767	4,601
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,335
9	1,383	1,833	2,262	2,821	3,325
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,974
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750