**California Polytechnic State University, San Luis Obispo**

**JAY L. DEVORE**

Tài liệu môn học

# Probability and Statistics for Engineering and Sciences

# XÁC SUẤT THỐNG KÊ ỨNG DỤNG

Người dịch:
Nguyễn Hồng Nhung
Hoàng Thị Minh Thảo
Lê Thị Mai Trang
Nguyễn Ngọc Tứ

**Bộ môn Toán - ĐH SPKT, Tp. Hồ Chí Minh - Năm 2017**

# Mục lục

# MỞ ĐẦU

# Chương 1

# TỔNG QUAN VÀ THỐNG KÊ MÔ TẢ

## Giới thiệu

# Chương 2

# PHÉP TÍNH XÁC SUẤT

# Chương 3

# BIẾN NGẪU NHIÊN RỜI RẠC VÀ PHÂN PHỐI XÁC SUẤT

# Chương 4

# BIẾN NGẪU NHIÊN LIÊN TỤC VÀ PHÂN PHỐI XÁC SUẤT

Giới thiệu

4.1 Hàm mật độ xác suất

4.2 Hàm phân phối tích lũy và các số đặc trưng

4.3 Phân phối chuẩn

4.4 Phân phối mũ và Gamma

4.5 Một số phân phối liên tục khác

4.6 Đồ thị xác suất

Tóm tắt và Bài tập

# Chương 5

# PHÂN PHỐI XÁC SUẤT ĐỒNG THỜI VÀ MẪU NGẪU NHIÊN

## Giới thiệu

Trong chương 3 và 4, chúng ta đã phát triển các mô hình xác suất cho một biến ngẫu nhiên đơn biến. Nhiều vấn đề về xác suất thống kê liên quan đến các biến ngẫu nhiên đồng thời. Trong chương này, đầu tiên chúng ta thảo luận về các mô hình xác suất đồng thời của các biến ngẫu nhiên, đặc biệt nhấn mạnh vào trường hợp trong đó các biến số độc lập với nhau. Sau đó chúng ta nghiên cứu các giá trị kì vọng của các hàm số của các biến ngẫu nhiên, gồm hiệp phương sai và tương quan để đo mức độ liên kết giữa hai biến.

Ba phần cuối của chương xem xét các hàm của n biến ngẫu nhiên $X_1, X_2, ..., X_n$, và tập trung đặc biệt vào trung bình của chúng $(X_1 + ... + X_n)/n$. Một hàm như vậy không những là một biến ngẫu nhiên mà còn là một thống kê. Các phương pháp từ xác suất sẽ được sử dụng để lấy thông tin cho phân phối của một thống kê. Ta được kết quả đầu tiên là Định lý giới hạn trung tâm (Central Limit Theorem), cơ sở cho nhiều quy trình suy luận liên quan đến kích cỡ mẫu lớn.

## 5.1 Phân phối đồng thời của các biến ngẫu nhiên

Có rất nhiều tình huống thử nghiệm trong đó có nhiều hơn một biến ngẫu nhiên được người quan sát quan tâm. Trước tiên chúng ta xem xét phân phối xác suất đồng thời cho hai biến ngẫu nhiên rời rạc, sau đó cho hai biến liên tục và cuối

cùng cho nhiều hơn hai biến.

### 5.1.1 Hai biến ngẫu nhiên rời rạc

Hàm xác suất khối của một biến ngẫu nhiên rời rạc đơn $X$ xác định xác suất khối được đặt trên mỗi giá trị có thể của $X$. Hàm xác suất khối đồng thời của hai biến rời rạc $X$ và $Y$ mô tả xác suất khối được đặt trên mỗi cặp giá trị $(x, y)$.

**Định nghĩa 5.1.1.** Cho $X$ và $Y$ là hai biến rời rạc được định nghĩa trên không gian mẫu $\Omega$ của một phép thử. Hàm xác suất khối $p(x, y)$ cho mỗi cặp số $(x, y)$ được định nghĩa bởi

$$p(x, y) = P(X = x \text{ and } Y = y)$$

thỏa $p(x, y) \geq 0$ và $\Sigma_x \Sigma_y p(x, y) = 1$.

**Ví dụ 5.1** Cho biến ngẫu nhiên rời rạc X là chiều cao sinh viên (mét) với tập giá trị 1,5 ; 1,6 ; 1,7, biến ngẫu nhiên Y là cân nặng của sinh viên (kg) với tập giá trị là 50; 60; 65 và bảng phân phối xác suất đồng thời của X và Y như sau:

| | | Y | | |
|---|---|---|---|---|
| | $p_{X,Y}$ | 50 | 60 | 65 |
| X | 1,5 | 0,1 | 0,2 | 0,1 |
| | 1,6 | 0,2 | 0,05 | 0,1 |
| | 1,7 | 0,05 | 0,1 | 0,1 |

Khi đó tính các xác suất
$p(1,5;60) = P(X = 1,5; Y = 60) = 0,2$
$P(Y \geq 60) = p(1,5;60)+p(1,6;60)+p(1,7;60)+p(1,5;65)+p(1,6;65)+p(1,7;65) = 0,2+0,05+0,1+0,1+0,1+0,1 = 0,65$
$P(X = 1,6) = 0,2+0,05+0,1 = 0,35$

**Định nghĩa 5.1.2. Hàm xác suất khối biên duyên** của $X$, được kí hiệu là $p_X$, được định nghĩa $p_X(x) = \Sigma_{y:p(x,y)>0} \quad p(x, y)$ với mỗi giá trị có thể của $x$.
Tương tự, **hàm xác suất khối biên duyên** của $Y$ là $p_Y(y) = \Sigma_{x:p(x,y)>0} p(x, y)$ với mỗi giá trị có thể của $y$.

Việc sử dụng từ "biên duyên" ở đây là kết quả của thực tế rằng nếu hàm xác suất khối đồng thời được biểu diễn trong một bảng chữ nhật, thì các tổng xác suất các hàng (cột) chính là xác suất khối biên duyên của $X$ và các tổng xác suất của các cột (hàng) chính là xác suất khối biên duyên của $Y$.

**Ví dụ 5.2** (sử dụng dữ liệu của ví dụ 5.1)

| | | Y | | |
|---|---|---|---|---|
| | $p_{X,Y}$ | 50 | 60 | 65 |
| X | 1,5 | 0,1 | 0,2 | 0,1 |
| | 1,6 | 0,2 | 0,05 | 0,1 |
| | 1,7 | 0,05 | 0,1 | 0,1 |

$p_X(1,5) = p(1,5;50) + p(1,5;60) + p(1,5;65) = 0,1 + 0,2 + 0,1 = 0,4$
$p_X(1,6) = p(1,6;50) + p(1,6;60) + p(1,56;65) = 0,2 + 0,05 + 0,1 = 0,35$
$p_X(1,7) = p(1,7;50) + p(1,7;60) + p(1,7;65) = 0,05 + 0,1 + 0,1 = 0,25$
Vậy hàm xác suất khối biên duyên của $X$ là:

$$p_X(x) = \begin{cases} 0,4 & \text{nếu} X = 1,5 \\ 0,35 & \text{nếu} X = 1,6 \\ 0,25 & \text{nếu} X = 1,7 \end{cases}$$

Tương tự hãy tính hàm xác suất khối biên duyên cho $Y$?

### 5.1.2 Hai biến ngẫu nhiên liên tục

Xác suất mà giá trị quan sát thấy của một biến ngẫu nhiên liên tục $X$ nằm trong một tập hợp một chiều A( ví dụ một khoảng)có được bằng cách lấy tích phân hàm mật độ $f(x)$ trên tập A. Tương tự như vậy, xác suất mà cặp $(X, Y)$ của các biến ngẫu nhiên liên tục rơi vào tập hai chiều A ( ví dụ hình chữ nhật) có được bằng cách lấy tích phân một hàm gọi tên là *hàm mật độ đồng thời* .

**Định nghĩa 5.1.3.** Cho $X$ và $Y$ là các biến ngẫu nhiên liên tục. Một **hàm mật độ đồng thời** $f(x, y)$ cho hai biến này là một hàm thỏa $f(x, y) \geq 0$ và $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)dxdy = 1$ . Khi đó với tập 2 chiều A ta có:

$$P[(X, Y) \in A] = \int \int_A f(x, y)dxdy$$

Đặc biệt, nếu A là hình chữ nhật 2 chiều $(x, y) : a \leq x \leq b, c \leq y \leq d$ thì

$$P[(X,Y) \in A] = P(a \le x \le b, c \le y \le d) = \int_a^b \int_c^d f(x,y)dxdy$$

Hàm mật độ biên duyên của mỗi biến có thể có được bằng cách tương tự như ta tìm cho trường hợp biến 2 chiều rời rạc. Hàm mật độ biên duyên của $X$ tại giá trị $x$ có được khi ta cố định $x$ trong cặp $(x,y)$ và lấy tích phân hàm mật độ đồng thời theo $y$. Còn nếu lấy tích phân hàm mật độ đồng thời theo $x$ thì được hàm mật độ biên duyên theo $Y$.

**Định nghĩa 5.1.4. Hàm mật độ xác suất biên duyên** của $X$ và $Y$ , được kí hiệu $f_X(x)$ và $f_Y(y)$ , được cho bởi công thức:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \text{ với } -\infty < x < \infty$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \text{ với } -\infty < y < \infty$$

**Ví dụ 5.3** Một ngân hàng mở hai loại dịch vụ A và B, ngẫu nhiên chọn 1 ngày, gọi $X$ là tỉ lệ thời gian mà dịch vụ A được sử dụng, và $Y$ là tỉ lệ thời gian dịch vụ B được dùng. Thì tập hợp $(X,Y)$ là một hình chữ nhật $D = \{(x,y) : 0 \le x \le 1, 0 \le y \le 1\}$. Giả sử hàm mật độ đồng thời của $(X,Y)$ cho bởi

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & \text{với} \quad 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

Chú ý rằng điều kiện của hàm mật độ đồng thời là $\int_{\infty}^{\infty} \int_{\infty}^{\infty} f(x,y)dxdy = 1$ và $f(x,y) \ge 0$ đã thỏa.

a/ Xác suất không dịch vụ nào bận hơn $1/4$ thời gian là:

$P(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4}) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x+y^2)dxdy = \frac{6}{5} \int_0^{1/4} \int_0^{1/4} xdxdy + \frac{6}{5} \int_0^{1/4} \int_0^{1/4} y^2 dxdy = 0,0109$

b/ Tính $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 \frac{6}{5}(x+y^2) = \frac{6}{5}x + \frac{2}{5}$

Hàm mật độ biên duyên của X ( chính là phân phối xác suất của thời gian bận của dịch vụ A mà không quan tâm đến dịch vụ B) là:

$$f_X(x) = \begin{cases} \frac{6}{5}x + \frac{2}{5} & \text{với} \quad 0 \le x \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

c/ Tìm hàm mật độ biên duyên của Y? Tính $P(\frac{1}{4} \le Y \le \frac{3}{4})$?

**Ví dụ 5.4** Một công ty hạt có thị trường hộp các loại hạt pha trộn chứa hạnh nhân, hạt điều và đậu phộng. Giả sử trọng lượng tịnh của mỗi hộp có thể là 1 kg, nhưng trọng lượng của từng loại hạt là ngẫu nhiên. Bởi vì ba trọng số cộng lại bằng 1, một mô hình xác suất đồng thời cho hai loại hạt sẽ cung cấp thông tin cần thiết về trọng lượng cho loại hạt thứ ba. Đặt $X$ là trọng lượng của hạnh nhân trong một chiếc hộp đã chọn và $Y$ là trọng lượng hạt điều. Thì vùng để hàm mật độ dương là $D = (x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1$ . Cho hàm mật độ đồng thời có dạng

$$f(x,y) = \begin{cases} kxy & \text{với} \quad 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1 \\ 0 & \text{nơi khác} \end{cases}$$

a/ Tính k ? (đs: 24)
b/ Tính xác suất mà hai loại hạt bất kì chiếm 50 phần trăm ? (Gợi ý: $A = \{(x, y) : 0 \leq x \leq 1, 0 \leq x \leq 1, x + y \leq 0, 5\}$) (đs: 0,0625)

### 5.1.3 Các biến ngẫu nhiên độc lập

Trong Chương 2, chúng ta đã chỉ ra một cách xác định tính độc lập của hai biến cố là thông qua điều kiện $P(A \cap B) = P(A).P(B)$. Đây là một định nghĩa tương tự cho sự độc lập của hai biến ngẫu nhiên.

**Định nghĩa 5.1.5.** Hai biến ngẫu nhiên $X$ và $Y$ là độc lập nếu mỗi cặp giá trị $x$ và $y$ thỏa

$$p(x, y) = p_X(x).p_Y(y) \text{ khi } X \text{ và } Y \text{ rời rạc}$$

hoặc

$$f(x, y) = f_X(x).f_Y(y) \text{ khi } X \text{ và } Y \text{ liên tục.}$$

Nếu hai công thức trên không thỏa cho tất cả $(x, y)$ thì $X$ và $Y$ được gọi là phụ thuộc.

**Ví dụ 5.5** Trong ví dụ 5.1 cho biến rời rạc, tồn tại $(p(1,6; 50) = 0,2) \neq (p(1,6) \times p(50) = 0.35 \times 0.35)$ nên X và Y không độc lập.
**Ví dụ 5.6** Trong ví dụ 5.3 ; tồn tại $f_X(\frac{3}{4}) = f_Y(\frac{3}{4}) = \frac{9}{16}; f(\frac{3}{4}, \frac{3}{4}) = 0 \neq \frac{9}{16}.\frac{9}{16}$ nên hai biến trên không độc lập.

Sự độc lập của hai biến ngẫu nhiên khá quan trọng khi mô tả thí nghiệm nghiên cứu khi mà $X$ và $Y$ không ảnh hưởng lẫn nhau. Ngoài ra ta còn có kết quả sau: $P(a \leq X \leq b, c \leq X \leq d) = P(a \leq X \leq b).P(c \leq X \leq d)$

### 5.1.4 Nhiều hơn 2 biến

Để mô hình hoá các hàm của hơn hai biến ngẫu nhiên, chúng ta mở rộng khái niệm về phân phối đồng thời của hai biến.

**Định nghĩa 5.1.6.** Nếu $X_1, X_2, ..., X_n$ là tất cả biến ngẫu nhiên rời rạc, hàm xác suất đồng thời của các biến ngẫu nhiên là :

$$p(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

Nếu các biến liên tục, hàm mật độ đồng thời của $X_1, X_2, ..., X_n$ là hàm $p(x_1, x_2, ..., x_n)$ chẳng hạn với bất kì $n$ khoảng $[a_1, b_1], ..., [a_n, b_n]$,

$$P(a_1 \leq X_1 \leq b_1, ..., a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f(x_1, ..., x_n) dx_n ... dx_1$$

Trong một phép thử nhị thức, mỗi thí nghiệm chỉ cho một kết quả trong 2 kết quả có thể xảy ra. Bây giờ xét một phép thử gồm $n$ thí nghiệm độc lập và giống nhau, mà trong đó mỗi phép thử chỉ cho một kết quả trong $r$ kết quả có thể xảy ra. Cho $p_i = P$ ( kết quả thứ i của bất kì phép thử cụ thể), và định nghĩa biến ngẫu nhiên $X_i =$ số lần thí nghiệm mà có kết quả $i (i = 1, ..., r)$. Một thí nghiệm như vậy được gọi là **phép thử đa thức - multinomial experiment** và hàm xác suất khối của $X_1, X_2, ..., X_n$ được gọi là **phân phối đa thức - multinomial distribution**. Bằng cách sử dụng lý luận tính tương tự với một biến trong trường hợp phân phối nhị thức, hàm xác suất đồng thời của $X_1, X_2, ..., X_n$ là

$$p(x_1, ..., x_r) = \begin{cases} \frac{n!}{(x_1!)(x_2!)...(x_r!)} . p_1^{x_1} ... p_r^{x_r}, & x_i = 0, 1, 2, ...; x_1 + ... + x_r = n \\ 0, \quad \text{nơi khác} \end{cases}$$

Trường hợp $r = 2$ sẽ cho phân phối nhị thức, với $X_1 =$số lần thành công và $X_2 = n = X_1 =$ số lần thất bại.

**Định nghĩa 5.1.7.** Biến ngẫu nhiên $X_1, X_2, ..., X_n$ được gọi là **độc lập** nếu với mỗi tập con $X_{i_1}, X_{i_2}, ..., X_{i_k}$ của các biến ( mỗi cặp, mỗi bộ ba , ...)thì hàm xác suất đồng thời hay hàm mật độ đồng thời bằng với tích của các hàm xác suất biên duyên hay các hàm mật độ biên duyên..

Vì vậy, nếu các biến độc lập với $n = 4$, thì hàm xác suất đồng thời hay hàm mật độ đồng thời của 2 biến là tích của hai hàm biên duyên, tương tự cho trường hợp 3 và 4 biến. Quan trong nhất là khi ta đã có $n$ biến độc lập thì hàm xác suất đồng thời hay hàm mật độ đồng thời là tích của $n$ hàm biên duyên.

### 5.1.5 Phân phối điều kiện

**Định nghĩa 5.1.8.** Cho $X$ và $Y$ là 2 biến ngẫu nhiên liên tục với hàm mật độ biên duyên $f(x,y)$ và hàm mật độ biên duyên của $X$ là $f_X(x)$. Thì với bất kì $X$ có giá trị $x$ mà $f_X(x) > 0$,**hàm mật độ điều kiện của Y cho bởi X=x** là

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \qquad -\infty < y < \infty$$

Nếu $X$ và $Y$ là biến rời rạc, thay hàm mật độ bằng hàm xác suất khối trong định nghĩa này thì được **hàm xác suất khối điều kiện của Y khi X=x.**

**Ví dụ 5.7** Dựa vào dữ liệu của ví dụ 5.3.
a/ Hàm mật độ điều kiện của $Y$ cho bởi $X = 0.8$ là
$f_{Y|X}(y|0,8) = \frac{f(0,8;y)}{f_X(0.8)} = \frac{1,2(0,8+y^2)}{1,2(0,8)+0,4} = \frac{1}{34}(24 + 30y^2)$ với $0 < y < 1$
b/ Xác suất để dịch vụ A bận trong nửa thời gian đầu khi biết $X = 0.8$
$P(Y \leq 0,5|X = 0,8) = \int_{-\infty}^{0,5} f_{Y|X}(y|0,8)dy = \int_0^{0,5} \frac{1}{34}(24 + 30y^2)dy = 0,39$
c/ Kì vọng có điều kiện $X = 0,8$ là:
$E(Y|X = 0,8) = \int_{-\infty}^{\infty} y.f_{Y|X}(y|0,8)dy = \frac{1}{34}\int_0^1 y(24 + 30y^2)dy = 0.574$

## 5.2 Giá trị Kỳ vọng, Phương sai và Hiệp phương sai

Bất kỳ hàm $h(X)$ của một biến ngẫu nhiên đơn thì chính nó cũng là một biến ngẫu nhiên. Tuy nhiên, để tính toán $E[h(X)]$ thì không cần phải có phân phối xác suất của $h(X)$; thay vào đó, $E[h(X)]$ được tính như một trung bình trọng số(a weighted average ) của các giá trị $h(X)$, trong đó hàm trọng số (the weight function) là hàm xác suất khối $p(x)$ hoặc hàm mật độ $f(x)$ của $X$. Kết quả tương tự cho một hàm phân phối đồng thời 2 biến $h(x,y)$.

**Mệnh đề 5.2.1.** *Cho $X$ và $Y$ là biến ngẫu nhiên đồng thời với hàm xác suất khối $p(x,y)$ hoặc hàm mật độ $f(x,y)$ tương ứng với trường hợp biến rời rạc hoặc biến liên tục. Thì giá trị kì vọng của hàm $h(X,Y)$ được kí hiệu là $E[h(X,Y)]$ hoặc $\mu_{h(X,Y)}$ được tính như sau:*

$$E[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y).p(x,y), & \text{nếu } X \text{ và } Y \text{ rời rạc} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y).f(x,y)dxdy, & \text{nếu } X \text{ và } Y \text{ liên tục} \end{cases}$$

**Ví dụ 5.8** Năm người bạn đã mua vé cho một buổi hòa nhạc. Nếu vé là chỗ ngồi từ 1-5 trong một hàng nào đó và vé được phân phối ngẫu nhiên cho năm người, thì kì vọng cho số ghế khi chia cho 2 người là bao nhiêu? Đặt $X$ và $Y$ kí hiệu cho số ghế của người đầu tiên và người thứ 2. Cặp (X,Y) có thể có các giá trị (1, 2), (1, 3),. . . , (5, 4), và hàm xác suất khối đồng thời của $(X, Y)$ là

$$p(x,y) = \begin{cases} \frac{1}{20} & x = 1,...,5; y = 1,...,5; x \neq y \\ 0 & \text{nơi khác} \end{cases}$$

Lúc đó số chỗ ngồi chia cho 2 người là hàm $h(X,Y) = |X-Y|-1$, ta có bảng tương ứng sau:

| | | x | | | | |
|---|---|---|---|---|---|---|
| | $h_{X,Y}$ | 1 | 2 | 3 | 4 | 5 |
| y | 1 | — | 0 | 1 | 2 | 3 |
| | 2 | 0 | — | 0 | 1 | 2 |
| | 3 | 1 | 0 | — | 0 | 1 |
| | 4 | 2 | 1 | 0 | — | 0 |
| | 5 | 3 | 2 | 1 | 0 | — |

Vì vậy $E[h(X,Y)] = \sum_( x,y) \sum h(x,y).p(x,y) = \sum_{x=1}^{5} \sum_{y=1,x\neq y}^{5} (|x-y|-1).\frac{1}{20} = 1$

**Ví dụ 5.9** Hàm mật độ đồng thời của $X$ ( lượng hạnh nhân) và $Y$ (lượng hạt điều) trong 1kg hạt là:

$$f(x,y) = \begin{cases} 24xy & \text{với} \quad 0 \leq x \leq 1, 0 \leq y \leq 1, x+y \leq 1 \\ 0 & \text{nơi khác} \end{cases}$$

Nếu 1kg hạt hạnh nhân có giá 1\$ ; 1 kg hạt điều có giá 1,5\$ ; 1 kg hạt đậu phụng có giá 0.5\$; thì tổng giá trị của 1 hộp là:
$h(X,Y) = 1.X + 1,5.Y + 0,5.(1-X-Y) = 0,5 + 0,5.X + Y$
Kì vọng cho tổng giá trị là:
$E[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y).f(x,y)dxdy = \int_0^1 \int_0^{1-x} (0,5 + 0,5x + y).24xydydx = 1,1\$$

### 5.2.1 Hiệp phương sai

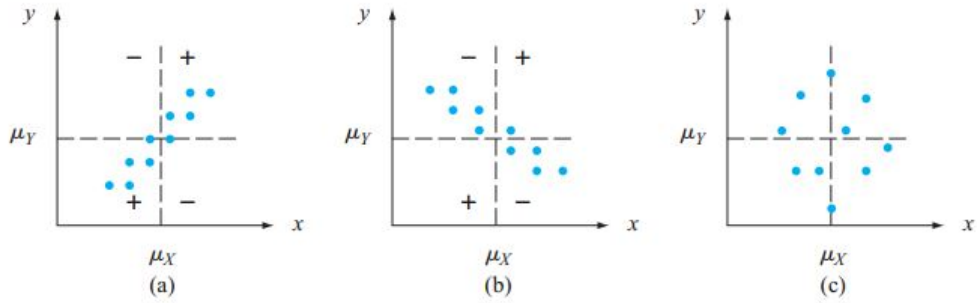
**Định nghĩa 5.2.2. Hiệp phương sai** giữa hai biến ngẫu nhiênX và $Y$ là

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(x,y) & , \quad X,Y \text{ rời rạc} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy, & X,Y \text{ liên tục} \end{cases}$$

Đó là, vì $X-\mu_X$ và $Y-\mu_Y$ là độ lệch của hai biến từ các giá trị trung bình tương ứng, hiệp phương sai là kì vọng của tích hai độ lệch. Chú ý rằng $Cov(X,X) = E[(X-\mu_X)^2] = V(X)$.

Giải thích định nghĩa: giả sử $X$ và $Y$ có một mối liên hệ dương với nhau, nghĩa là các giá trị lớn của $X$ có khuynh xảy ra với các giá trị lớn của $Y$ và các giá trị nhỏ của $X$ với giá trị nhỏ của $Y$. Thì hầu như xác suất khối hay hàm mật độ sẽ được liên quan đến $x-\mu_X$ và $y-\mu_Y$, khi cả hai giá trị $X,Y$ đều âm hay cả hai đều dương. Do đó với liên kết dương mạnh, $Cov(X,Y)$ sẽ hầu như dương (quite positive). Đối với một liên kết âm mạnh dấu của $X-\mu_X$ và $Y-\mu_Y$ sẽ có xu hướng ngược nhau, lúc đó thì $Cov(X,Y)$ hầu như âm. Nếu X,Y không có liên kết mạnh, tích âm và dương có xu hướng triệt tiêu lẫn nhau thì $Cov(X,Y) = 0$.

Hình dưới chỉ ra những khả năng khác nhau. Hiệp phương sai phụ thuộc vào các tập kết quả và xác suất. Hiệp phương sai có thể được thay đổi mà không cần điều chỉnh những tập kết quả, và có thể làm thay đổi nhiều đến giá trị của $Cov(X,Y)$.




p(x, y) = 1/10 for each of ten pairs corresponding to indicated points:
(a) positive covariance; (b) negative covariance; (c) covariance near zero

**Ví dụ 5.10** Cho hàm xác suất khối và hàm phân phối biên duyên của hai biến X,Y như sau:

| | | y | | |
|---|---|---|---|---|
| | $p_{X,Y}$ | 0 | 100 | 200 |
| x | 100 | 0,2 | 0,1 | 0,2 |
| | 250 | 0,05 | 0,15 | 0,3 |

| $x$ | 100 | 250 |
|---|---|---|
| $p_X(x)$ | 0,5 | 0,5 |

| $y$ | 0 | 100 | 200 |
|---|---|---|---|
| $p_Y(y)$ | 0,25 | 0,25 | 0,5 |

Áp dụng công thức tính Cov (X,Y) ? (Đs: 1875)

**Mệnh đề 5.2.3.** $Cov(X,Y) = E(X,Y) - \mu_X.\mu_Y$

**Ví dụ 5.11**

$$f(x,y) = \begin{cases} 24xy & \text{với} \quad 0 \le x \le 1, 0 \le y \le 1, x+y \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

$$f_X(x) = \begin{cases} 12x(1-x^2) & \text{với} \quad 0 \le x \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

với $f_Y(y)$ cũng có được tương tự bằng cách thay $x$ bởi $y$ trong $f_X(x)$.
Hãy tính $\mu_X$ , $\mu_Y$, $E(X,Y)$ từ đó suy ra $Cov(X,Y)$ ? (đs: -2/75)

### 5.2.2 Tương quan

**Định nghĩa 5.2.4. Hệ số tương quan** của $X$ và $Y$, được kí hiệu là $Corr(X,Y), \rho_{X,Y}$ hay chỉ là $\rho$ và được định nghĩa như sau

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

**Ví dụ 5.12** Từ dữ kiện của ví dụ 5.10. Hãy tính hệ số tương quan của $X$ và $Y$ ? (Đs: 0,301)

**Mệnh đề 5.2.5.**

1. *Nếu $a$ và $c$ cùng âm hoặc cùng dương thì*

$$Corr(aX+b, cY+d) = Corr(X,Y)$$

2. *Với bất kì hai biến ngẫu nhiên $X$ và $Y$ thì $-1 \le Corr(X,Y) \le 1$*

**Mệnh đề 5.2.6.**

1. *Nếu $X$ và $Y$ độc lập, thì $\rho = 0$ nhưng $\rho = 0$ thì không suy ra được là có sự độc lập.*

2. *$\rho = 1$ hay $-1$ khi và chỉ khi $Y = aX + b$ với $a, b$ là số nào đó và $a \neq 0$.*

Mệnh đề này chỉ ra rằng $\rho$ là một thước đo về bậc của mối quan hệ **tuyến tính** giữa $X$ và $Y$, và chỉ khi hai biến này quan hệ tuyến tính thì $\rho$ có thể âm hoặc dương tùy ý. Giá trị tuyệt đối của $\rho < 1$ chỉ ra rằng mối quan hệ không tuyến tính hoàn toàn, nhưng vẫn có mối quan hệ phi tuyến tính trong đó. Cũng như vậy khi $\rho = 0$ thì không chỉ ra rằng $X$ và $Y$ độc lập, mà chỉ là $X$ và $Y$ không có quan hệ tuyến tính. Khi $\rho = 0$, $X$ và $Y$ được gọi là **không tương quan**. Hai biến có thể không tương quan nhưng chưa chắc phụ thuộc vì có quan hệ phi tuyến tính mạnh ở đây, vì vậy cần cẩn thận đưa ra kết luận khi gặp $\rho = 0$.

## 5.3 Các phân phối thống kê

Các quan sát trong một mẫu đơn được kí hiệu trong chương 1 là $x_1, x_2, ..., x_n$. Xem như chọn hai mẫu khác nhau có kích thước $n$ từ cùng một phân phối tổng thể. $x_i$ trong mẫu thứ hai sẽ hầu như luôn khác nhau một chút với các giá trị của mẫu đầu tiên. Ví dụ, một mẫu đầu tiên của $n = 3$ xe ô tô loại đặc biệt có thể cho hiệu quả nhiên liệu $x_1 = 30,7, x_2 = 29,4, x_3 = 31,1$ , trong khi một mẫu thứ hai có thể cho $x_1 = 28,8, x_2 = 30.0$ và $x_3 = 32.5$. Trước khi chúng ta thu được dữ liệu, có sự không chắc chắn về giá trị của mỗi $x_i$. Vì điều này không chắc chắn, trước khi thu thập dữ liệu, chúng ta xem mỗi quan sát như là một biến ngẫu nhiên và kí hiệu mẫu bằng $X_1, X_2, ..., X_n$ (chữ in hoa cho biến ngẫu nhiên).
Sự thay đổi này trong các giá trị được quan sát giúp suy ra các giá trị của bất kỳ hàm nào của các quan sát mẫu - chẳng hạn như trung bình mẫu, độ lệch tiêu chuẩn mẫu. Trước tiên là để có $x_1, ..., x_n$, thì có một tất định cho giá trị $\overline{x}$, giá trị $s$...

**Định nghĩa 5.3.1.** Một **thống kê** là bất kì lượng nào mà giá trị có thể được tính từ dữ liệu mẫu. Trước khi thu thập dữ liệu, có một tất định về giá trị của bất kỳ thống kê đặc biệt sẽ cho kết quả. Do đó, thống kê là một biến ngẫu nhiên và sẽ được kí hiệu bởi một kí tự viết hoa; một kí tự viết thường được sử dụng để đại diện cho giá trị đã tính toán hoặc đã quan sát của thống kê.

Vì vậy trung bình mẫu trong thống kê được kí hiệu là $\overline{X}$, và giá trị tính được của thống kê này là $\overline{x}$. Tương tự $S$ đại diện cho độ lệch chuẩn mẫu và giá trị được tín kí hiệu là $s$.
Bất kì thống kê nào ( bản thân nó là một biến ngẫu nhiên) đều có một phân phối xác suất. Đáng chú ý là phân phối xác suất của trung bình mẫu $\overline{X}$ .

Phân phối xác suất của một thống kê đôi khi gọi là phân phối mẫu để nhấn mạnh rằng nó mô tả cách mà thống kê khác nhau về giá trị trên tất cả các mẫu được lựa chọn.

### 5.3.1 Mẫu ngẫu nhiên

Phân bố xác suất của bất kỳ thống kê cụ thể phụ thuộc không chỉ vào phân phối tổng thể (chuẩn, đều,...) và kích thước mẫu $n$ mà còn phụ thuộc phương pháp lấy mẫu. Xem xét lựa chọn một mẫu kích thước $n = 2$ từ tổng thể chỉ gồm ba giá trị 1, 5 và 10, và giả sử rằng thống kê được quan tâm là phương sai mẫu. Nếu lấy mẫu được thực hiện "có hoàn lại," thì $S^2 = 0$ khi $X_1 = X_2$. Tuy nhiên, $S^2$ không thể bằng 0 nếu mẫu "không hoàn lại". Vì vậy, $P(S^2 = 0) = 0$ với một phương pháp lấy mẫu, và xác suất này là dương với phương pháp lấy mẫu khác. Định nghĩa tiếp theo của chúng ta mô tả một phương pháp lấy mẫu thường gặp trong thực tế.

**Định nghĩa 5.3.2.** Các biến ngẫu nhiên $X_1, X_2, ..., X_n$ được gọi là mẫu ngẫu nhiên (đơn) cỡ mẫu $n$ nếu:

1. Các $X_i$ là biến ngẫu nhiên độc lập.

2. Các $X_i$ có cùng phân phối mẫu.

Nếu lấy mẫu có hoàn lại từ tổng thể vô hạn, thì điều kiện 1 và 2 được thỏa mãn. Những điều kiện này cũng gần như thỏa mãn nếu lấy mẫu không hoàn lại, tuy nhiên mẫu cỡ $n$ thì nhỏ hơn nhiều so với tổng thể cỡ $N$. Trong thực tế, nếu $n/N < 0.05$ ( nhiều nhất 5% tổng thể được lấy mẫu), chúng ta có thể tiến hành khi $X_i$ là biến ngẫu nhiên. Hiệu quả của phương pháp lấy mẫu này là phân phối xác suất của bất kỳ thống kê nào có thể dễ dàng thu được hơn so với bất kỳ phương pháp lấy mẫu nào khác.

Có hai phương pháp chung để thu thập thông tin về phân phối mẫu của một thống kê. Một phương pháp liên quan đến tính toán dựa trên các quy tắc xác suất, và thứ hai là thực hiện một thí nghiệm mô phỏng.

### 5.3.2 Tìm phân phối mẫu

Các quy tắc xác suất có thể được sử dụng để có được phân phối của một thống kê miễn là nó là một hàm "khá đơn giản" của các biến ngẫu nhiên $X_i$ và hoặc là có vài giá trị $X$ khác nhau trong tổng thể hoặc phân phối xác suất có dạng "tốt".

### 5.3.3 Những thí nghiệm mô phỏng

Phương pháp thứ hai để thu thập thông tin về phân phối mẫu của thống kê là thực hiện một thí nghiệm mô phỏng. Phương pháp này thường được sử dụng khi lấy đạo hàm thông qua các quy tắc xác suất là quá khó hoặc phức tạp để thực hiện. Một thí nghiệm như vậy luôn luôn được thực hiện với sự trợ giúp của máy tính. Các đặc điểm sau của thí nghiệm yêu cầu:

1. Thống kê của lãi suất ($\overline{X}, S$, một giá trị trung bình đã được cắt xén (trimmed),...)
2. Phân phối của tổng thể ( chuẩn với $\mu = 100, \sigma = 15$ ; phân phối đều với giới hạn dưới $A = 5$ và giới hạn trên $B = 10$ , ...)
3. Kích cỡ mẫu $n$ (ví dụ, $n = 10$ hoặc $n = 50$)
4. Số lần lặp lại $k$ (số mẫu cần lấy)

Sau đó sử dụng phần mềm thích hợp để lấy $k$ mẫu ngẫu nhiên khác nhau, mỗi kích thước $n$, từ phân phối tổng thể đã được thiết kế. Đối với mỗi mẫu, tính toán giá trị của thống kê và xây dựng một biểu đồ tần số của các giá trị $k$ . Biểu đồ này cung cấp một xấp xỉ phân phối mẫu của thống kê. Giá trị càng lớn hơn của $k$, thì xấp xỉ càng tốt hơn (phân phối mẫu thât sự sẽ xuất hiện khi $k \to \infty$ ). Trong thực hành, $k = 500$ hoặc 1000 thường là đủ nếu thống kê "khá đơn giản".

## 5.4 Phân phối của trung bình mẫu

Tầm quan trọng của trung bình mẫu $\overline{X}$ là để rút ra kết luận cho trung bình tổng thể $\mu$. Một số phương pháp suy luận dựa trên tính chất của phân phối mẫu $\overline{X}$ . Các tính chất tính toán và thí nghiệm mô phỏng đã nêu trong các phần trước khi ta đề cấp đến mối quan hệ $E(\overline{X})$ và $\mu$ hay $V(\overline{X}), \sigma^2$, và $n$

**Mệnh đề 5.4.1.** *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ một phân phối với trung bình $\mu$ và độ lệch chuẩn $\sigma$. Thì*

1. $E(\overline{X}) = \mu_{\overline{X}} = \mu$
2. $V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n$ *và* $\sigma_{\overline{X}} = \sigma^2/\sqrt{n}$

Thêm nữa, với $T_o = X_1 + ... + X_n$ ( tổng của mẫu), $E(T_o) = n\mu, V(T_o) = n\sigma^2, \sigma_{T_0} = \sqrt{n}\sigma$

Các chứng minh cho những kết quả trên sẽ được trình bày ở phần tiếp theo. Theo kết quả 1, phân phối mẫu của $\overline{X}$ (tức là xác suất) tập trung chính xác ở trung bình của tổng thể mà mẫu được chọn từ đó. Kết quả 2 cho thấy rằng phân phối $\overline{X}$ trở nên tập trung hơn về $\mu$ khi kích thước mẫu $n$ tăng lên. Sự khác biệt quan trọng là phân phối của $T_o$ trở nên lan rộng ra khi cỡ mẫu $n$ tăng lên.

**Ví dụ 5.13** Trong một thí nghiệm kiểm tra sức chịu mài mòn trên một mẫu titan, kì vọng số chu kỳ để phát ra âm thanh đầu tiên (dùng chỉ ra vết nứt đầu tiên) là $\mu = 28.000$ và độ lệch tiêu chuẩn của số chu kỳ là $\sigma = 5000$. Đặt $X_1, X_2, ..., X_{25}$ là một mẫu ngẫu nhiên có kích thước 25, trong đó mỗi $X_i$ là số chu kỳ trên một mẫu titan khác nhau ngẫu nhiên được chọn. Thì giá trị kì vọng của số trung bình mẫu của chu kỳ cho đến khi phát âm thanh đầu tiên là $E(\overline{X}) = \mu = 28.000$ và tổng kì vọng cho số chu kỳ cho 25 mẫu tiatan là $E(T_0) = n\mu = 25(28000) = 700000$. Độ lệch chuẩn của $\overline{X}$ (sai số tiêu chuẩn của trung bình) và của $T_0$ là:

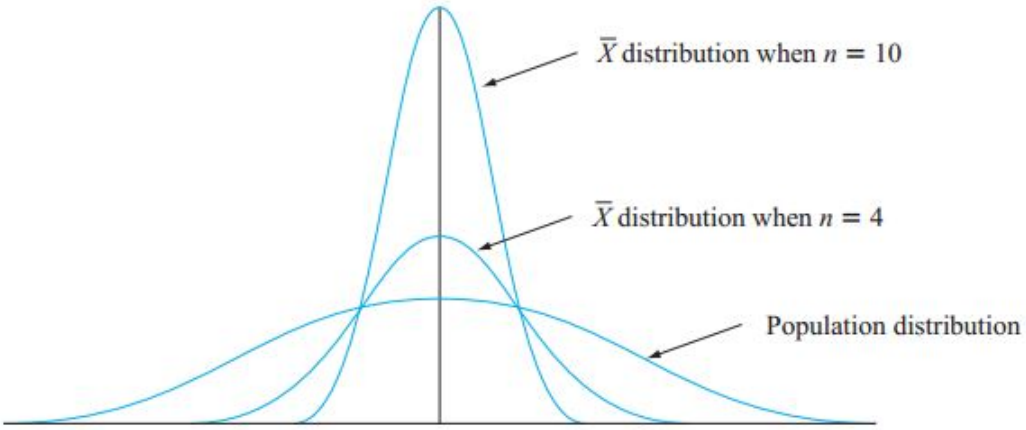
$$\sigma_{\overline{X}} = \sigma/\sqrt{n} = \frac{5000}{\sqrt{25}} = 1000$$
$$\sigma_{T_0} = \sqrt{n}\sigma = \sqrt{25}.5000 = 25000$$

Nếu cỡ mẫu tăng lên $n = 100$, $E(\overline{X})$ thì không đổi, nhưng $\sigma_{\overline{X}} = 500$ (bằng một nửa giá trị trước đó ) (kích cỡ mẫu phải được tăng gấp bốn lần để giảm một nửa độ lệch tiêu chuẩn của $\overline{X}$).

### 5.4.1 Trường hợp phân phối chuẩn tổng thể

**Mệnh đề 5.4.2.** *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ phân phối chuẩn với trung bình $\mu$ và độ lệch chuẩn $\sigma$. Thì với mọi n, $\overline{X}$ được phân phối chuẩn (với trung bình $\mu$ và độ lệch chuẩn $\sigma/\sqrt{n}$), như là $T_o$ ( với trung bình $n\mu$ và độ lệch chuẩn $\sqrt{n}\sigma$) .*

Chúng ta có dữ kiện để biết phân phối $\overline{X}$ và $T_o$ khi phân phối tổng thể là phân phối chuẩn. Đặc biệt, xác suất như $P(a \leq \overline{X} \leq b)$ và $P(c \leq T_o \leq d)$ có thể tính được bằng cách chuẩn hóa. Hình dưới giải thích cho mệnh đề




A normal population distribution and $\overline{X}$ sampling distributions

**Ví dụ 5.14** Thời gian cho một con chuột của một phân loài nhất định tìm đường ra một mê cung là một phân phối chuẩn với $\mu = 1,5$ phút và $\sigma = 0,35$ phút. Giả sử năm con chuột được chọn để quan sát. Đặt $X_1, ..., X_5$ cho thời gian của chúng trong mê cung.

a/ Giả sử $X_i$ là một mẫu ngẫu nhiên từ phân bố chuẩn này, tìm xác suất mà tổng thời gian $T_0 = X_1 + ... + X_5$ nằm trong khoảng từ 6 đến 8 phút?

Theo mệnh đề, $T_0$ có phân phối chuẩn với $\mu_{T_0} = n\mu = 5.(1,5) = 7,5$ và phương sai $\sigma^2_{T_0} = n\sigma^2 = 5(0,1225) = 0.6125$, vì vậy $\sigma_{T_0} = 0,783$.

$P(6 \leq T_0 \leq 8) = P(\frac{6-5,7}{0,783} \leq Z \leq \frac{8-7,5}{0,783}) = P(-1,92 \leq Z \leq 6.4) = \Phi(0,64) - \Phi(-1,92) = 0,7115$

b/ Xác định xác suất mà thời gian trung bình mẫu $\overline{X}$ nhiều nhất là 2 phút , với $\mu_{\overline{X}} = \mu = 1,5$ và $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 0,35/\sqrt{5} = 0,1565$. Thì

$P(\overline{X} \leq 2) = P(Z \leq \frac{2-1,5}{0,1565}) = P(Z \leq 3,19) = \Phi(3,19) = 0,9993$

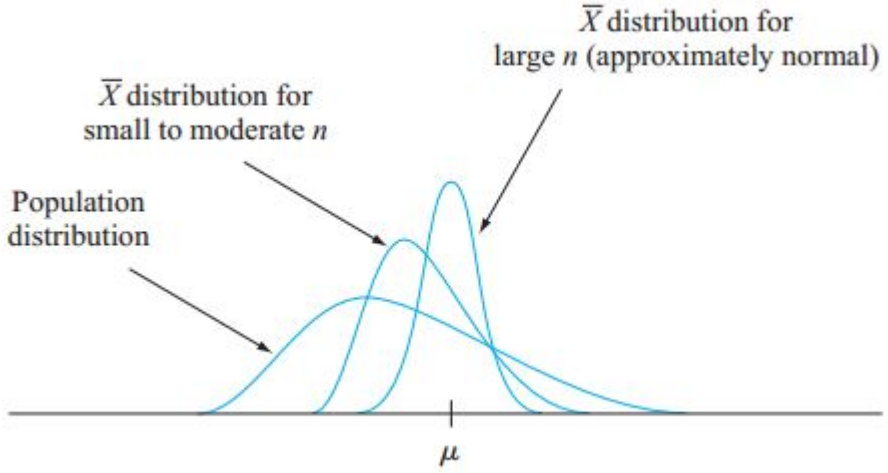
### 5.4.2 Định lý giới hạn trung tâm

Khi các $X_i$ có phân phối chuẩn, thì $\overline{X}$ cũng có phân phối chuẩn cho mỗi mẫu cỡ $n$ . Các kết quả từ ví dụ trên chỉ ra rằng khi phân phối tổng thể không chuẩn , lấy trung bình sẽ cho một phân phối tạo ra hình chuông hơn là mẫu được lấy. Nếu $n$ lớn, một đường cong chuẩn thích hợp sẽ xấp xỉ gần đúng phân phối thật của $\overline{X}$.

Phát biểu cho ý này được trình bày trong định lý xác suất quan trọng sau đây.

**Định lý 5.4.3.** *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ phân phối với trung bình $\mu$ và phương sai $\sigma^2$. Thì nếu $n$ đủ lớn, thì $\overline{X}$ có xấp xỉ gần đúng một phân phối chuẩn với $\mu_{\overline{X}} = \mu$ và $\sigma^2_{\overline{X}} = \sigma^2/n$ và $T_o$ cũng có xấp xỉ gần đúng một phân phối chuẩn với $\mu_{T_o} = n\mu, \sigma^2_{T_o} = n\sigma^2$. Giá trị $n$ càng lớn thì xấp xỉ càng tốt.*

Hình dưới minh họa Định lý Giới hạn Trung tâm. Theo ĐLGHTT, khi $n$ lớn và ta muốn tính xác suất như $P(a \leq \overline{X} \leq b)$, ta chỉ cần "giả vờ" là $\overline{X}$ là chuẩn và chuẩn hóa nó, sử dụng bảng phân phối chuẩn. Kết quả sẽ được xấp xỉ đúng.




The Central Limit Theorem illustrated

**Ví dụ 5.15** Lượng tạp chất cụ thể trong một lô sản phẩm hoá học nhất định là biến ngẫu nhiên có giá trị trung bình 4,0 g và độ lệch chuẩn 1,5 g. Nếu 50 lô được chuẩn bị độc lập, tính xác suất để trung bình mẫu tạp chất $\overline{X}$ nằm giữa 3,5 và 3,8g?
Ta có $n = 50$ đủ lớn để áp dụng ĐLGHTT. $\overline{X}$ xấp xỉ về phân phối chuẩn với trung bình $\mu_{\overline{X}} = 4$ và $\sigma_{\overline{X}} = 1,5/\sqrt{50} = 0,2121$ vì vậy
$P(3,5 \leq \overline{X} \leq 3,8) \approx P(\frac{3,5-4}{0,2121} \leq Z \leq \frac{3,8-4}{0,2121}) = \Phi(-0,94) - \Phi(-2,36) = 0,1645$

**Ví dụ 5.16** Một tổ chức tiêu dùng nhất định thường báo cáo số lượng khiếm khuyết lớn của mỗi chiếc ô tô mới mà nó thử nghiệm. Giả sử số lỗi này là một biến ngẫu nhiên với giá trị trung bình là 3.2 và độ lệch chuẩn 2.4. Trong số 100 xe được lựa chọn ngẫu nhiên, tính xác suất trung bình mẫu của những lỗi này nhiều hơn 4? (đs: 0,0004)

Định lý giới hạn trung tâm cung cấp cái nhìn sâu sắc về lý do tại sao nhiều biến ngẫu nhiên có phân phối xác suất gần với phân phối chuẩn. Ví dụ, sai số đo lường trong một thí nghiệm khoa học có thể được coi là tổng của một số điểm nhiễu loạn đáng kể và sai số của độ đo nhỏ.
Một khó khăn thực tế trong việc áp dụng ĐLGHTT là khi nào $n$ là đủ lớn. Vấn đề là độ chính xác của xấp xỉ cho một $n$ cụ thể phụ thuộc vào hình dạng phân phối cơ bản ban đầu được lấy mẫu. Nếu phân phối cơ bản gần với một đường cong chuẩn, thì xấp xỉ tốt ngay cả đối với một $n$ nhỏ, trong khi nếu nó xa phân phối chuẩn, thì cần $n$ lớn.

**Nguyên tắc chung:** Nếu $n > 30$ định lý giới hạn trung tâm được sử dụng.

Có những phân phối tổng thể mà thậm chí $n$ từ 40 hoặc 50 thì không đủ, nhưng những phân phối này hiếm gặp trong thực hành. Mặt khác, nguyên tắc chung thì bảo toàn đối với nhiều phân phối tổng thể, khi đó giá trị $n$ ít hơn 30 thì đã đủ. Ví dụ, trong trường hợp phân phối đều của tổng thể, ĐLGHTT cho xấp xỉ tốt với $n \geq 12$ .

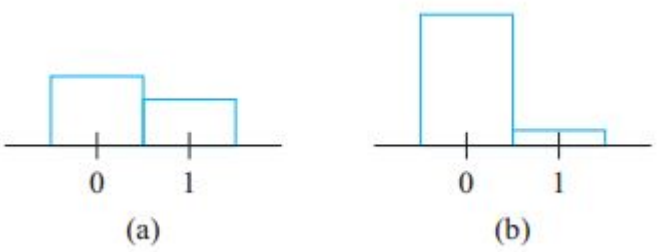
### 5.4.3 Những ứng dụng khác của định lý giới hạn trung tâm

ĐLGHTT có thể được sử dụng để làm rõ cho xấp xỉ chuẩn đối với phân phối nhị thức được thảo luận trong Chương 4. Nhớ lại rằng một biến nhị thức $X$ là số lần thành công trong một thí nghiệm nhị thức bao gồm $n$ thử nghiệm thành công / thất bại độc lập với $p = P(S)$ cho bất kỳ thử nghiệm cụ thể nào. Xác định một biến ngẫu nhiên mới $X_1$ bởi

$$X_1 = \begin{cases} 1 & , \quad \text{kết quả đầu thành công} \\ 0 & , \quad \text{kết quả đầu thất bại} \end{cases}$$

và định nghĩa $X_1, X_2, X_3, ..., X_n$ tương tự cho $n - 1$ thử nghiệm khác. Mỗi $X_i$ cho biết một phép thử tương ứng có thành công hay không.
Bởi vì các thử nghiệm là độc lập và $P(S)$ là hằng số qua các thử nghiệm, các biến ngẫu nhiên $X_i$ thì độc lập cùng phân phối ( mẫu ngẫu nhiên được lấy từ phân phối Bernoulli). ĐLGTTT chỉ ra rằng nếu $n$ là đủ lớn, cả tổng và trung bình của các $X_i$ đều xấp xỉ phân phối chuẩn. Khi các $X_i$ được cộng lại, tổng $S$ được thêm 1 nếu biến đó xuất hiện và $F$ được thêm 0 , vì vậy $X_1 + X_2 + ... + X_n = X$. Trung bình mẫu của các $X_i$ là $X/n$ chính là tỷ lệ thành công của mẫu. Nghĩa là, cả $X$ và $X/n$ đều

xấp xỉ chuẩn khi $n$ lớn. Kích cỡ mẫu cần thiết cho phép xấp xỉ này phụ thuộc vào giá trị của $p$: Khi $p$ gần 0,5, sự phân bố của mỗi $X_i$ thì đối xứng (xem hình dưới), trong khi phân phối là khá lệch khi $p$ gần 0 hoặc 1. Sử dụng xấp xỉ chỉ khi $np \geq 10$ và $n(1-p) \geq 10$ đảm bảo rằng $n$ đủ lớn để vượt qua bất kỳ sự sai lệch nào trong phân phối Bernoulli bên dưới:




Nhắc lại: $X$ có phân phối log chuẩn nếu $ln(X)$ có phân phối chuẩn.

**Mệnh đề 5.4.4.** *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ một phân phối mà chỉ có giá trị dương thỏa $[P(X_i > 0) = 1]$. Thì với n đủ lớn, tích $Y = X_1X_2...X_n$ sẽ xấp xỉ phân phối log chuẩn*

## 5.5 Phân phối của tổ hợp tuyến tính

Trung bình mẫu $\overline{X}$ và tổng của mẫu $T_o$ là những trường hợp đặc biệt của một loại biến ngẫu nhiên xuất hiện rất thường xuyên trong các ứng dụng thống kê.

**Định nghĩa 5.5.1.** Cho một tập $n$ các biến ngẫu nhiên $X_1, ..., X_n$ và $n$ hằng số $a_1, ..., a_n$ của biến ngẫu nhiên.

$$Y = a_1X_1 + ... + a_nX_n = \sum_{i=1}^{n} a_iX_i$$

được gọi là **tổ hợp tuyến tính** của biến ngẫu nhiên $X_i$

**Mệnh đề 5.5.2.** *Cho $X_1, X_2, ..., X_n$ có giá trị trung bình tương ứng $\mu_1, ..., \mu_n$ và phương sai tương ứng $\sigma_1^2, ..., \sigma_n^2$ .*

1. *$X_i$có độc lập hay không đều có:*

$$E(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n) = a_1\mu_1 + ... + a_n\mu_n$$

2. *Nếu* $X_1, ..., X_n$ *độc lập*

$$V(a_1X_1 + a_2X_2 + ... + a_nX_n = a_1^2V(X_1) + a_2^2V(X_2) + ... + a_n^2V(X_n) = a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2$$

*và*

$$\sigma_{a_1X_1+a_2X_2+...+a_nX_n} = \sqrt{a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2}$$

3. *Với bất kì* $X_1, ...X_n$
$V(a_1X_1 + a_2X_2 + ... + a_nX_n) = \Sigma_{i=1}^n\Sigma_{j=1}^n a_ia_jCov(X_i, X_j)$

**Ví dụ 5.17** Một trạm xăng bán ba loại xăng: regular, extra và super. Giá tương ứng các loại xăng này là 3 \$ ; 3,2\$ và 3,4\$ cho 1 gallon (4,4 lít). Gọi $X_1, X_2, X_3$ kí hiệu cho lượng gallon bán được trong 1 ngày nào đó. Giả sử mẫu $X_i$ độc lập với $\mu_1 = 1000$ , $\mu_2 = 500$ , $\mu_3 = 300, \sigma_1 = 100, \sigma_2 = 80, \sigma_3 = 50$. Doanh thu kí hiệu là $Y = 3.X_1 + 3, 2X_2 + 3, 4X_3$ . Hãy tính $E(Y), V(Y), \sigma(Y)$? (đs: 5620 ; 184,436 , 42,46)

### 5.5.1 Hiệu của hai biến ngẫu nhiên

Một trường hợp đặc biệt quan trọng của tổ hợp tuyến tính khi cho $n = 2, a_1 = 1, a_2 = -1$:

$$Y = a_1X_1 + a_2X_2 = X_1 - X_2$$

Từ đó ta có hệ quả sau:

**Hệ quả 5.5.3.** $E(X_1 - X_2) = E(X_1) - E(X_2)$ *với hai biến ngẫu nhiên bất kì* $X_1$ *và* $X_2$.
$V(X_1 - X_2) = V(X_1) + V(X_2)$ *nếu* $X_1$ *và* $X_2$ *độc lập.*

**Ví dụ 5.18** Một nhà sản xuất ô tô trang bị một mô hình với một động cơ có 6 xy-lanh và động cơ 4 xy-lanh. Đặt $X_1$ và $X_2$ là lượng dầu cần tương ứng cho động cơ 6 xy-lanh và 4 xy-lanh. Với $\mu_1 = 22, \mu_2 = 26, \sigma_1 = 1, 2; \sigma_2 = 1, 5$ hãy tính $E(X_1 - X_2), V(X_1 - X_2), \sigma_{X_1-X_2}$? (đs: -4 ; 3,69 ; 1,92)

### 5.5.2 Trường hợp của biến ngẫu nhiên liên tục

Khi $X_i$ là mẫu ngẫu nhiên lấy từ phân phối chuẩn, $\overline{X}$ và $T_o$ thì cũng có phân phối chuẩn. Đây là kết quả tổng quát hơn liên quan đến tổ hợp tuyến tính.

**Mệnh đề 5.5.4.** *Nếu $X_1, X_2, ..., X_n$ độc lập, có phân phối chuẩn ( có thể là trung bình khác nhau, phương sai khác nhau), thì bất kỳ tổ hợp tuyến tính của $X_i$ cũng có phân phối chuẩn. Đặc biệt, hiệu của $X_1 - X_2$ cũng được phân phối chuẩn.*

**Ví dụ 5.19** Theo ví dụ trên, cho tổng thu nhập khi bán 3 loại xăng ở 1 trạm là $Y = 3X_1 + 3,2X_2 + 3,4X_3$ và ta tính $\mu_Y = 5620$ , $\sigma_Y = 429,46$. Nếu $X_i$ có phân phối chuẩn thì xác suất để doanh thu hơn 4500 là:
$P(Y > 45000) = P(Z > \frac{4500-5620}{429,26}) = P(Z > -2,61) = 1 - \Phi(-2,61) = 0,9955$

CHÚ Ý: Trường hợp $n = 2$ thì ta có kết quả sau
$V(a_1X_1 + a_2X_2) = a_1^2V(X_1) + a_2^2V(X_2) + 2a_1a_2Cov(X_1, X_2)$

# Chương 6

# ƯỚC LƯỢNG ĐIỂM

## Giới thiệu

## 6.1 Một số khái niệm tổng quát về ước lượng điểm

## 6.2 Các phương pháp ước lượng điểm

# Chương 7

# ƯỚC LƯỢNG KHOẢNG

## Giới thiệu

## 7.1 Các tính chất cơ bản của khoảng tin cậy

## 7.2 Khoảng tin cậy mẫu lớn cho trung bình tổng thể

## 7.3 Các khoảng dựa trên phân phối chuẩn

## 7.4 Khoảng tin cậy của phương sai và độ lệch chuẩn của phân phối chuẩn

# Chương 8

# KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

Giới thiệu

8.1 Giả thiết và thủ tục kiểm định

8.2 Kiểm định về trung bình tổng thể

8.3 Kiểm định về tỷ lệ

8.4 P giá trị

8.5 Một số chú ý về chọn thủ tục kiểm định

# Chương 9

# CÁC KẾT LUẬN DỰA TRÊN HAI MẪU

Giới thiệu

9.1 Tiêu chuẩn z và khoảng tin cậy cho hiệu giữa hai trung bình

9.2 Tiêu chuẩn t và khoảng tin cậy

9.3 Phân tích số liệu ghép đôi

9.4 Các kết luận liên quan đến hiệu hai tỷ lệ

9.5 Các kết luận liên quan đến hai phương sai

# Chương 10

# PHÂN TÍCH PHƯƠNG SAI

## Giới thiệu

## 10.1 Một nhân tố ANOVA

## 10.2 So sánh trong ANOVA

## 10.3 Nhiều hơn trên một nhân tố ANOVA

# Chương 11

# PHÂN TÍCH PHƯƠNG SAI NHIỀU NHÂN TỐ

**Giới thiệu**

**11.1 Hai nhân tố ANOVA với $K_{ij} = 1$**

**11.2 Hai nhân tố ANOVA với $K_{ij} > 1$**

**11.3 Ba nhân tố ANOVA**

**11.4 NHân tố $2^p$**

# Chương 12

# TƯƠNG QUAN VÀ HỒI QUI TUYẾN TÍNH

Giới thiệu

12.1 Mô hình hồi qui tuyến tính

12.2 Ước lượng tham số mô hình

12.3 Suy luận về hệ số dốc $\beta_1$

12.4 Suy luận về sự liên quan giữa $\mu_{Y,x}$ và giá trị dự đoán của $Y$

12.5 Hệ số tương quan

# TỪ KHÓA

# Tài liệu tham khảo