CHƯƠNG 5: TỔNG CÁC BIẾN NGẪU NHIÊN VÀ CÁC ĐỊNH LÝ GIỚI HẠN

Bài 1, 2, 3: Luật số lớn và Định lý giới hạn Trung tâm

Nhiều bài toán dẫn đến việc đếm số lần xảy ra các biến cố, ví dụ như phép đo các hiệu quả tích luỹ hoặc việc tính trung bình số học của dãy các kết quả đo. Thông thường các bài toán này được dẫn tới bài toán tìm phân phối chính xác hoặc xấp xỉ của các biến ngẫu nhiên được tạo bởi tổng các biến ngẫu nhiên độc lập cùng phân phối. Trong chương này, chúng ta nghiên cứu tổng của các biến ngẫu nhiên và tính chất của chúng khi n tiến ra vô cùng.

Trong phần 5.1 chúng ta sẽ chỉ ra hàm đặc trưng được sử dụng như thế nào để tính hàm mật độ xác suất của tổng các biến ngẫu nhiên độc lập.

Trong phần 5.2 chúng ta thảo luận việc sử dụng trung bình mẫu để ước lượng kỳ vọng của biến ngẫu nhiên và sử dụng tần suất tương đối để ước lượng xác suất của một sự kiện. Chúng ta giới thiệu độ đo để đánh giá sự phù hợp của các ước lượng này. Sau đó chúng ta xét luật số lớn, đó là các định lý phát biểu rằng trung bình mẫu và tần suất tương đối hội tụ tới giá trị kỳ vọng và xác suất tương ứng khi cỡ mẫu tăng lên vô hạn. Các kết quả lý thuyết này chứng minh tính nhất quán đáng chú ý giữa lý thuyết xác suất và kết quả quan trắc, và chúng nhấn mạnh rằng tần suất tương đối là sự thể hiện của xác suất.

Trong phần 5.3, chúng ta trình bày định lý giới hạn trung tâm, mà nó phát biểu rằng, với những điều kiện rất chung, hàm phân phối của tổng các biến ngẫu nhiên xấp xỉ với hàm phân phối của biến ngẫu nhiên Gauss, mặc dù biến ngẫu nhiên ban đầu khác rất xa biến ngẫu nhiên Gauss. Kết quả này cho phép chúng ta giải thích tại sao biến ngẫu nhiên xuất hiện trong nhiều ứng dụng khác nhau.

Trong phần 5.4, chúng ta thảo luận khái niệm khoảng tin cậy đóng, mà nó đóng vai trò quan trọng trong việc xác định các giá trị thực nghiệm của các tham số của biến ngẫu nhiên. Trong phần 5.5 chúng ta xét các biến ngẫu nhiên và các tính chất hội tụ. Trong phần 5.6 chúng ta thảo luận các thí nghiệm ngẫu nhiên khi mà các biến cố xảy ra tại các thời điểm ngẫu nhiên. Trong các thí nghiệm này chúng ta quan tâm đến tốc độ trung bình mà tại đó các biến cố xảy ra cũng như tốc độ mà tại đó các đại lượng kết hợp với các biến cố xảy ra. Cuối cùng, phần 5.7 giới thiệu các phương pháp tính dựa trên phép biến đổi Fourier rời rạc mà nó tỏ ra rất hữu ích khi tính hàm khối lượng xác suất và hàm mật độ xác suất từ các phép biến đổi này.

5.1 TỔNG CỦA CÁC BIẾN NGẪU NHIÊN

Giả sử $X_1, X_2, \ldots, X_n$ là dãy các biến ngẫu nhiên và giả sử S_n là tổng của chúng:

$$S_n = X_1 + X_2 + \ldots + X_n . \qquad (5.1)$$

Trong phần này, chúng ta tìm giá trị trung bình và phương sai của S_n cũng như mật độ xác suất của S_n trong trường hợp đặc biệt quan trọng khi X_j là các biến ngẫu nhiên độc lập.

Trung bình và phương sai của tổng các biến ngẫu nhiên:

Trong phần 4.7 đã chứng tỏ rằng, giá trị kỳ vọng của tổng các biến ngẫu nhiên độc lập là các biến ngẫu nhiên độc lập:

$$E[X_1 + X_2 + \ldots + X_n] = E[X_1] + \ldots + E[X_n] \quad (5.2)$$

Như vậy thông tin về giá trị trung bình của các biến ngẫu nhiên X_j đủ để xác định giá trị trung bình của S_n.

Ví dụ sau chứng tỏ rằng để tính phương sai của tổng các biến ngẫu nhiên chúng ta cần phải biết phương sai và hiệp phương sai của các biến ngẫu nhiên X_j.

VÍ DỤ 5.1 Hãy tìm phương sai của Z = X + Y

Từ hệ thức (5.2) E[Z] = E [X + Y] = E[X] + E[Y].

Do đó phương sai của Z bằng:

$$\begin{aligned}
VAR(Z) &= E[(Z - E[Z])^2] = E[(X + Y - E[X] - E[Y])^2] \\
&= E[\{(X - E[X]) + (Y - E[Y])\}^2] \\
&= E[(X - E[X])^2 + (Y - E[Y])^2 + (X - E[X])(Y - E[Y]) \\
&\quad + (Y - E[y])(X - E[X])] \\
&= VAR[X] + VAR[Y] + COV(X, Y) + COV(Y, X) \\
&= VAR[X] + VAR[Y] + 2\,COV(X,Y).
\end{aligned}$$

Nói chung hiệp phương sai COV (X,Y) không bằng 0, bởi vậy phương sai của tổng các biến ngẫu nhiên không nhất thiết bằng tổng của các phương sai.

Phương sai của các biến ngẫu nhiên :

$$\mathbf{VAR}(X_1, +X_2 + \cdots + X_n)$$

$$= E\left\{\sum_{j=1}^{n}\left(X_j - E[X_j]\right)\sum_{k=1}^{n}\left(X_k - E[X_k]\right)\right\}$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} E\left[\left(X_j - E[X_j]\right)\left(X_k - E[X_k]\right)\right]$$

$$= \sum_{k=1}^{n}\mathbf{VAR}(X_k) + \sum_{\substack{j=1 \\ j\neq k}}^{n}\sum_{k=1}^{n}\mathbf{COV}(X_j, X_k) \quad . \quad (5.3)$$

Như vậy, nói chung phương sai của tổng các biến ngẫu nhiên không bằng tổng phương sai của các biến ngẫu nhiên thành phần.

Trường hợp đặc biệt quan trọng là khi các biến X_j độc lập. Nếu $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên độc lập thì $COV(X_j, \ldots, X_k) = 0 \quad \forall\, j \neq k$ và

$$VAR(X_1 + X_2 + \ldots + X_n) = VAR(X_1) + \ldots + VAR(X_n). \qquad (5.4)$$

VÍ DỤ 5.2
Tổng các biến ngẫu nhiên độc lập

Hãy tìm giá trị trung bình và phương sai của tổng n biến ngẫu nhiên độc lập cùng phân phối (i i d), mà mỗi bên có giá trị trung bình μ và phương sai σ^2.

Giá trị trung bình của S_n nhận được từ hệ thức (5.2) :

$$E[S_n] = E[X_1] + \ldots + E[X_n] = n\mu .$$

Hiệp phương sai của cặp biến ngẫu nhiên độc lập bằng 0, như thế từ hệ thức (5.4)

$$VAR[S_n] = n\, VAR[X_j] = n\sigma^2 ,$$

do $VAR[X_j] = \sigma^2$ đối với $j = 1, \ldots, n$.

Hàm mật độ xác suất của các biến ngẫu nhiên độc lập

Giả sử $X_1 , X_2 , \ldots, X_n$ là n biến ngẫu nhiên độc lập. Trong phần này chúng ta sẽ chứng tỏ bằng phương pháp biến đổi nào đó có thể được sử dụng để tìm hàm mật độ xác suất của $S_n = X_1 + X_2 + \ldots + X_n$.

Trước hết, chúng ta xét trường hợp n = 2, Z = X + Y, khi X và Y là các biến ngẫu nhiên độc lập. Hàm đặc trưng của Z được cho bởi:

$$\begin{aligned}\Phi_Z(\omega) &= E[e^{j\omega Z}] \\ &= E[e^{j\omega(X+Y)}] \\ &= E[e^{j\omega X} e^{j\omega Y}] \\ &= E[e^{j\omega X}]E[e^{j\omega Y}] \\ &= \Phi_X(\omega)\,\Phi_Y(\omega), \qquad (5.5)\end{aligned}$$

ở đây hệ thức thứ tư sụy ra từ thực tế là hàm số của các biến ngẫu nhiên độc lập (tức là $e^{j\omega X}$, $e^{j\omega Y}$) cũng là các biến ngẫu nhiên độc lập như được xét trong ví dụ 4.40. Như vậy, hàm đặc trưng của Z là tích của các hàm đặc trưng của X và Y.

Trong ví dụ 4.31 chúng ta nhận thấy rằng hàm mật độ xác suất của Z = X + Y được cho bởi tích chập của các hàm mật độ xác suất của X và Y:

$$f_Z(z) = f_X(x) * f_Y(y). \qquad (5.6)$$

Nhắc lại rằng $\Phi_Z(\omega)$ cũng có thể được coi như là phép biến đổi **Fourier** của hàm mật độ xác suất của Z :

$$\Phi_Z(\omega) = \mathcal{F}\{f_Z(z)\}.$$

Bằng việc thay phương trình (5.6) vào phương trình (5.5) ta nhận được :

$$\Phi_Z(\omega) = \mathcal{F}\{f_Z(z)\} = \mathcal{F}\{f_X(x) * f_Y(y)\} = \Phi_X(\omega)\,\Phi_Y(\omega). \qquad (5.7)$$

Phương trình (5.7) phát biểu một kết quả nổi tiếng là biến đổi Fourier của một tích chập hai hàm bằng tích các phép biến đổi Fourier từng hàm riêng.

Bây giờ ta xét tổng n biến ngẫu nhiên độc lập :

$$S_n = X_1 + X_2 + \ldots + X_n$$

Hàm dặc trưng của S_n là

$$\Phi_{S_n}(\omega) = E\left[e^{j\omega S_n}\right] = E\left[e^{j\omega(X_1+X_2+\ldots+X_n)}\right]$$
$$= E\left[e^{j\omega X_i}\right]\cdots E\left[e^{j\omega X_n}\right]$$
$$= \Phi_{X_1}(\omega)\cdots\Phi_{X_n}(\omega). \qquad (5.8)$$

Như vậy hàm mật độ xác suất của S_n có thể tìm được nhờ tìm phép đổi Fourier ngược của tích các hàm đặc trưng thành phần của các X_j.

$$f_{S_n}(X) = \Im^{-1}\left\{\Phi_{X_1}(\omega)\cdots\Phi_{X_n}(\omega)\right\}. \qquad (5.9).$$

VÍ DỤ 5.3
Tổng của các biến ngẫu nhiên Gauss độc lập

Giả sử S_n là tổng n biến ngẫu nhiên Gauss độc lập với kỳ vọng và phương sai tương ứng là $m_1, \ldots, m_n$ và $\sigma_1^2, \ldots, \sigma_n^2$. Hãy tìm hàm mật độ xác suất của S_n .

Hàm đặc trưng của X_k là :

$$\Phi_{X_k}(\omega) = e^{+j\omega m_k - \omega^2\sigma_k^2/2}$$

và khi đó hệ thức (5.8),

$$\Phi_{S_n}(\omega) = \prod_{k=1}^{n} e^{+j\omega m_k - \omega^2\sigma_k{}^2/2}$$
$$= e^{+j\omega(m_1+\cdots+m_n) - \omega^2(\sigma_1^2+\cdots+\sigma_n^2)/2}$$

Đây là hàm đặc trưng của một biến ngẫu nhiên Gauss. Như vậy S_n là biến ngẫu nhiên Gauss với trung bình $m_1 + m_2 + \ldots + m_n$ và phương sai $\sigma_1^2 + \ldots + \sigma_n^2$.

VÍ DỤ 5.4
Tổng của các biến ngẫu nhiên độc lập

Hãy tìm hàm mật độ xác suất của tổng n biến ngẫu nhiên độc lập cùng phân phối với các hàm đặc trưng :

$\Phi_{X_k}(\omega) = \Phi_X(\omega)$ với k = 1, . . . , n.

Từ hệ thức của hàm đặc trưng của tổng các biến ngẫu nhiên độc lập :

$$\Phi_{S_n}(\omega) = \left\{\Phi_X(\omega)\right\}^n. \qquad (5.10)$$

VÍ DỤ 5.5
Tổng của các biến ngẫu nhiên mũ độc lập

Hãy tìm hàm mật độ xác suất của tổng n biến ngẫu nhiên có phân phối mũ độc lập tất cả có tham số α.

Hàm đặc trưng của một biến ngẫu nhiên mũ là :

$$\Phi_X(\omega) = \frac{\alpha}{\alpha - j\omega}.$$

Từ ví dụ trên chúng ta có

$$\Phi_{S_n}(\omega) = \left\{\frac{\alpha}{\alpha - j\omega}\right\}^n$$

Từ bảng 3.2, chúng ta có thể nhận thấy rằng S_n là một biến ngẫu nhiên *m* – Erlang.

Khi làm việc với biến ngẫu nhiên nhận giá trị nguyên để thuận lợi chúng ta làm việc với hàm sinh xác suất.

$G_N(z) = E[z^N]$.

Hàm sinh của tổng các biến ngẫu nhiên rời rạc độc lập,

$N = X_1 + X_2 + \ldots + X_n$, là :

$$G_N(z) = E\left[z^{X_1 + X_2 + \ldots + X_n}\right] = E\left[z^{X_1}\right]\ldots E\left[z^{X_n}\right]$$
$$= G_{X_1}(z)\ldots G_{X_n}(z). \qquad (5.11)$$

VÍ DỤ 5.6

Hãy tìm hàm sinh của tổng *n* biến ngẫu nhiên có phân phối hình học có cùng giá trị p và độc lập.

Hàm sinh của biến ngẫu nhiên hình học riêng lẻ được cho bởi :

$$G_X(z) = \frac{pz}{1 - pz}.$$

Do đó hàm sinh của tổng của *n* biến ngẫu nhiên độc lập như vậy là :

$G_N(z) =$ **Error!**.

Từ bảng 3.1, chúng ta thấy rằng đây là hàm sinh của một biến ngẫu nhiên nhị thức âm với tham số *p* và *n*.

***Tổng của một số ngẫu nhiên các biến ngẫu nhiên**

Trong một số bài toán chúng ta quan tâm đến tổng của một số ngẫu nhiên N các biến ngẫu nhiên độc lập cùng phân phối:

$$S_N = \sum_{k=1}^{N} X_k. \qquad (5.12)$$

ở đây N là biến ngẫu nhiên độc lập với X_k' *s*. Ví dụ N là số các thao tác máy tính được thực hiện một giờ và X_k là thời gian cần thiết để thao tác thứ *k* :

Giá trị trung bình của S_N tìm được bằng việc sử dụng kỳ vọng có điều kiện :

$$E[S_N] = E[E[S_N \mid N]]$$
$$= E[NE[X]]$$
$$= E[N]E[X]. \qquad (5.13)$$

Hệ thức thứ 2 suy ra từ hệ thức sau

$$E[S_N \mid N = n] = E\left[\sum_{k=1}^{N} X_k\right] = nE[X],$$

Bởi vậy $E[S_N \mid N] = NE[X]$.

Hàm đặc trưng của S_N cũng có thể tìm được bằng việc sử dụng kỳ vọng có điều kiện. Từ hệ thức (5.10), chúng ta có:

$$E\left[e^{j\omega S_N} \mid N = n\right] = E\left[e^{j\omega(X_1 + \cdots + X_n)}\right] = \Phi_X(\omega)^n,$$

Bởi vậy

$$E\left[e^{j\omega S_N} \mid N\right] = \Phi_X(\omega)^N.$$

Do đó

$$\Phi_{S_N}(\omega) = E\left[E\left[e^{j\omega S_N} \mid N\right]\right]$$
$$= E[\Phi_X(\omega)^N]$$
$$= E[z^N] \mid_{z = \Phi X(\omega)}$$
$$= G_N(\Phi_X(\omega)). \qquad (5.14)$$

Nghĩa là hàm đặc trưng của S_N tìm được bằng cách tính hàm sinh của N tại $z = \Phi_X(\omega)$.

VÍ DỤ 5.7 Số N thao tác mà máy tính thực hiện trong 1 giờ là biến ngẫu nhiên hình học với tham số p và thời gian thực hiện thao tác là các biến ngẫu nhiên mũ độc lập với trung bình $1/\alpha$. Hãy tìm hàm xác suất của tổng thời gian thực hiện thao tác trong một giờ.

Hàm sinh cho N là:

$$G_N(z) = \frac{p}{1-qz},$$

và hàm đặc trưng của một biến ngẫu nhiên có phân phối mũ là :

$$\Phi_X(\omega) = \frac{\alpha}{\alpha - j\omega}.$$

Từ hệ thức (5.14), hàm đặc trưng của S_N là:

$$\Phi_{S_N}(\omega) = \frac{p}{1 - q[\alpha/(\alpha - j\omega)]}$$
$$= p(\alpha - j\omega) / (p\alpha - j\omega)$$
$$= p + (1-p)\frac{p\alpha}{p\alpha - j\omega}.$$

Hàm mật độ xác suất của S_N tìm được bởi việc thực hiện phép biến đổi ngược của biểu thức trên:

$$f_{S_N}(x) = p\delta(x) + (1-p)p\alpha e^{-p\alpha x} \qquad x \geq 0.$$

Hàm mật độ xác suất có sự giải thích trực tiếp như sau với xác suất p không có một yêu cầu nào và do vậy tổng thời gian thực hiện là bằng 0; với xác suất (1–p) có một hoặc hơn một yêu cầu đến, và tổng thời gian thực hiện là biến ngẫu nhiên có phân phối mũ với trung bình là $1/p\alpha$.

5.2 TRUNG BÌNH MẪU VÀ LUẬT SỐ LỚN

Giả sử X là biến ngẫu nhiên với giá trị trung bình, E[X] = μ là chưa biết. Giả sử $X_1, \ldots,$ X_n ký hiệu n phép đo độc lập đại lượng X, nghĩa là X_j là các biến ngẫu nhiên có độc lập có cùng phân phối với cùng hàm phân phối như X. Trung bình mẫu của dãy được sử dụng để ước lượng E[X]:

$$M_n = \frac{1}{n}\sum_{j=1}^{n} X_j .$$

(5.15)

Trong phần này chúng ta tính giá trị kỳ vọng và phương sai của M_n để đánh giá hiệu quả của M_n với tư cách là ước lượng của E[X]. Chúng ta cũng nghiên cứu dáng điệu của M_n khi n tiến ra vô cùng.

Ví dụ sau đây chứng tỏ rằng tần số tương đối để ước lượng xác suất của một biến cố là trường hợp riêng của trung bình mẫu. Như vậy các kết quả được suy ra sau đây cho trung bình mẫu cũng được áp dụng cho ước lượng tần số tương đối.

VÍ DỤ 5.8

Tần số tương đối

Xét dãy lặp lại thí nghiệm ngẫu nhiên độc lập nào đó và đặt biến ngẫu nhiên Ij là hàm chỉ số để chỉ sự xảy ra biến cố A trong phép thử thứ j. Khi đó tổng số lần xảy ra biến cố A trong n phép thử là Bernoulli là :

$N_n = I_1 + I_2 + \ldots + I_n$.

Tần suất tương đối của biến cố A trong n lặp lặp lại thí nghiệm đầu tiên là:

$$f_A(n) = \frac{1}{n}\sum_{j=1}^{n} I_j . \qquad (5.16)$$

Như vậy tần số tương đối $f_A(n)$ là trung bình mẫu của biến ngẫu nhiên I_j.

Trung bình mẫu bản thân nó cũng là một biến ngẫu nhiên, bởi vậy nó sẽ thể hiện biến thiên một cách ngẫu nhiên. Một ước lượng tốt cần phải có hai tính chất sau: (1) Kỳ vọng của nó đúng bằng giá trị cần ước lượng nghĩa là $E[M_n] = \mu$; (2) Dao động của nó cần phải không quá lớn xung quanh giá trị đúng của tham số cần ước lượng, nghĩa là, $E[(M_n - \mu)^2]$ là nhỏ.

Giá trị kỳ vọng của trung bình mẫu được cho bởi:

$$E[M_n] = E\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} E[X_j] = \mu \qquad (5.17)$$

do $E[X_j] = E[X] = \mu$ với $\forall$ j. Như vậy trung bình mẫu bằng E[X] = μ về giá trị trung bình. Vì lý do này, chúng ta nói rằng trung bình mẫu là ước lượng không chệch cho μ.

Hệ thức (5.17) suy ra rằng sai số trung bình bình phương của trung bình mẫu xung quanh μ là bằng phương sai của M_n, nghĩa là,

$E[(M_n - \mu)^2] = E[(M_n - E[M_n])^2]$.

Chú ý rằng $M_n = S_n/n$ trong đó $S_n = X_1 + X_2 + \ldots + X_n$. Từ hệ thức (5.4), $VAR[S_n] = n\,VAR[X_j] = n\sigma^2$, do X_j là các biến ngẫu nhiên độc lập cùng phân phối. Như vậy,

$$\mathbf{VAR}[M_n] = \frac{1}{n^2}\mathbf{VAR}[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \qquad (5.18)$$

Hệ thức (5.18) phát biểu rằng phương sai của trung bình mẫu tiến dần đến 0 khi cỡ mẫu dần ra vô hạn. Điều này suy ra rằng xác suất để trung bình mẫu tiến dần đến giá trị trung bình thực là xấp xỉ 1 khi n tiến ra vô cùng. Chúng ta có thể chứng minh khẳng định này bằng việc sử dụng bất đẳng thức Chebyshev, hệ thức (3.73):

$$P\{|M_n - E[M_n]| \geq \varepsilon\} \leq \frac{VAR[M_n]}{\varepsilon^2}$$

Thay giá trị của $E[M_n]$ và $VAR[M_n]$ vào công thức, chúng ta nhận được:

$$P\{|M_n - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

Nếu chúng ta xét biến cố đối của biến cố xét ở (5.19), chúng ta nhận được:

$$P[|M_n - \mu| < \varepsilon] \geq 1 - \frac{\sigma^2}{n\varepsilon^2}. \qquad (5.20)$$

VÍ DỤ 5.9 Điện áp là hằng số nhưng chúng ta chưa biết, chúng ta tiến hành đo điện áp. Kết quả mỗi một phép đo X_j thực sự là tổng của điện áp mong muốn *v* và điện áp nhiễu N_j có kỳ vọng 0 và độ lệch chuẩn 1 micro vôn (μV):

$X_j = v + N_j$.

giả sử rằng các điện áp nhiễu là những biến ngẫu nhiên độc lập. Đòi hỏi phải có bao nhiêu độ đo sao cho xác suất để M_n sai khác với kỳ vọng đúng trong vòng bán kính ε = 1 μV tối thiểu bằng 0.99 ?

Mỗi độ đo X_j có kỳ vọng 0 và phương sai 1. Bởi vậy từ phương trình (5.20) chúng ta đòi hỏi n phải thoả mãn

$$1 - \frac{\sigma^2}{n\varepsilon^2} = 1 - \frac{1}{n} = .99.$$

Từ đó suy ra n = 100.

Như vậy chúng ta lặp đi lặp lại đo 100 lần và tính trung bình mẫu về điện áp ít nhất 99 lần rút ra từ 100 lần, kết quả mẫu sẽ sai khác 1 μV so với kỳ vọng đúng.

Chú ý rằng chúng ta cho $n \to \infty$ từ phương trình (5.20) chúng ta nhận được

$$\lim_{n\to\infty} P[|M_n - \mu| < \varepsilon] = 1.$$

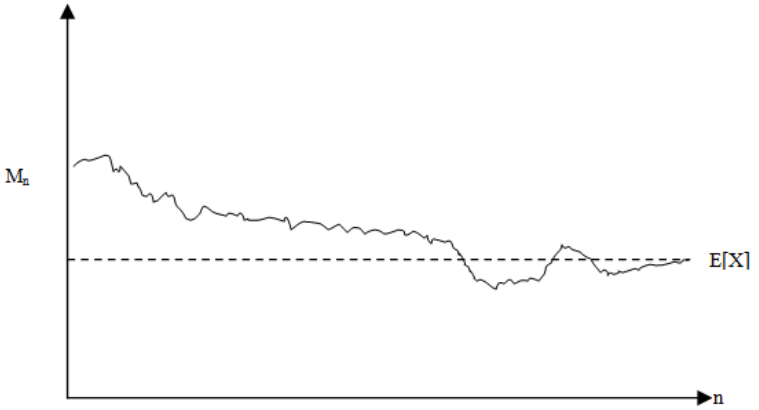
Phương trình (5.20) đòi hỏi các X_j có phương sai hữu hạn. Có thể chứng tỏ rằng giới hạn này ($\lim_{n\to\infty} P[|M_n - \mu|] = 1$) đúng ngay cả khi phương sai của các X_j không tồn tại (Gnedenko, 1976, 203). Chúng ta phát biểu kết quả này tổng quát hơn

LUẬT SỐ LỚN YẾU

Cho $X_1, X_2 \ldots$ là dãy các biến ngẫu nhiên phân phối bằng nhau độc lập với kỳ vọng hữu hạn $E[X] = \mu$, đối với $\varepsilon > 0$ cho trước thì,

$$\lim_{n\to\infty} P[|M_n - \mu| < \varepsilon] = 1. \qquad (5.21)$$

HÌNH 5.1
Sự hội tụ các trung bình mẫu tới E[X].



Luật số lớn yếu [illegible] inh mẫu khi dùng n mẫu sẽ gần với kỳ vọng đúng với xác suất cao. Luật số lớn yếu không nhằm vào câu hỏi điều gì xảy ra đối với trung bình mẫu là một hàm của n khi chúng ta bổ sung vào các giá trị đó. Câu hỏi này tiếp tục đưa ra bởi luật số lớn mạnh mà chúng ta sẽ thảo luận sau.

Giả sử tạo ra một dãy các độ độc lập của cùng một biến ngẫu nhiên. Cho $X_1, X_2 \ldots$là dãy kết quả các biến ngẫu nhiên như nhau, độc lập với kỳ vọng μ. Bây giờ xét dãy các trung bình mẫu là kết quả từ các độ đo trên: $M_1, M_2, \ldots$, trong đó M_j là trung bình mẫu tính khi được dùng X_1 đến X_j . Khái niệm tính đều thống kê đã thảo luận trong chương 1 đưa chúng ta hy vọng rằng dãy trung bình mẫu này hội tụ tới μ, có nghĩa là chúng ta hy vọng rằng với xác suất cao, mỗi dãy thực nghiệm của trung bình mẫu tiến gần tới μ và dừng tại đó như được thể hiện ở hình 5.1. Theo thuật ngữ xác suất, chúng ta hy vọng điều sau đây:

$$P\left[\lim_{n\to\infty} M_n = \mu\right] = 1;$$

có nghĩa là với một độ tin cậy thực, mỗi dãy trung bình mẫu tính toán hội tụ về kỳ vọng đúng của đại lượng quan tâm. Chứng minh kết quả này sẽ vượt quá cấp độ đặt ra của giáo trình này.(xem Gnedenko, 1976,216). Nhưng chúng ta sẽ có dịp trong các phần sau áp dụng kết quả trong các tình huống khác nhau.

LUẬT MẠNH SỐ LƠN

Cho X_1, X_2,… là một dãy các biến ngẫu nhiên cùng phân phối, độc lập, với kỳ vọng hữu hạn E[X] = μ và phương sai hữu hạn, thì

$$P\left[\lim_{n\to\infty} M_n = \mu\right] = 1;. \qquad (5.22)$$

Phương trình (5.22) xuất hiện tượng đương với phương trình (5.21) nhưng thực sự nó tạo nên một phát biểu khác biệt một cách …… .Nó tuyên bố rằng với xác suất 1, mỗi dãy tính toán các trung bình mẫu rốt cuộc dần tới và ở lại gần với E[X] = μ. Đây là một kiểu hội tụ, chúng ta hy vọng trong những điều kiện vật lý mà ở đó tính đều của thống kê được duy trì.

Với luật số lớn mạnh chúng ta có đầy đủ khả năng trong việc mô hình hoá tiến trình. Chúng ta bắt đầu bởi việc chú rằng tính đều thống kê quan trắc được trong nhiều hiện tượng vật lý và tự việc này chúng ta suy luận ra một số tính chất của tần suất tương đối. Những tính chất này được dùng để thành lập một bộ các tiên đề mà từ đó chúng ta đã phát triển ra lý thuyết toán học của xác suất. Bây giờ chúng ta đã có đủ hiểu biết và được sáng tỏ rằng dưới những điều kiện xác định, lý thuyết dự báo được hội tụ của trung bình mẫu tới giá trị kỳ vọng. Vẫn còn có lỗ hổng giữa lý thuyết toán học và thế giới thực (tức là không bao giờ chúng ta thực sự tiến hành một số vô hạn các lần đo và tính một số vô hạn các trung bình mẫu). Do đó luật số lớn mạnh trình bày một sự nhất quán đáng chú ý giữa lý thuyết và cách hành xử vật lý quan trắc được.

Chúng ta đã được chỉ cho thấy rằng là tần suất tương đối là trường hợp đặc biệt của trung bình mẫu. Nếu chúng ta áp dụng luật số lớn yếu cho tần suất tương đối của biến cố A, $f_A(n)$, trong một dãy thực hiện lặp lại một cách độc lập một thí nghiệm ngẫu nhiên, chúng ta nhận được

$$\lim_{n\to\infty}\left[\,|\, f_A(n) - P[A]| < \varepsilon\right] = 1 \qquad (5.23)$$

Nếu chúng ta áp dụng luật mạnh số lớn thì

$$P\left[\lim_{n\to\infty} f_A(n) = P[A]\right] = 1. \qquad (5.24)$$

VÍ DỤ 5.10 Để ước lượng xác suất của biến cố A, một dãy phép thử Bernoulli được tiến hành và tần suất tương đối của A là quan trắc được. *n* cần lớn bằng bao nhiêu để tần suất tương đối sai khác với p = P[A] bé hơn 0.01 có xác suất 0.95.

Giả sử X = I_A là hàm số chỉ báo của A. Từ bảng 3.1 chúng ta có kỳ vọng của I_A là μ = p và phương sai là $\sigma^2 = p(1 - p)$. Vì p chưa biết nên σ^2 cũng chưa biết. Tuy nhiên dễ dàng chứng minh rằng p(1 – p) lớn nhất là 1/4 đối với mọi p sao cho $0 \le p \le 1$. Do đó, theo phương trình (5.19)

$$P\left\{|\, f_A(n) - p \,| \ge \varepsilon\right\} \le \frac{\sigma^2}{\varepsilon^2} \le \frac{1}{4n\varepsilon^2}$$

Độ chính xác mong muốn là ε = 0.01 và xác suất mong muốn là

$$1-.95=\frac{1}{4n\varepsilon^2}.$$

Chúng ta giải phương trình cho *n*, chúng ta nhận được n = 50000. Chứng tỏ rằng Chebyshev đưa ra các biên rất lỏng, bởi vậy chúng ta hy vọng rằng giá trị này đối với *n* chắc có lẽ khá bảo toàn. Trong phần tiếp theo chúng ta trình bày ước lượng tốt hơn cho giá trị *n* theo yêu cầu.

5.3 ĐỊNH LÝ GIỚI HẠN TRUNG TÂM

Giả sử $X_1, X_2, \ldots$ là dãy các biến ngẫu nhiên phân phối bằng nhau độc lập với kỳ vọng hữu hạn μ và phương sai hữu hạn σ^2, và giả sử S_n là tổng *n* biến ngẫu nhiên trên trong dãy :

$$S_n = X_1 + X_2 + \cdots + X_n .$$

(5.25)

Trong Phần 5.1, chúng ta đã phát triển các phương pháp để xác định các hàm mật độ xác suất chính xác của S_n. Bây giờ chúng ta trình bày định lý giới hạn trung tâm phát biểu rằng khi *n* trở nên lớn, hàm phân phối của S_n đã chuẩn hoá hoàn toàn tiến gần tới hàm phân phối của biến ngẫu nhiên Gauss. Điều này tạo khả năng cho chúng ta xấp xỉ hàm phân phối của S_n vô hàm tích luỹ của biến ngẫu nhiên Gauss.

Định lý giới hạn trung tâm giải thích tại sao biến ngẫu nhiên Gauss xuất hiện trong nhiều ứng dụng đến như vậy. Trong tự nhiên, nhiều hiện tượng vĩ mô rút ra từ các quá trình vi mô độc lập. Điều đó phát sinh ra biến ngẫu nhiên Gauss. Trong nhiều vấn đề nhân tạo, chúng ta quan tâm về các giá trị trung bình của tổng các biến ngẫu nhiên độc lập. Một lần nữa lại phát sinh ra biến ngẫu nhiên Gauss.

Từ ví dụ 5.2, chúng ta biết rằng nếu các X_j là phân phối bằng nhau, độc lập thì S_n có kỳ vọng *n* μ và phương sai $n\sigma^2$. Định lý giới hạn trung tâm phát biểu rằng hàm phân phối luỹ tiến của bản sao chuẩn hoá phù hợp của S_n tiến gần tới hàm phân phối của một biến ngẫu nhiên Gauss.

Định lý giới hạn trung tâm

Giả sử S_n là tổng của *n* biến ngẫu nhiên độc lập cùng phân phối với kỳ vọng hữu hạn $E[X] = \mu$ và phương sai hữu hạn σ^2, và giả sử Z_n là biến ngẫu nhiên phương sai 1, kỳ vọng không được định nghĩa bởi

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

(5.26)

thì :

$$\lim_{n\to\infty} P[Z_n \le z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx .$$

(5.27)

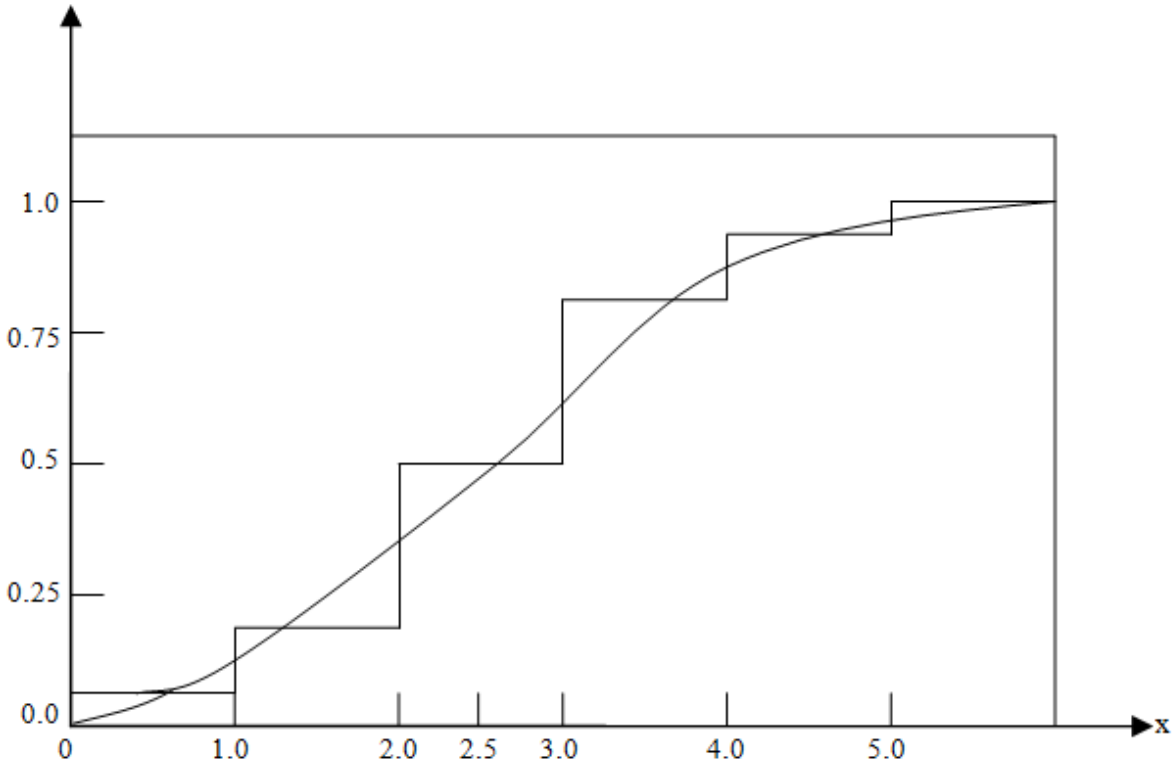
Phần đặc biệt nhất của định lý giới hạn trung tâm là các số hạng X_j có thể có phân phối bất kỳ miễn là chúng có kỳ vọng hữu hạn và phương sai hữu hạn. Điều đó đưa ra ứng dụng rộng rãi các kết quả của nó.

Phần đặc biệt nhất của định lý giới hạn trung tâm là các số lượng X_j có thể có phân phối bất kỳ, miễn là chúng có kỳ vọng hữu hạn và phương sai hữu hạn. Điều đó đưa ra ứng dụng rộng rãi các kết quả của nó.

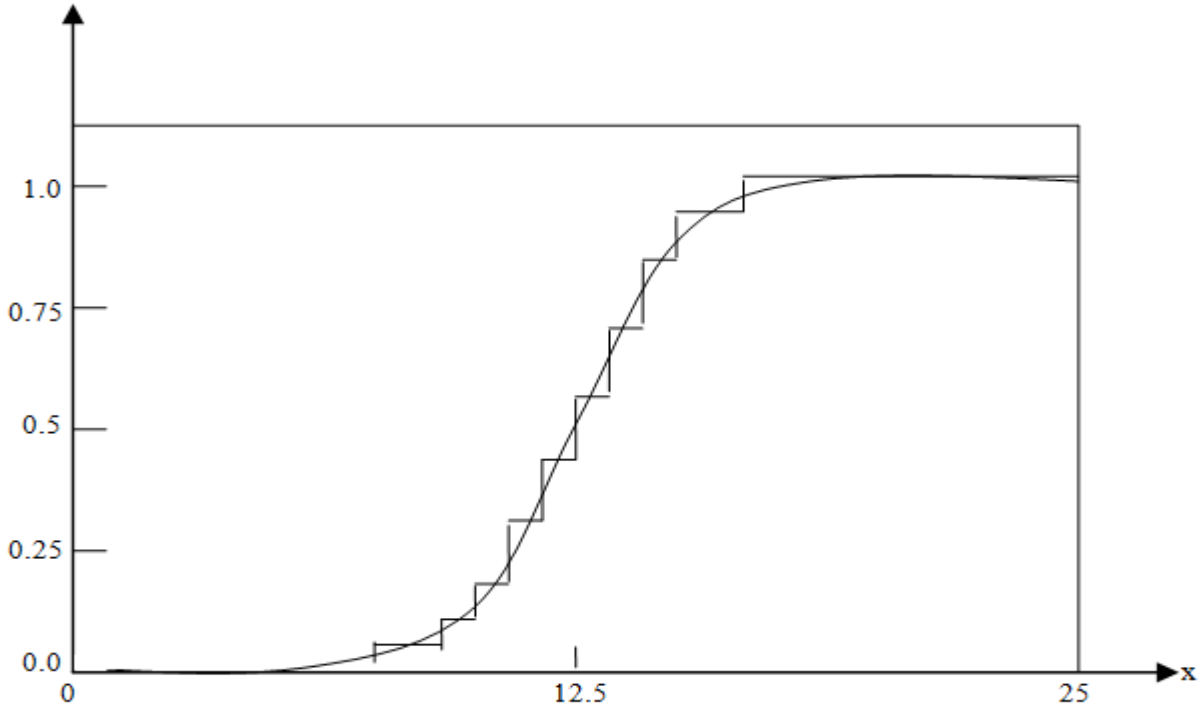
Các hình 5.2 đến 5.4 so sánh hàm phân phối và xấp xỉ Gauss cho tổng các biến ngẫu nhiên phân phối mũ, đều, Bernoulli tương ứng. Trong tất cả 3 trường hợp, có thể thấy rằng, xấp xỉ được cải thiện khi số các số hạng trong tổng tăng lên. Chứng minh định lý giới hạn trung tâm sẽ được đề cập tới trong phân đoạn cuối của phần này.

HÌNH 5.2a

Hàm phân phối của tổng 5 biến ngẫu nhiên Bernoulli độc lập với $\rho = 1/2$ và hàm phân phối của biến ngẫu nhiên Gauss cùng kỳ vọng và phương sai.

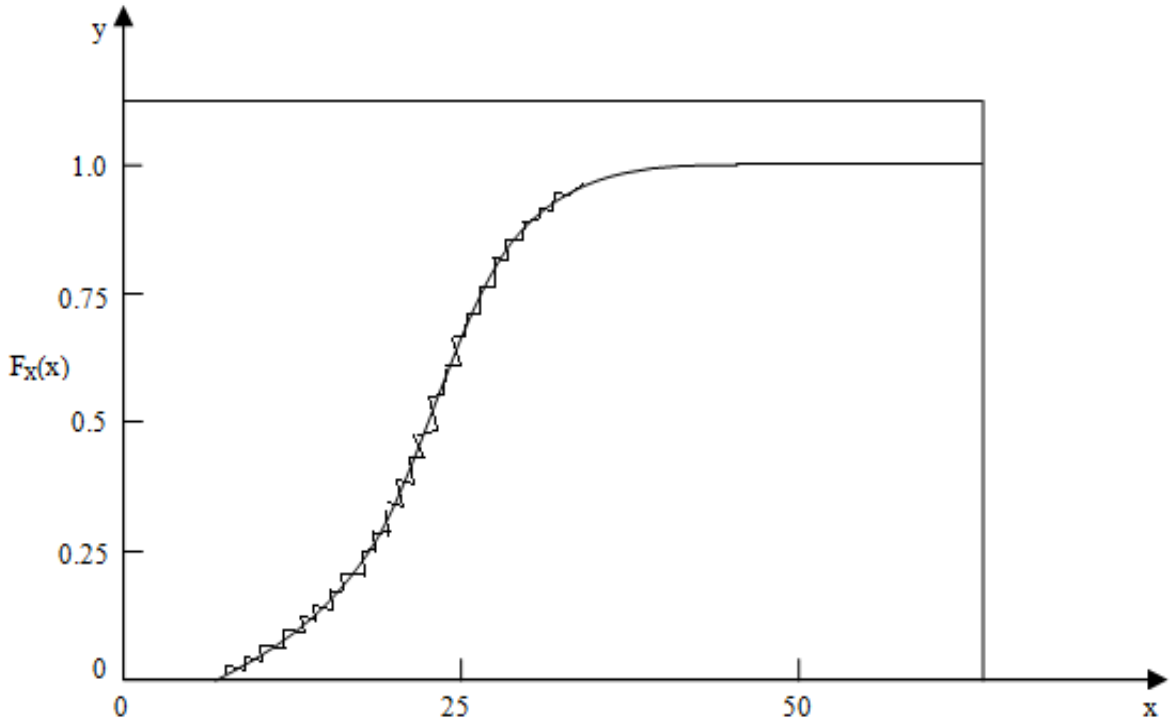


HÌNH 5.2b
Hàm phân phối của tổng 25 biến ngẫu nhiên Bernoulli độc lập với $\rho = 1/2$ và hàm phân phối của biến ngẫu nhiên Gauss cùng kỳ vọng và phương sai.



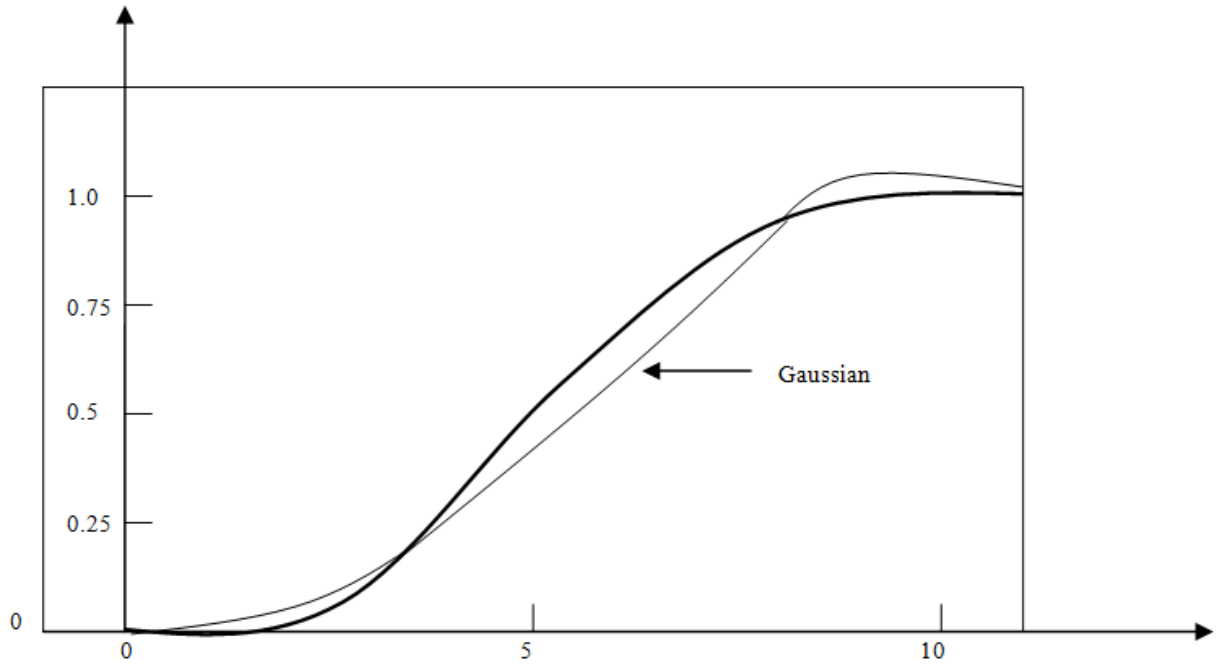
HÌNH 5.3

Hàm phân phối của tổng 5 biến ngẫu nhiên đồng đều rời rạc, độc lập từ một tập {0, 1, 2, …, 9} và hàm phân phối của một biến ngẫu nhiên Gauss cùng kỳ vọng và phương sai



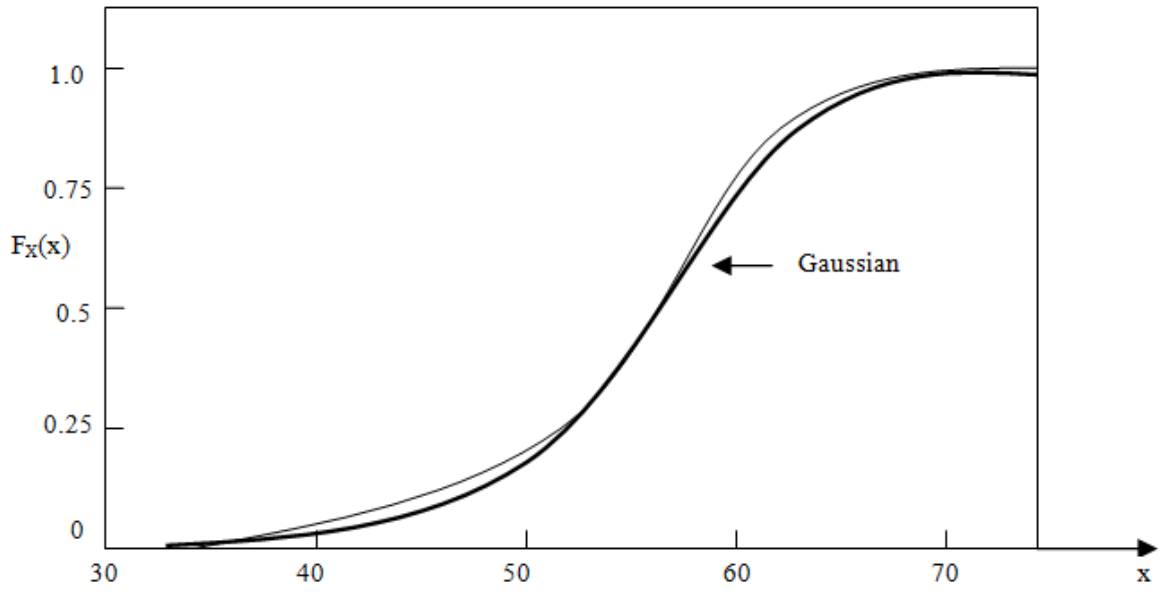
HÌNH 5.4a

Hàm phân phối của tổng 5 biến ngẫu nhiên phân phối mũ kỳ vọng 1 và hàm phân phối của một biến ngẫu nhiên Gauss cùng kỳ vọng và phương sai.



HÌNH 5.4b

Hàm phân phối của tổng 50 biến ngẫu nhiên mũ độc lập có kỳ vọng 1 và hàm phân phối của một biến ngẫu nhiên Gauss cùng kỳ vọng và phương sai.



VÍ DỤ 5.11. Giả sử rằng các đơn đặt hàng tại một nhà hàng là các biến ngẫu nhiên có độc lập cùng phân phối với kỳ vọng μ = \$8 và độ lệch chuẩn σ = \$2. Ước lượng xác suất mà 100 khách hàng đầu tiên tiêu tổng cộng số tiền lớn hơn \$840. Ước lượng xác suất mà 100 khách hàng đầu tiên chi tiêu tổng cộng số tiền nằm giữa \$780 và \$820.

Giả sử X_k ký hiệu số tiền chi tiêu của khách hàng thứ k thì số tiền chi tiêu tổng cộng của 100 khách hàng là :

$S_{100} = X_1 + X_2 + ... + X_{100}$.

Kỳ vọng của S_{100} là $n\mu = 800$ và phương sai của S_{100} là $n\sigma^2 = 400$. Hình 5.5 cho thấy hàm phân phối tập trung khá cao xung quanh kỳ vọng. Dạng chuẩn tắc của S_{100} là

$$Z_{100} = \frac{S_{100} - 800}{20}$$

Như vậy :

$$P[S_{100} > 840] = P\left[Z_{100} > \frac{840-800}{20}\right]$$

$$\approx Q(2) = 2.28(10^{-2}).$$

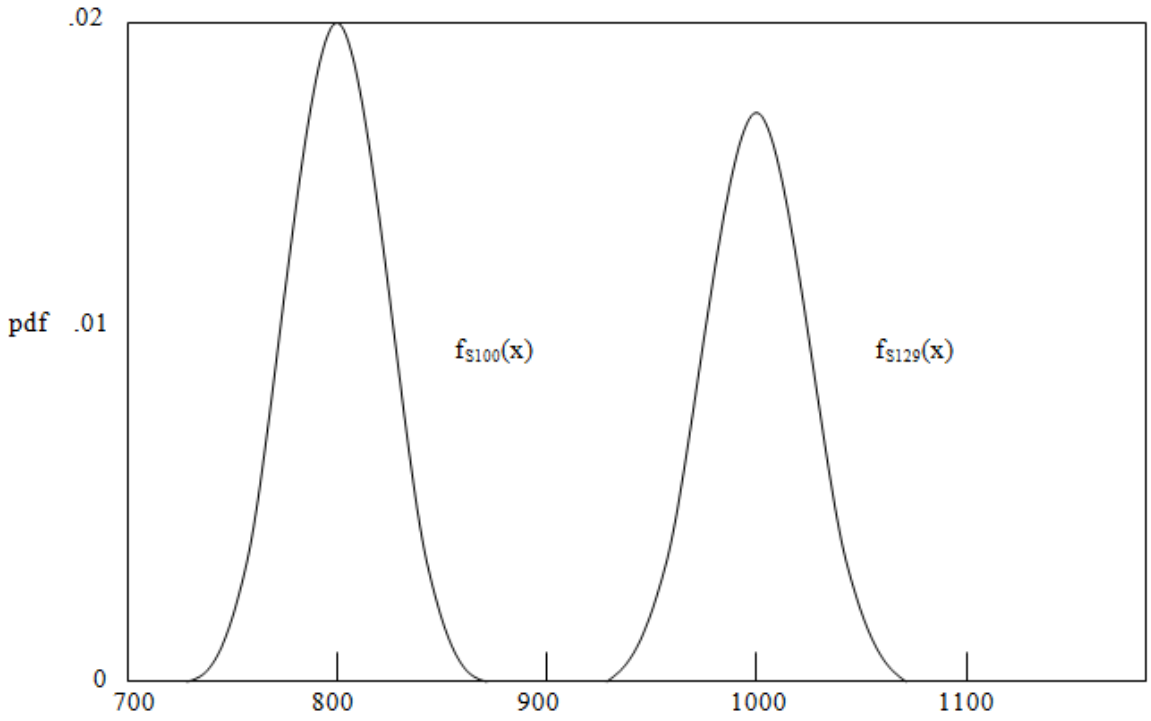
ở đây chúng ta đã sử dụng Bảng 3.3 để tính Q(2). Tương tự :

$$P[780 \le S_{100} \le 820] = P[-1 \le Z_{100} \le 1]$$

$$\approx 1 - 2Q(1)$$

$$= .682.$$

HÌNH 5.5
Các xấp xỉ hàm mật độ xác suất Gauss S_{100} và S_{129} trong các ví dụ 5.11 và 5.12.



VÍ DỤ 5.12. Trong ví dụ 5.11, sau bao nhiêu đơn đặt hàng, thì chúng ta có thể tin tưởng 90% rằng chi tiêu toàn bộ của tất cả các khách hàng lớn hơn $1000?

Vấn đề ở đây là tìm ra giá trị của n sao cho

$$P[S_n > 1000] = .90$$

S_n có kỳ vọng 8n và phương sai 4n. Tiếp tục như trong ví dụ trước chúng ta có :

$$P[S_n > 1000] = P\left[Z_n > \frac{1000-8n}{2\sqrt{n}}\right] = .90$$

Sử dụng công thức Q(–x) = 1 – Q(x), Bảng 3.4 cho thấy rằng n phải thoả mãn

$$\frac{1000-8n}{2\sqrt{n}} = -1.2815$$

đẳng thức tạo ra phương trình bậc hai đối với $\sqrt{n}$:

$8n - 1.2815(2)\sqrt{n} - 1000 = 0.$

Nghiệm dương của phương trình là $\sqrt{n} = 11{,}34$ hay $n = 128{,}6$. Hình 5.5 biểu diễn mật độ xác suất cho S_{129}.

VÍ DỤ 5.13. Thời gian giữa các biến cố trong một phép thử ngẫu nhiên nào đó là các biến ngẫu nhiên mũ có phân phối bằng nhau và độc lập với kỳ vọng m giây. Tìm xác suất để biến cố thứ 1000 xảy ra trong khoảng thời gian (1000± 50)m.

Giả sử X_j là thời gian giữa các biến cố và giả sử S_n là thời gian của biến cố thứ n thì S_n được cho bởi phương trình (5.25). Từ Bảng 3.2, kỳ vọng và phượng sai của X_j được cho bởi $E[X_j] = m$ và $VAR[X_j] = m^2$. Kỳ vọng và phương sai của S_n là $E[S_n] = nE[X_j] = nm$ và $VAR[S_n] = nVAR[X_j] = nm^2$. Từ Định lý giới hạn trung bình có

P[950m ≤ S_{1000} ≤ 1050m]

$$= P\left[\frac{950m - 1000m}{m\sqrt{1000}} \le Z_n \le \frac{1050m - 1000m}{m\sqrt{1000}}\right]$$

≈ Q(1.58) – Q(–1.58)

= 1 – 2Q(–1.58)

= 1 – 2(0.0567) = 0.8866

Như vậy, khi n lớn lên, S_n sẽ rất gần với kỳ vọng của nó là nm. Do đó chúng ta có thể phỏng đoán rằng tỷ suất trung bình một số lớn các số hạng (long-term) mà các biến cố xảy ra là :

$$\frac{n\,bc}{S_n\,s} = \frac{n}{nm} = \frac{1}{m} \text{ (biến cố/giây).} \qquad (5.28))$$

Tính toán các tỷ suất về sự xuất hiện của biến cố và các giá trị trung bình liên quan được bàn tới trong phần 5.6.

Xấp xỉ Gauss cho các xác suất nhị thức

Chúng ta đã tìm thấy trong chương hai là biến ngẫu nhiên nhị thức khó tính trực tiếp đối với n lớn vì cần phải tính các số hạng dưới dạng giai thừa. Một ứng dụng quan trọng trong thực hành của định lý giới hạn trung tâm là trong việc lấy xấp xỉ các xác suất nhị thức. Vì biến ngẫu nhiên nhị thức là tổng các biến ngẫu nhiên Bernoulli độc lập cùng phân phối (có kỳ vọng và phương sai hữu hạn). Hàm phân phối của nó dần tiến tới hàm phân phối của biến ngẫu nhiên Gauss. Giả sử X là biến ngẫu nhiên nhị thức với phương sai np và kỳ vọng là np(1 – p) và giả sử Y là biến ngẫu nhiên Gauss với cùng kỳ vọng và phương sai, thì theo định lý giới hạn trung tâm với n lớn xác suất X = k xấp xỉ bằng tích phân hàm mật độ xác suất Gauss trong khoảng độ dài đơn vị chứa k như được biểu diễn trong hình 5.6 :

P[X = k] ≈ P[k – 1/2 < Y < k + 1/2]

$$= = \frac{1}{\sqrt{2\pi np(1-p)}} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} e^{-(x-np)^2/2np(1-p)}\,dx \qquad (5.29)$$

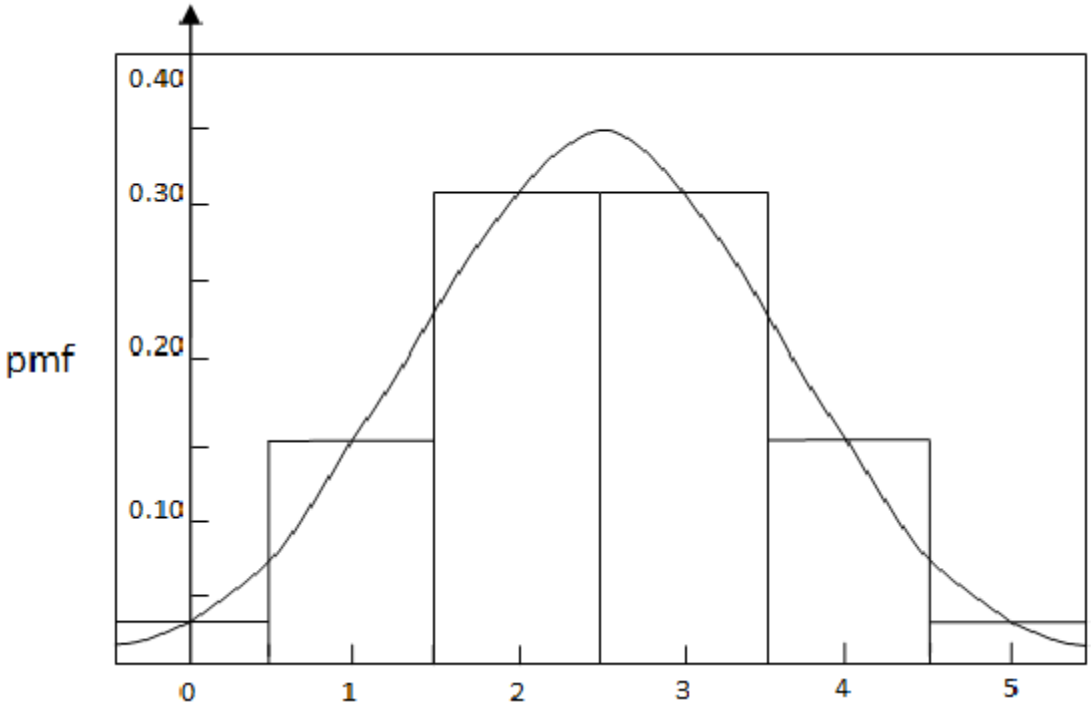
Phép xấp xỉ trên có thể được giản lược nhờ xấp xỉ tích phân bởi tích hàm dưới dấu tích phân tại tâm khoảng lấy tích phân (có nghĩa là x = k) và độ dài khoảng lấy tích phân bằng 1 :

$$P\left[X = k\right] \approx \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-(k-np)^2/2np(1-p)} \qquad (5.30)$$

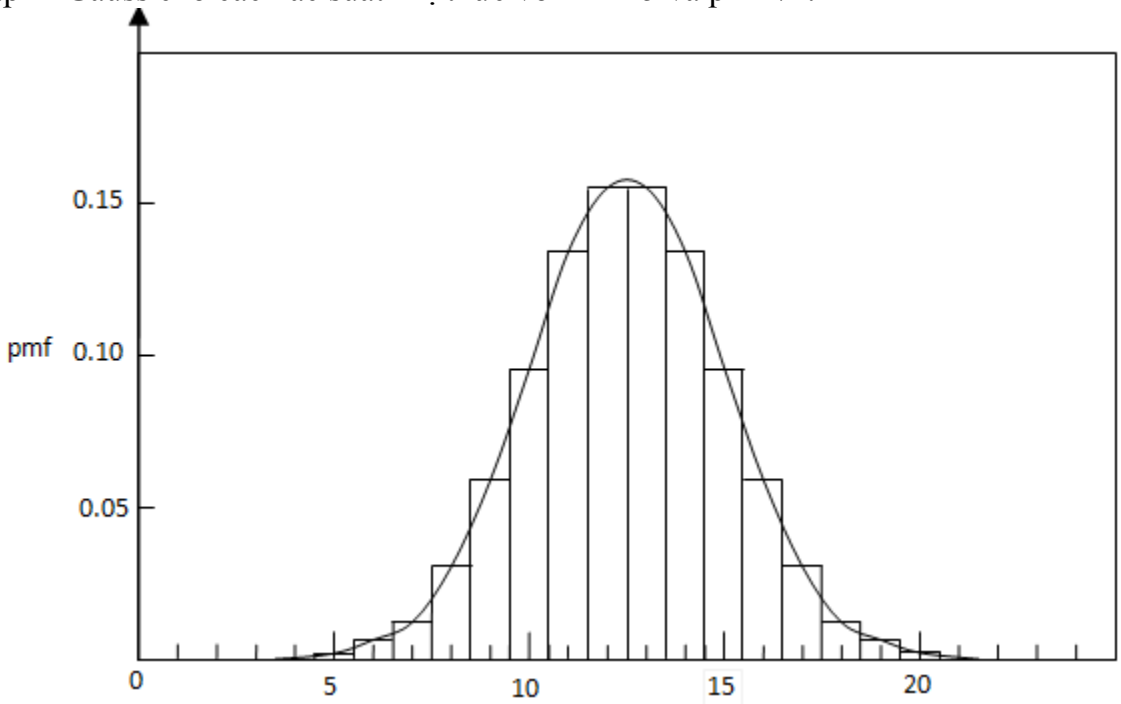
Các hình 5.6(a) và 5.6(b) so sánh các xác suất nhị thức và xấp xỉ Gauss dùng phương trình (5.30).

Hình 5.6a

Xấp xỉ Gauss cho các xác suất nhị thứcvới n = 5 và p = 1/2.



Hình 5.6b
Xấp xỉ Gauss cho các xác suất nhị thức với n = 25 và p = 1/2.



VÍ DỤ 5.14. Trong ví dụ 5.10 ở phần 5.2 chúng ta đã dùng bất đẳng thức Chebyshev để ước lượng số mẫu đòi hỏi sao cho với xác suất 0,95 ước lượng tần suất tương đối cho xác suất của biến cố A : P[A] bé hơn 0,01. Bây giờ chúng ta ước lượng số mẫu đòi hỏi dùng xấp xỉ Gauss cho phân phối nhị thức.

Giả sử $f_n(A)$ là tần suất tương đối của A trong n phép thử Bernoulli độc lập. Vì $f_n(A)$ có kỳ vọng n và phương sai p(1 – p)/n, nên :

$$Z_n = \frac{f_n(A) - p}{\sqrt{p(1-p)/n}}$$

có kỳ vọng 0 và phương sai 1, và là xấp xỉ Gauss cho n đủ lớn. Xác suất quan tâm là :

$P[|f_n(A) - p| < \varepsilon] \approx$ P**Error!** = 1 – 2Q**Error!**.

Xác suất trên không thể tính được vì p chưa được biết. Tuy nhiên, có thể dễ dàng chứng minh rằng $p(1-p) \le 1/4$ đối với p trong khoảng đơn vị. Từ đó suy ra $\sqrt{p(1-p)} \le ½$ và vì Q giảm khi đối số tăng

$P[|f_n(A) - p| < \varepsilon] > 1 - 2Q(2\varepsilon\sqrt{n})$

Chúng ta muốn xác suất trên bằng 0,95. Điều đó nói lên rằng $Q(2\varepsilon\sqrt{n}) =$ $= (1 - 0.95)/2 = 0.025$. Từ Bảng 3.3, chúng ta thấy rằng đối của Q(x) cần xấp xỉ bằng 1.95. Do đó

$2\varepsilon\sqrt{n} = 1.95.$

Giải với n chúng ta nhận được :

$n = (0.98)^2/\varepsilon^2 = 9506.$

*** Chứng minh của định lý giới hạn trung tâm :**

Bây giờ chúng ta phác thảo chứng minh của định lý giới hạn trung tâm. Trước hết chú ý rằng :

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu)$$

Hàm đặc trưng của Z_n được cho bởi :

$$\Phi_{Z_n}(\omega) = E[e^{j\omega Z_n}]$$

$$= E\text{**Error!**}$$

$$= E\text{**Error!**}$$

$$= \prod_{k=1}^{n} E\text{**Error!**}$$

$$= \{ E[e^{j\omega(X_k - \mu)/\sigma\sqrt{n}}] \}^n. \qquad (5.31)$$

Đẳng thức thứ ba trên đúng do tính độc lập các X_k' và đẳng thức cuối cùng đúng do các X_k' có cùng phân phối.

Bằng các khai triển hàm mũ chúng ta nhận được một biểu thức theo n và các momen trung tâm của X :

$$E[e^{j\omega(X_k - \mu)/\sigma\sqrt{n}}]$$

$$= E[1 + \frac{j\omega}{\sigma\sqrt{n}}(X - \mu) + \frac{(j\omega)^2}{2!n\sigma^2}(X - \mu)^2 + R(\omega)]$$

$$= 1 + \frac{j\omega}{\sigma\sqrt{n}}E[(X - \mu)] + \frac{(j\omega)^2}{2!n\sigma^2}E[(X - \mu)^2] + E[R(\omega)].$$

Để ý rằng $E[(X - \mu)] = 0$ và $E[(X - \mu)^2] = \sigma^2$, chúng ta có :

$$E[e^{j\omega(X_k - \mu)/\sigma\sqrt{n}}] = 1 - \frac{\omega^2}{2n} + E[R(\omega)]. \qquad (5.32)$$

Hạng thức $E[R(\omega)]$ có thể được bỏ đi khi n đủ lớn. Nếu chúng ta thế phương trình (5.32) vào (5.31) chúng ta nhận được

$$\Phi_{Z_n}(\omega) = \left\{1 - \frac{\omega^2}{2n}\right\}^n \quad \to e^{-\omega^2/2} \quad \text{khi} \quad n \to \infty.$$

Biểu thức cuối cùng này là hàm đặc trưng của biến ngẫu nhiên Gauss kỳ vọng 0, phương sai 1. Như vậy hàm phân phối của Z_n dần tới hàm phân phối của biến ngẫu nhiên Gauss kỳ vọng 0, phương sai 1.