

Lý Thuyết Xác Suất (2): Phân Phối Xác Suất

Khái quát

- Biến ngẫu nhiên.
- Phân phối xác suất của biến ngẫu nhiên.
- Phân phối binomial.
- Các đặc trưng của phân phối xác suất.
- Phân phối chuẩn.
- Một số phân phối thường gặp.

Biến Ngẫu Nhiên – Phân Phối Xác Suất Biến Ngẫu Nhiên

Biến Ngẫu Nhiên

- **Biến ngẫu nhiên (random variable):** biến ngẫu nhiên $X: S \rightarrow \mathbb{R}$ là một hàm số gán giá trị cho các quan sát của phép thử ngẫu nhiên.
- **S là tập hợp không gian mẫu (sample space).** Tập hợp S của biến ngẫu nhiên có hữu hạn hoặc vô hạn, rời rạc hoặc liên tục.
- Tên gọi 'biến ngẫu nhiên' dễ gây nhầm lẫn, do một biến ngẫu nhiên không phải là một biến (a variable) hay có tính ngẫu nhiên (random). Nó mà là một **hàm số ánh xạ các biến cố tới các số**.
- Chúng ta có thể kí hiệu: biến ngẫu nhiên X, Y, Z hay $X_1, X_2, \dots$

Biến Ngẫu Nhiên – Ví Dụ 1

- Khi tung một đồng xu. Hai khả năng sắp ngửa xảy ra (S và N). Biến ngẫu nhiên X :

$$X(S) = 0$$

$$X(N) = 1$$

- Rót nước vào 1 can có dung tích 2 lít. X là biến ngẫu nhiên:

$$X = \text{lượng nước đổ vào}$$

$$0 \leq X \leq 2$$

- Liên tục hay rời rạc? Hữu hạn hay vô hạn?

Biến Ngẫu Nhiên – Ví Dụ 2

- Một người bắn vào bia cho đến khi nào trúng thì dừng (trúng kí hiệu T , không trúng kí hiệu là $\bar{T}$).
 - Không gian mẫu $S = \{T, \bar{T}T, \bar{T}\bar{T}T, \bar{T}\bar{T}\bar{T}T \dots\}$
 - Biến ngẫu nhiên: $X = \{1, 2, 3, \dots, n, \dots\}$
- Quan sát thời gian giữa 2 cuộc gọi đến văn phòng Shopee:
 $X = \text{thời gian} \in [0, +\infty)$
- Liên tục hay rời rạc? Hữu hạn hay vô hạn?

Phân Phối Xác Suất (1)

- Đối với biến ngẫu nhiên X , **phân phối xác suất của X** hiểu như là tập hợp tất cả các giá trị xác suất dưới dạng $\mathbf{P}(X(s) \in \mathcal{C})$ cho các biến cố $\mathcal{C}$ thuộc tập không gian mẫu S .
- Xác suất này có thể biến đổi bằng hàm số:
$$f(x) = \mathbf{P}(s: X(s) \in \mathcal{C})$$
- Với tất cả các biến cố có thể xảy ra ($f(x_i) > 0$), tổng xác suất các biến cố đó phải bằng 1.

$$\sum f(x_i) = 1, \quad \int f(x)dx = 1$$

Phân Phối Xác Suất (2)

- Với ví dụ tung đồng xu một lần:

	Sấp	Ngửa
X	0	1
$P(X = x) \equiv f(x)$	0.5	0.5

- Tung đồng xu hai lần liên tiếp:

	SS	SN	NS	NN
Y	1	2	3	4
$P(Y = y) \equiv f(y)$	0.25	0.25	0.25	0.25

Phân Phối Nhị Thức (Phân Phối Binomial) (1)

- Phép thử binomial:
 - Dẫn tới hai kết quả, có thể tạm gọi thành công và không thành.
 - Bao gồm n lần giống hệt nhau.
 - Xác suất của từng kết quả cố định là p và $1 - p$.
 - Các phép thử độc lập với nhau.
 - Chúng ta muốn tính số lượng xảy ra của một kết quả (thành công) trên n lần.
- Ví dụ:
 - Sản xuất điện thoại mỗi dây chuyền 25000 sản phẩm và tính số lượng không đạt chất lượng.
 - Tung đồng xu 1000 lần và tính tỷ lệ sấp hoặc ngửa.

Phân Phối Nhị Thức (Phân Phối Binomial) (2)

- Giả sử chúng ta chọn $n = 6$, các kết quả có thể xảy ra là:
010001, 1001110, 101010, 0110110, 000000 ...
- X là biến ngẫu nhiên cho số lượng thành công ở ví dụ trên:
 $X = 2$ cho thí nghiệm đầu tiên
 $X = 0$ cho thí nghiệm cuối cùng
- Không gian mẫu gồm bao nhiêu quan sát với bao nhiêu giá trị X ?

Phân Phối Nhị Thức (Phân Phối Binomial) (3)

- Công thức tổng quát cho phân phối xác suất binomial $B(n, p)$, với n lần xảy ra, xác suất thành công là p , thì tỷ lệ xảy ra x lần:

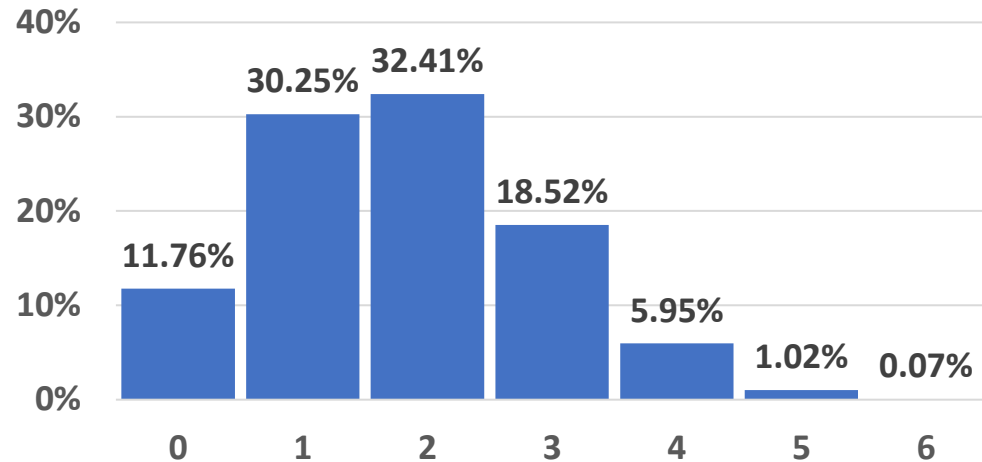
$$f(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

với $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

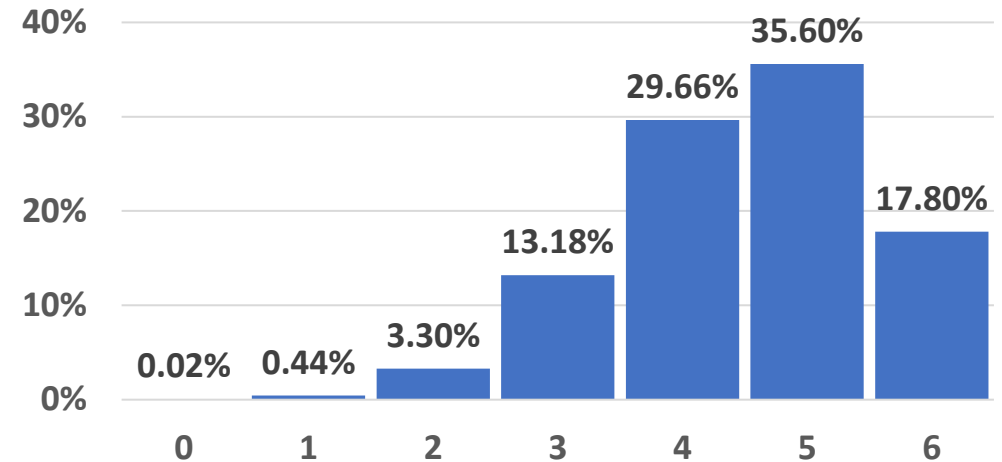
- Chú ý: $f(x) = 0$ khi $x \leq 0$ hoặc $x > n$.

Biểu Diễn Phân Phối Binomial

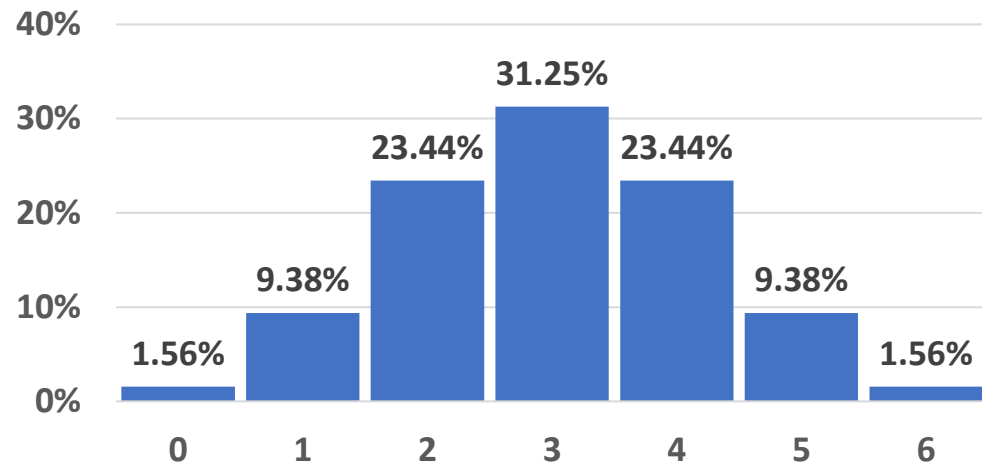
Binominal Distribution ($n=6, p=0.3$)



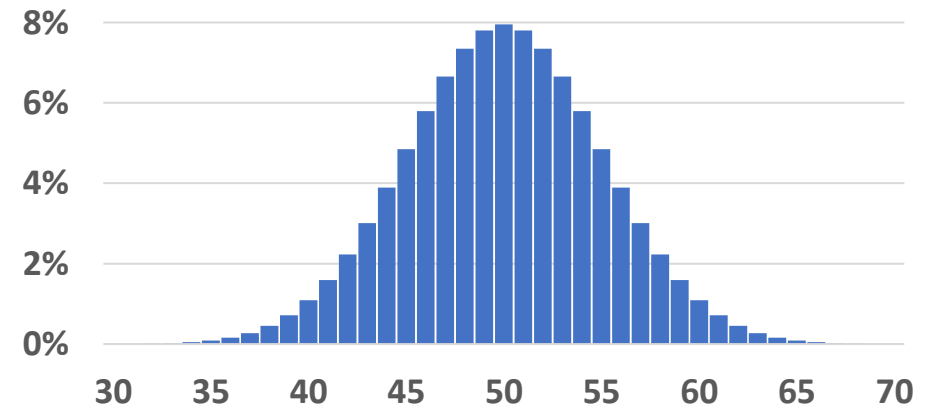
Binominal Distribution ($n=6, p=0.75$)



Binominal Distribution ($n=6, p=0.5$)



Binominal Distribution ($n=100, p=0.5$)



Phân Phối Liên Tục (1)

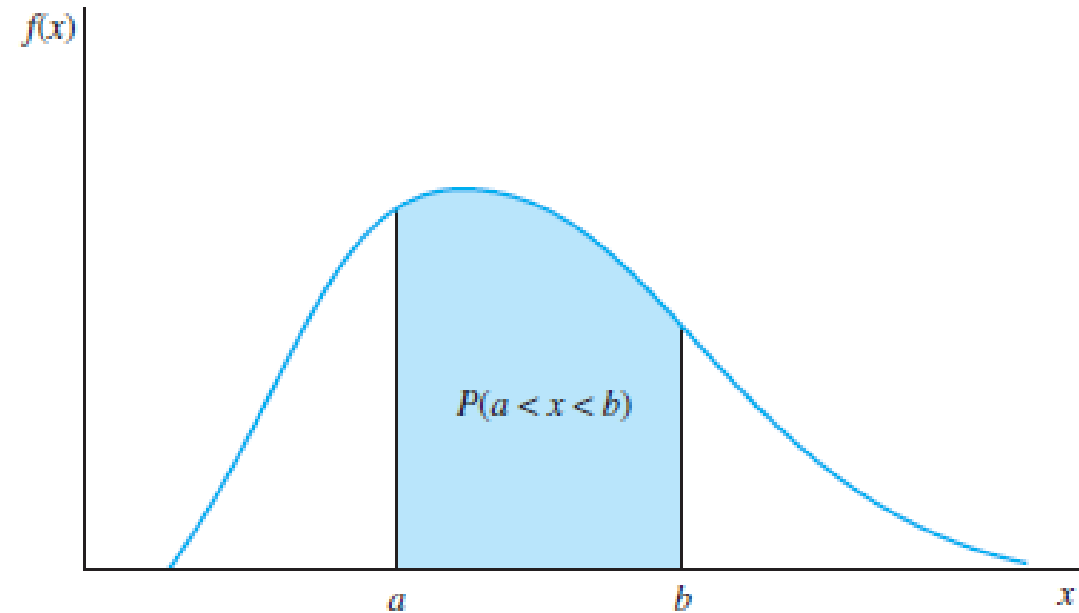
- **Phân phối liên tục** là một phân mà biến ngẫu nhiên có giá trị liên tục, ví dụ: $X \in (-\infty, +\infty)$
- Giá trị xác suất được định trên một đoạn bằng một hàm **probability density function (pdf)** – $f(x) \geq 0$ – như sau:

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

$$P(X \geq a) = \int_a^{\infty} f(x)dx$$

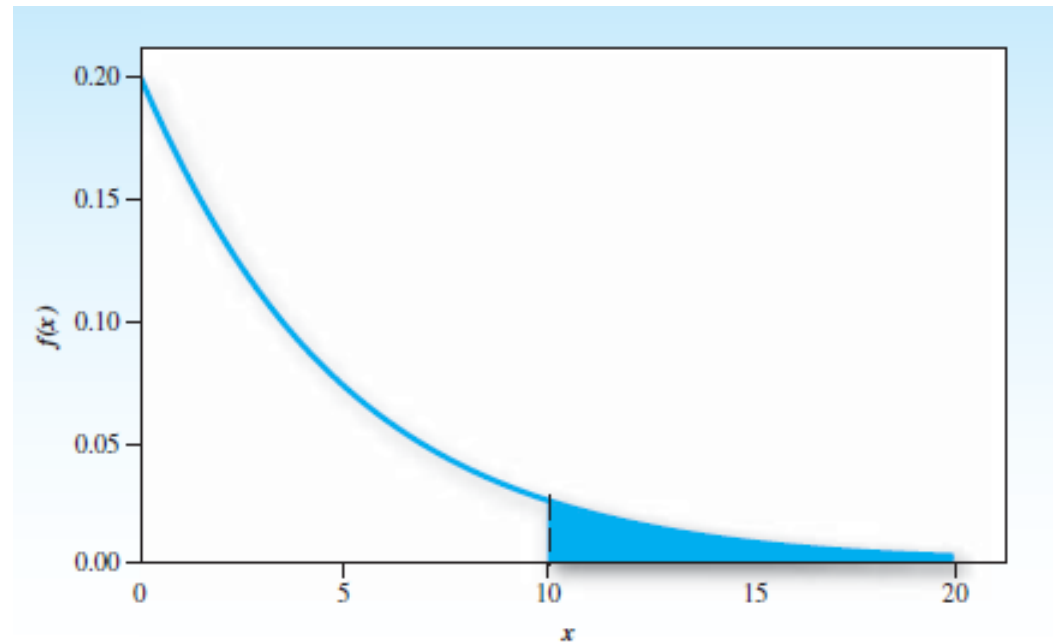
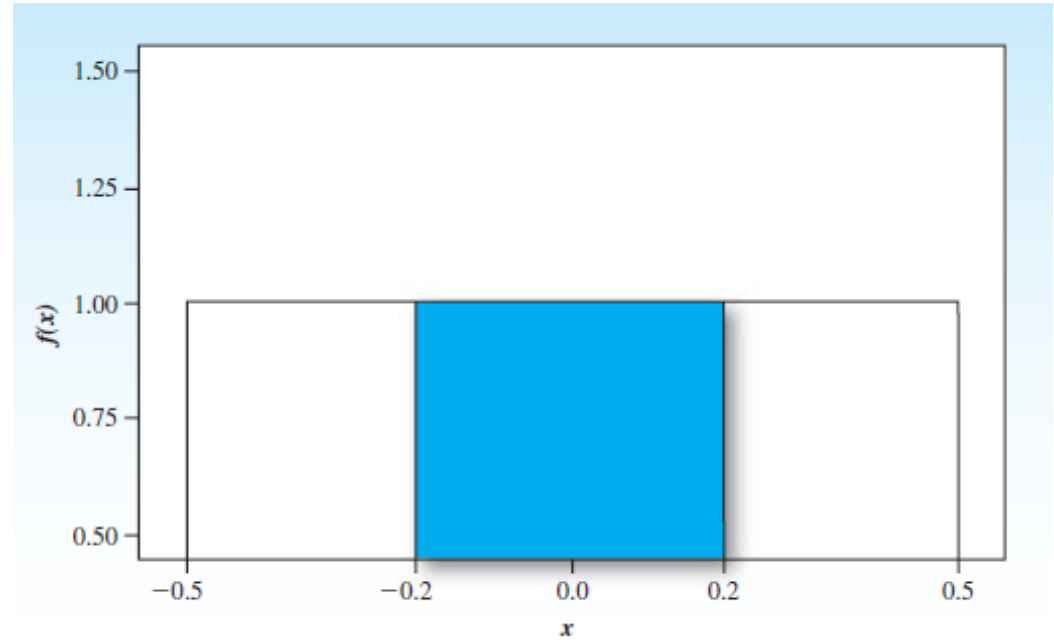
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

$$P(a) = \int_a^a f(x)dx = 0$$



Phân Phối Liên Tục (2)

- Làm tròn một con số đến giá trị trong khoảng từ -0.5 đến 0.5 có **phân phối đều liên tục** (continuous uniform distribution) vì tất cả đều được làm có kết quả là không, $f(x) = 1$.
- Thời gian chờ đợi ở siêu thị có phân phối theo hàm $f(x) = 0.2e^{-.2x}$ là một **phân phối mũ** (exponential distribution).



Các Đặc Trưng Của Phân Phối Xác Suất

Giá Trị Kỳ Vọng (1)

- **Giá trị kỳ vọng (expectation)** – kí hiệu: $E(X)$ hay μ – phản ánh giá trị trung tâm của một phân phối.
- Đối với phân phối rời rạc:

$$\begin{aligned} E(X) &= \sum_{i=1}^n x_i P_i = x_1 P_1 + x_2 P_2 + \cdots + x_n P_n \\ &= \sum x_i f(x_i) \end{aligned}$$

- Đối với phân phối liên tục:

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

Giá Trị Kỳ Vọng (2)

- Đồ xí ngẫu 12 mặt, giá trị kì vọng:

$$E(X) = 1 * \frac{1}{12} + 2 * \frac{1}{12} + \dots + 12 * \frac{1}{12} = 6.5$$

- Giá trị kì vọng của phân phối binomial:

$$E(X) = np$$

Phương Sai, Độ Lệch Chuẩn

- **Phương sai (variance)** của biến ngẫu nhiên X được tính:

$$\text{Var}(X) = E \left[(X - E(X))^2 \right] = E(X^2) - [E(X)]^2$$

trong đó:

$$E(X^2) = \sum_{i=1}^n x_i^2 P(x_i) \quad \text{nếu } X \text{ rời rạc}$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx \quad \text{nếu } X \text{ liên tục}$$

- **Độ lệch chuẩn (standard deviation)** là căn bậc hai của phương sai:

$$\sigma = \sqrt{\text{Var}(X)}$$

Ví Dụ:

- Đồ xí ngẫu 6 mặt, phương sai được tính (với $E(X) = 3.5$):

$$\begin{aligned}\text{Var}(X) &= (1 - 3.5)^2 * \frac{1}{6} + (2 - 3.5)^2 * \frac{1}{6} + (3 - 3.5)^2 * \frac{1}{6} \\ &\quad + (4 - 3.5)^2 * \frac{1}{6} + (5 - 3.5)^2 * \frac{1}{6} + (6 - 3.5)^2 * \frac{1}{6} \\ &= 2.917\end{aligned}$$

- Độ lệch chuẩn:

$$\sigma = \sqrt{\text{Var}(X)} = 1.708$$

Ý Nghĩa Của Phương Sai

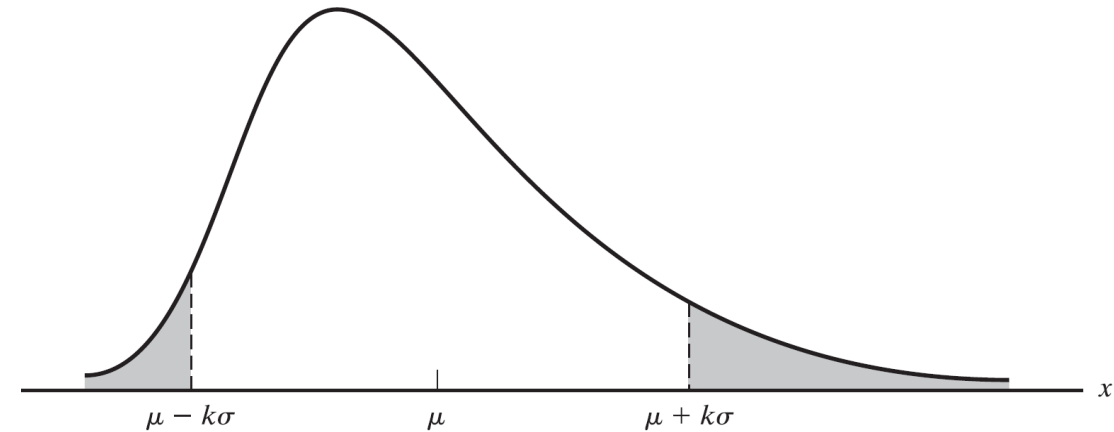
- Ta có $X - E(X)$ đo lường độ lệch giữa giá trị của X và giá trị trung bình của nó. Phương sai $Var(X)$ đo lường sự phân tán khỏi giá trị trung bình của X .
- Nếu cùng một giá trị kì vọng $E(X)$, phương sai lớn thì độ phân tán lớn và độ tập trung nhỏ; và ngược lại nếu phương sai nhỏ.
- Phương sai có thể dùng để tính toán độ rủi ro của sản phẩm tài chính hay quyết định.

Định Lý Chebyshev

- Một phân phối bất kì X có kì vọng μ và độ lệch chuẩn σ , đối với bất kì số k dương nào, chúng ta có:

$$P(|x - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

- Xác suất của tối đa nằm ngoài k độ lệch chuẩn (phần xám) của $\mu - k\sigma$:
 - $k = 1.5$: 44%
 - $k = 2$: 25%
 - $k = 3$: 11.11%



Tính Chất Của $E(X)$

- $E(c) = c$, c là hằng số.
- $E(aX) = aE(X)$
- $E(X + Y) = E(X) + E(Y)$
- Nếu X, Y độc lập thì $E(XY) = E(X)E(Y)$

Tính Chất Của $\text{Var}(X)$

- $\text{Var}(c) = 0$, c là hằng số.
- $\text{Var}(cX) = c^2 \text{Var}(X)$
- $\text{Var}(X + c) = \text{Var}(X)$
- Nếu X, Y độc lập $\text{Var}(X \pm Y) = \text{Var}(X) \pm \text{Var}(Y)$

Các Đặc Trưng Phổ Biến Khác (1)

- **Trung vị (median):** giá trị của X phân chia phân phối thành hai phần bằng nhau.
- **Số yếu vị (Mod(X)):** là giá trị x_i của X mà $f(x_i)$ có giá trị lớn nhất. Lưu ý Mod(X) có thể nhận nhiều giá trị một lúc.

Ví dụ:

- Tính giá trị trung bình, phương sai, độ lệch chuẩn, median và mode của biến ngẫu nhiên X rời rạc có phân phối như sau:

X	-4	0	2	4	8
$P(X)$	0.3	0.1	0.1	0.3	0.2

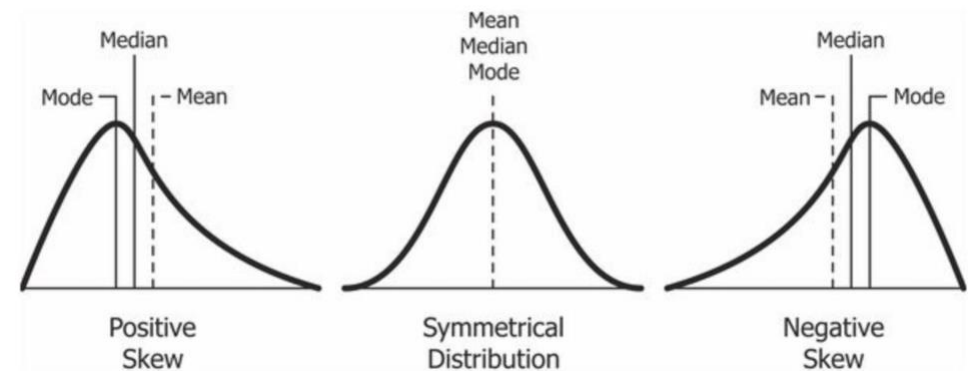
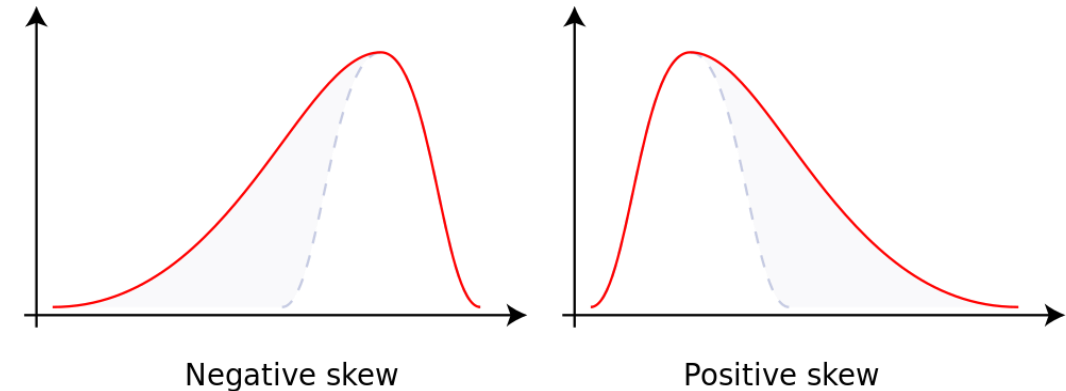
Các Đặc Trưng Phổ Biến Khác (2)

- **Độ xiên (skewness)** đo độ đối xứng của X :

$$\text{Skew}(X) = E \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right]$$

- **Độ nhọn (kurtosis)** đo mức độ tập trung xung của các quan sát trung tâm so với hai đuôi:

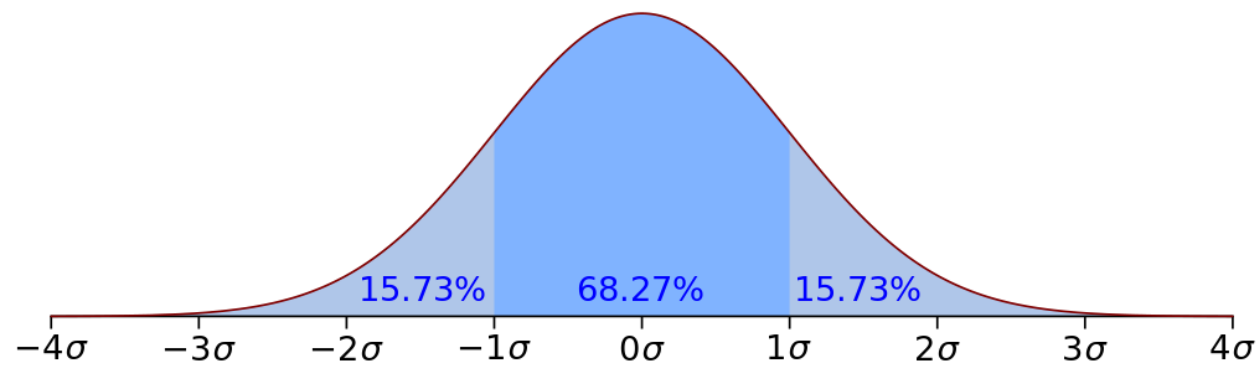
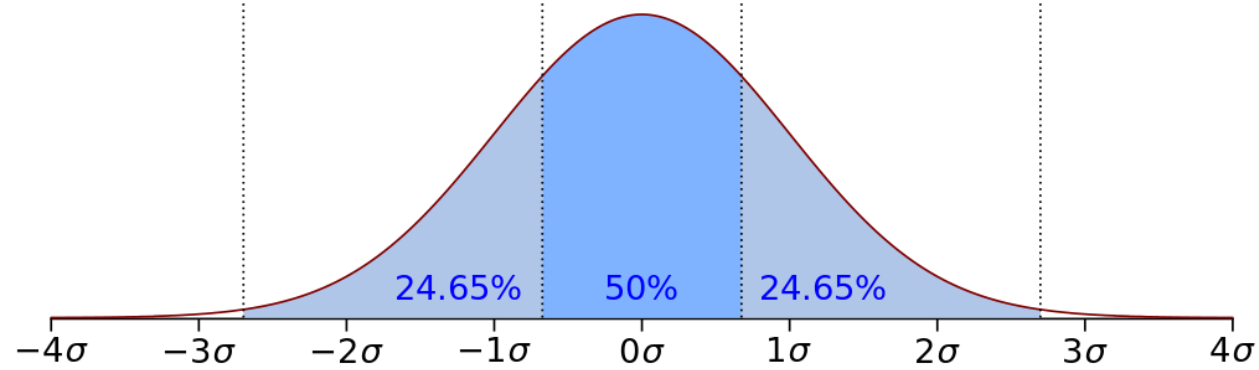
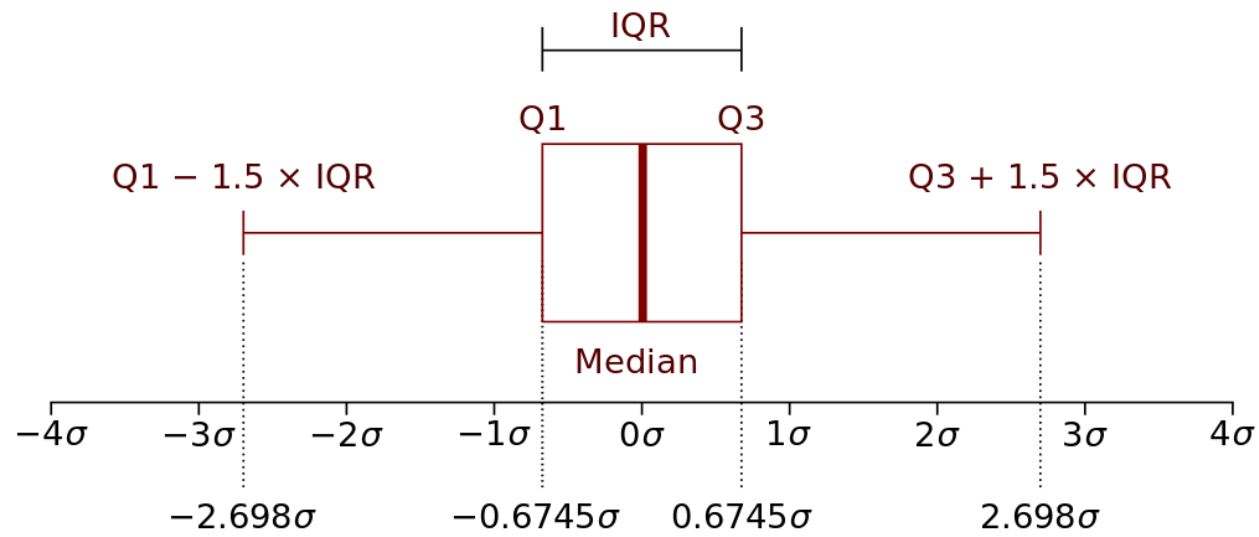
$$\text{Kurt}(X) = E \left[\left(\frac{X - \mu}{\sigma} \right)^4 \right]$$



Source: <https://en.wikipedia.org/wiki/Skewness>

Các Đặc Trưng Phổ Biến Khác (3)

- **Điểm tứ phân vị (quartile):** chia dữ liệu thành bốn phần có tỷ suất phân bố bằng nhau:
 - Giá trị tứ phân vị thứ nhất Q1 (first quartile): 25% phân phối nhỏ hơn Q1.
 - Giá trị tứ phân vị thứ nhì Q2 (second quartile): đúng bằng trung vị (median), 50% phân phối nằm hai bên Q2.
 - Giá trị tứ phân vị thứ ba Q3 (third quartile): 75% phân phối nhỏ hơn Q3.
- Khoảng cách giữa Q1 và Q3 gọi là **độ trải giữa (interquartile range IQR)**.



Phân Phối Chuẩn (Normal Distribution)

Phân Phối Chuẩn (1)

- **Phân phân chuẩn** $X \sim N(\mu, \sigma^2)$ bao gồm hai tham số giá trị kì vọng μ và phương sai σ^2 ($\sigma > 0$).

$$X \sim N(\mu, \sigma^2)$$

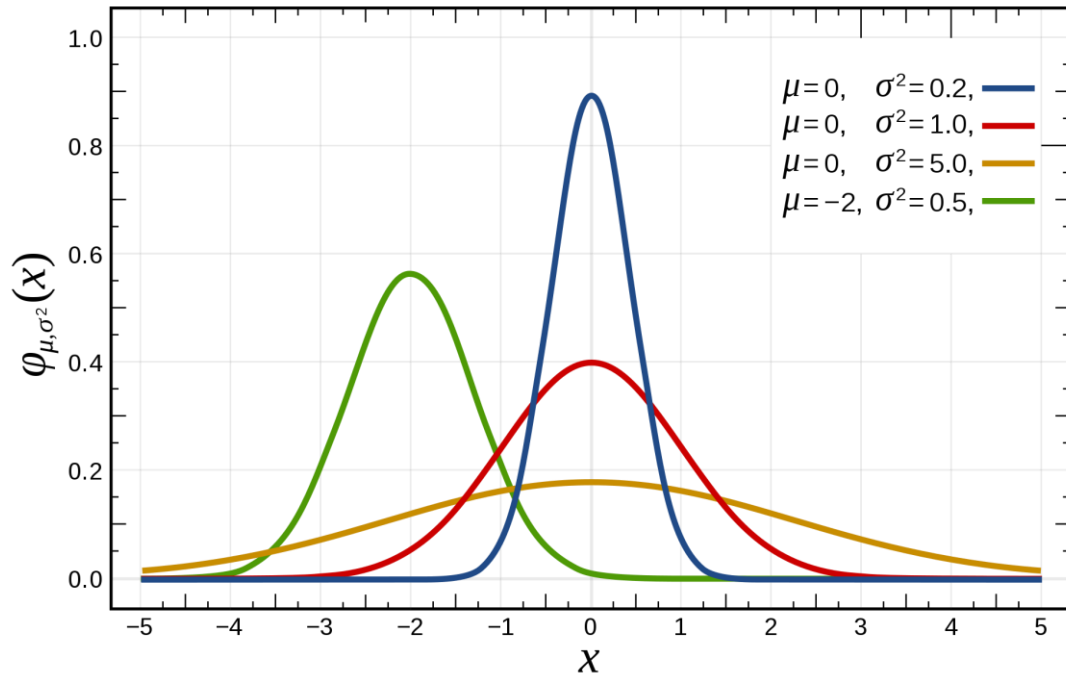
$$E(X) = \mu, \quad \text{Var}(X) = \sigma^2$$

- Phân phối chuẩn có hàm p.d.f. như sau:

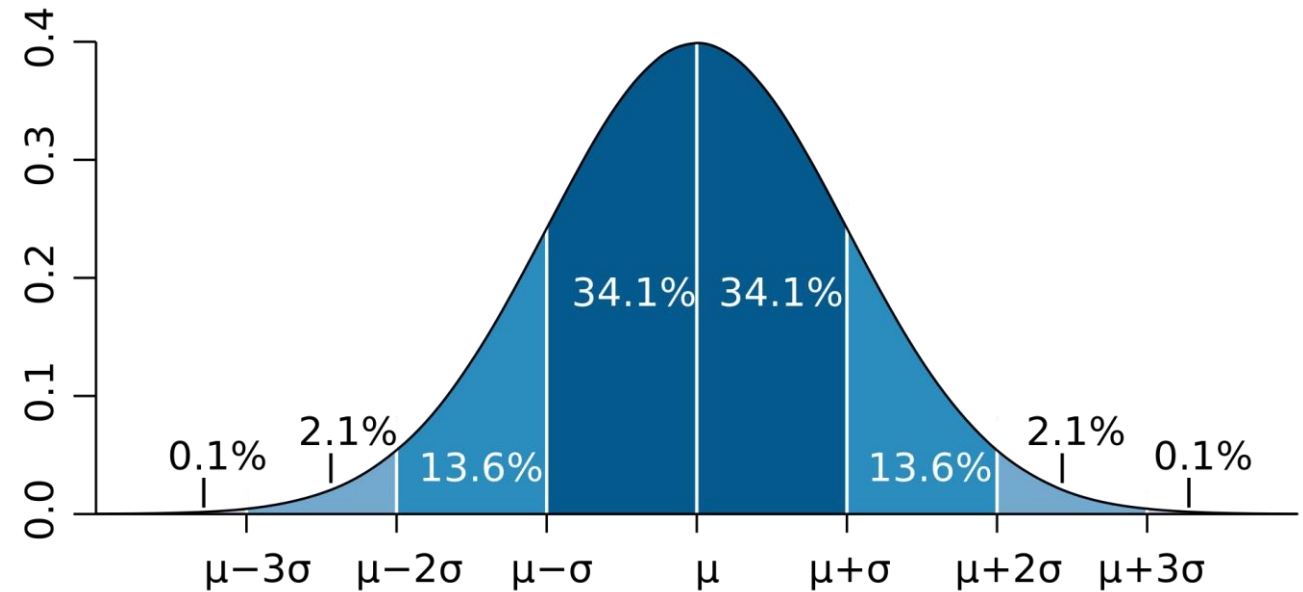
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \forall x \in \mathbb{R}$$

- Phân phối chuẩn rất thường gặp trong tự nhiên và xã hội (chiều cao, cân nặng người, chỉ số IQ, huyết áp...)

Phân Phối Chuẩn (2)



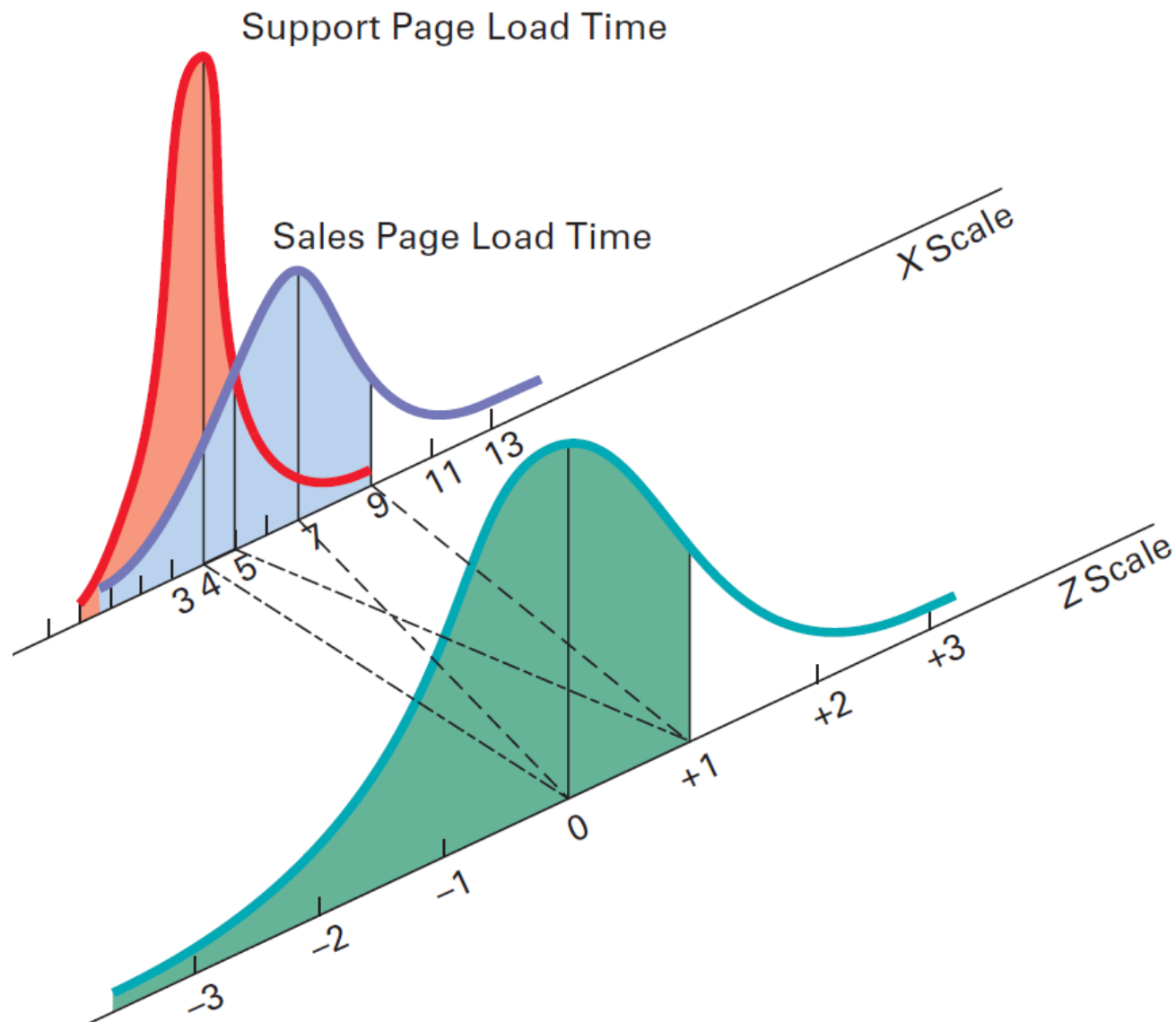
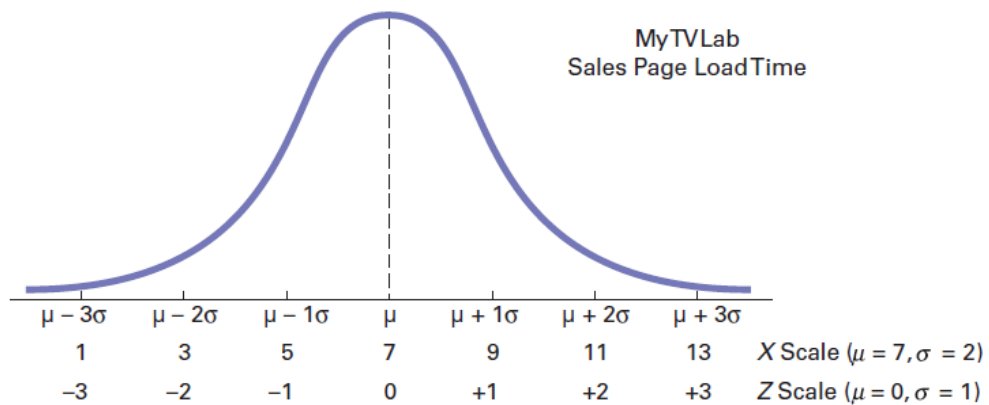
Source: https://en.wikipedia.org/wiki/Normal_distribution



Phân Phối Chuẩn Tắc (Standard Normal Distribution)

- Khi $\mu = 0$ và $\sigma = 1$, thì $Z \sim N(0, 1)$. Ta nói Z có **phân phối chuẩn tắc (standard normal distribution)**.
- Khi $X \sim N(\mu, \sigma^2)$, ta có thể biến đổi $Z = \frac{X - \mu}{\sigma}$ thì Z sẽ có phân phối chuẩn tắc $Z \sim N(0, 1)$ – áp dụng tính chất của $E(X)$ và $Var(X)$ ở trên.
- Phân phối chuẩn tắc có skewness = 0, và kurtosis = 3.

- Sales Page Load Time: $X \sim N(7, 2^2)$
- Support Page Load Time: $Y \sim N(4, 1^2)$
- Phân phối chuẩn tắc: $Z \sim N(0,1)$



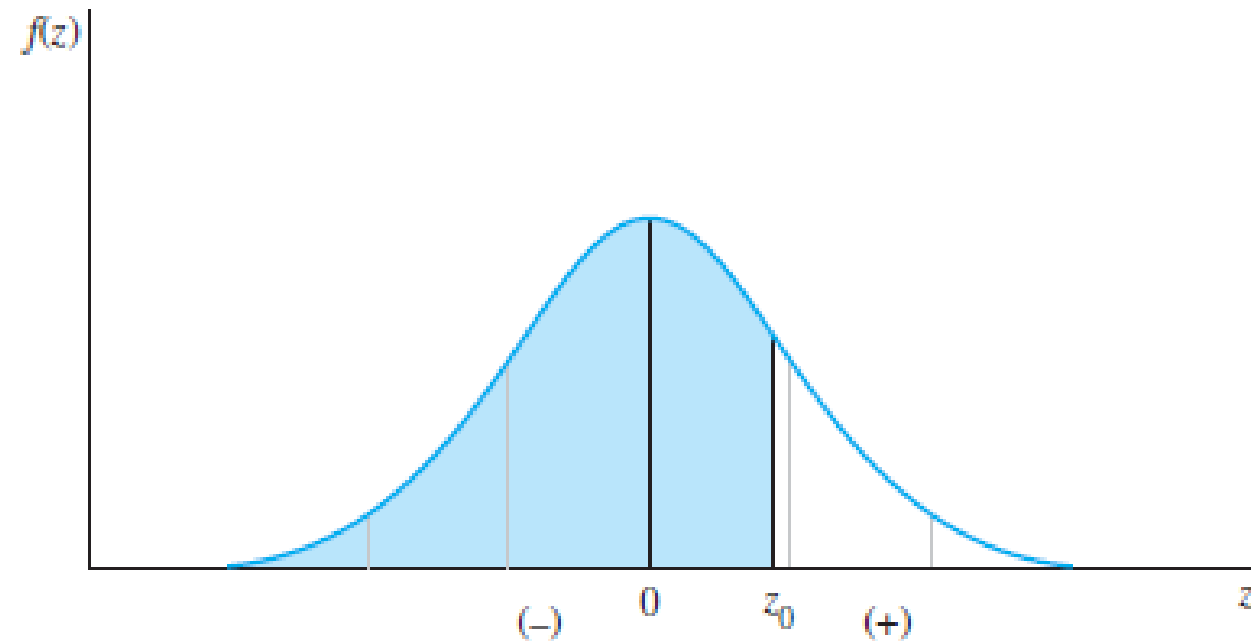
Cách Tính Giá Trị Phân Phối Chuẩn (1)

- Bảng giá trị cung cấp cho chúng ta tại từng giá trị xác định z_0 :

$$P(z_0 < Z)$$

- Sử dụng hàm Excel NORM.DIST ()
- Hàm $P(z_0 < Z)$ gọi là hàm phân phối tích lũy. Tổng quát:

$$F(x) = P(X \leq x)$$



Cách Tính Giá Trị Phân Phối Chuẩn (2)

- Ta có X là một phân phối chuẩn với giá trị $\mu = 10$ và $\sigma = 2$. Hãy tính
 - Xác suất X nằm giữa 9,4 và 10,6.
 - Xác suất $X > 12$.
 - Xác suất $X < 9$.

Cách Tính Giá Trị Phân Phối Chuẩn (3)

- Tính giá trị Z.
- Xác suất X nằm giữa 9,4 và 10,6.

$$Z_l = \frac{9,4 - 10}{2} = -0,3$$

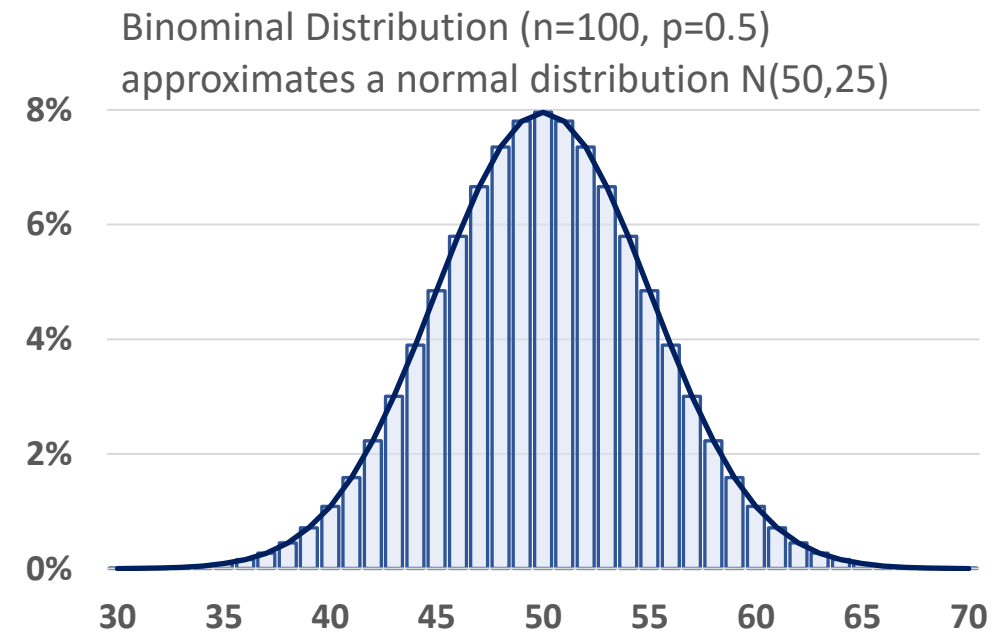
$$Z_u = \frac{10,6 - 10}{2} = 0,3$$

- Tra bảng:

$$P(9,4 < X < 10,6) = 0,618 - 0,382 = 0,236$$

Xấp Xỉ Của Phân Phối Binomial

- Cho một hàm phân phối binomial với tham số n và p .
- Khi giá trị np đủ lớn thì phân phối này xấp xỉ một phân phối chuẩn với giá trị kì vọng $\mu = np$ và phương sai $\sigma^2 = np(1 - p)$.
- Để xấp xỉ này có thể diễn ra thì n phải lớn và p phải không quá gần 0 và 1.

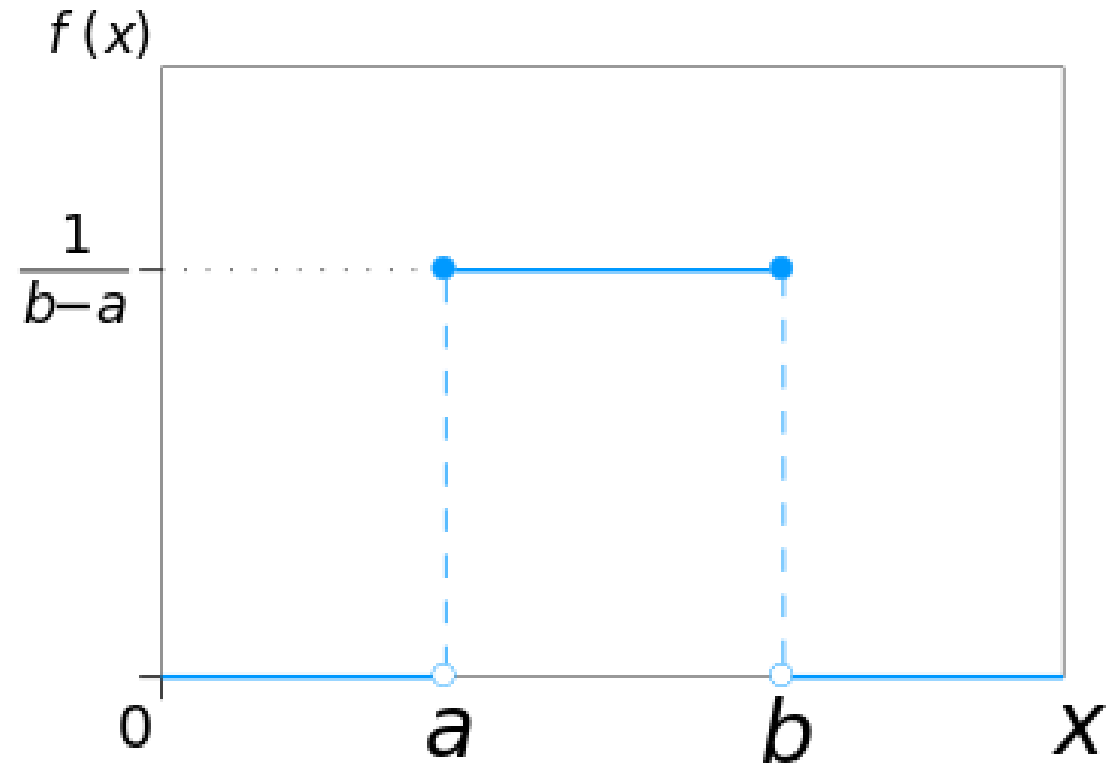


Một Số Phân Phối Thường Gặp (1)

- **Phân phối đều liên tục (the continuous uniform distribution)** là một phân phối mà xác suất xảy ra là bằng nhau cho tất cả các kết quả.

- Hàm phân phối xác suất có dạng:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{khi } a \leq x \leq b \\ 0 & \text{khi } x < a \text{ hoặc } x > b \end{cases}$$

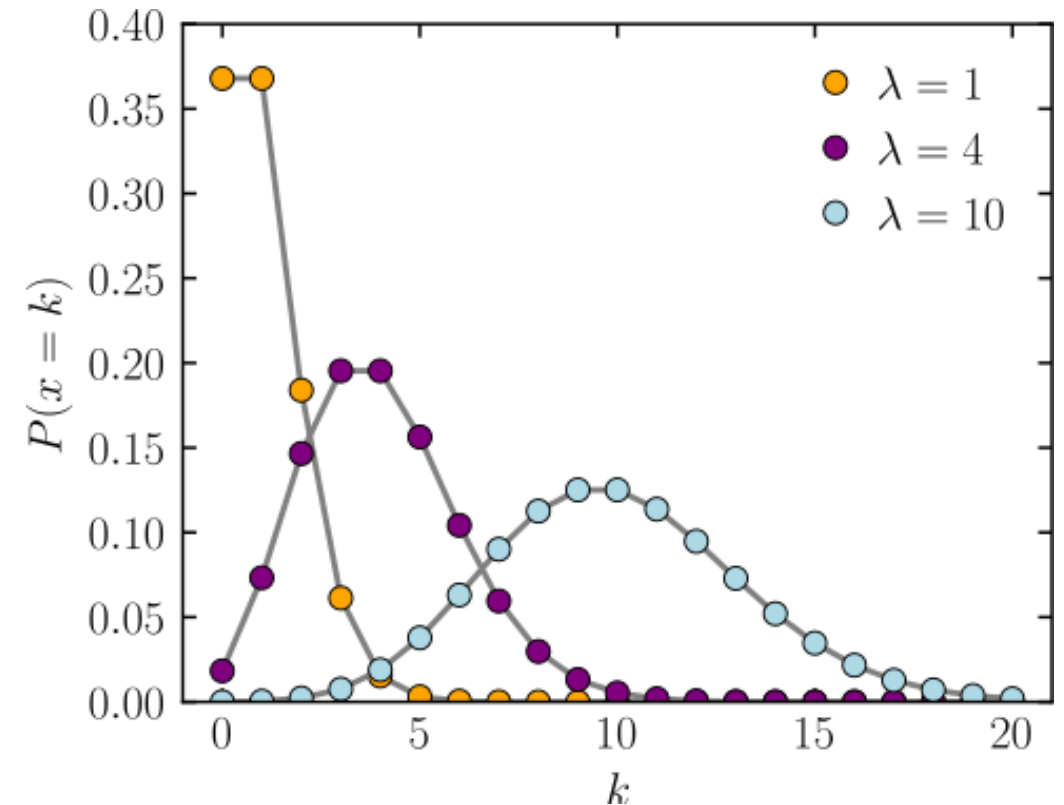


Một Số Phân Phối Thường Gặp (2)

- **Phân phối Poisson** đo lường số lượng hiện tượng đồng nhất, ngẫu nhiên và độc lập xảy ra trong một khoảng thời gian nhất định.
- Ví dụ:
 - Số lượng tai nạn giao thông trong 24 giờ tại TP HCM.
 - Số lượng khách đi xe bus tại một trạm Bến Thành trong 1 giờ.
- Phân phối Poisson X có hàm phân phối rời rạc:

$$f(k; \lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

với k là số lần xảy ra, λ là một số thực bằng với giá trị kì vọng và phương sai của phân phối $\lambda = E(X) = \text{Var}(X)$.



Source: https://en.wikipedia.org/wiki/Poisson_distribution

Ví dụ: Phân Phối Poisson

- Trung bình có 1 trận lụt lớn xảy ra trong vòng 10 năm, xác suất xảy ra:
- Trung bình 1 trận bóng đá World Cup có 2.5 bàn thắng, xác suất xảy ra:

k	$P(k \text{ lần lụt xảy ra})$
0	0.368
1	0.368
2	0.184
3	0.061
4	0.015
5	0.003
6	0.0005

k	$P(k \text{ bàn thắng})$
0	0.082
1	0.205
2	0.257
3	0.213
4	0.133
5	0.067
6	0.028

Một Số Phân Phối Thường Gặp (3)

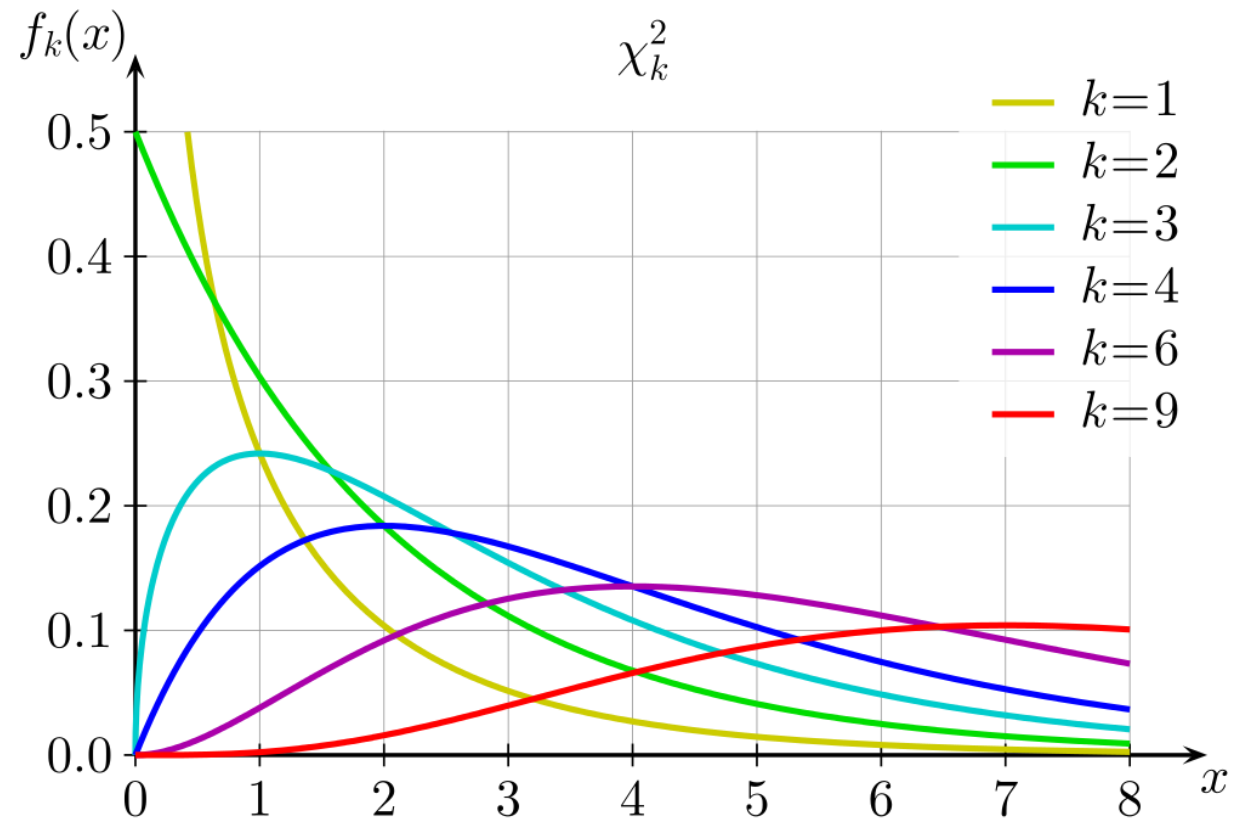
- **Phân Phối χ :** nếu chúng ta có k phân phối chuẩn tắc $Z_1, Z_2, \dots, Z_k$ ($N(0,1)$) độc lập với nhau, thì tổng của chúng:

$$Q = \sum_{i=1}^k Z_i^2$$

sẽ có phân phối:

$$Q \sim \chi^2(k)$$

với k là một số dương, và được gọi là **bậc tự do (degrees of freedom/d.o.f.)** của phân phối.



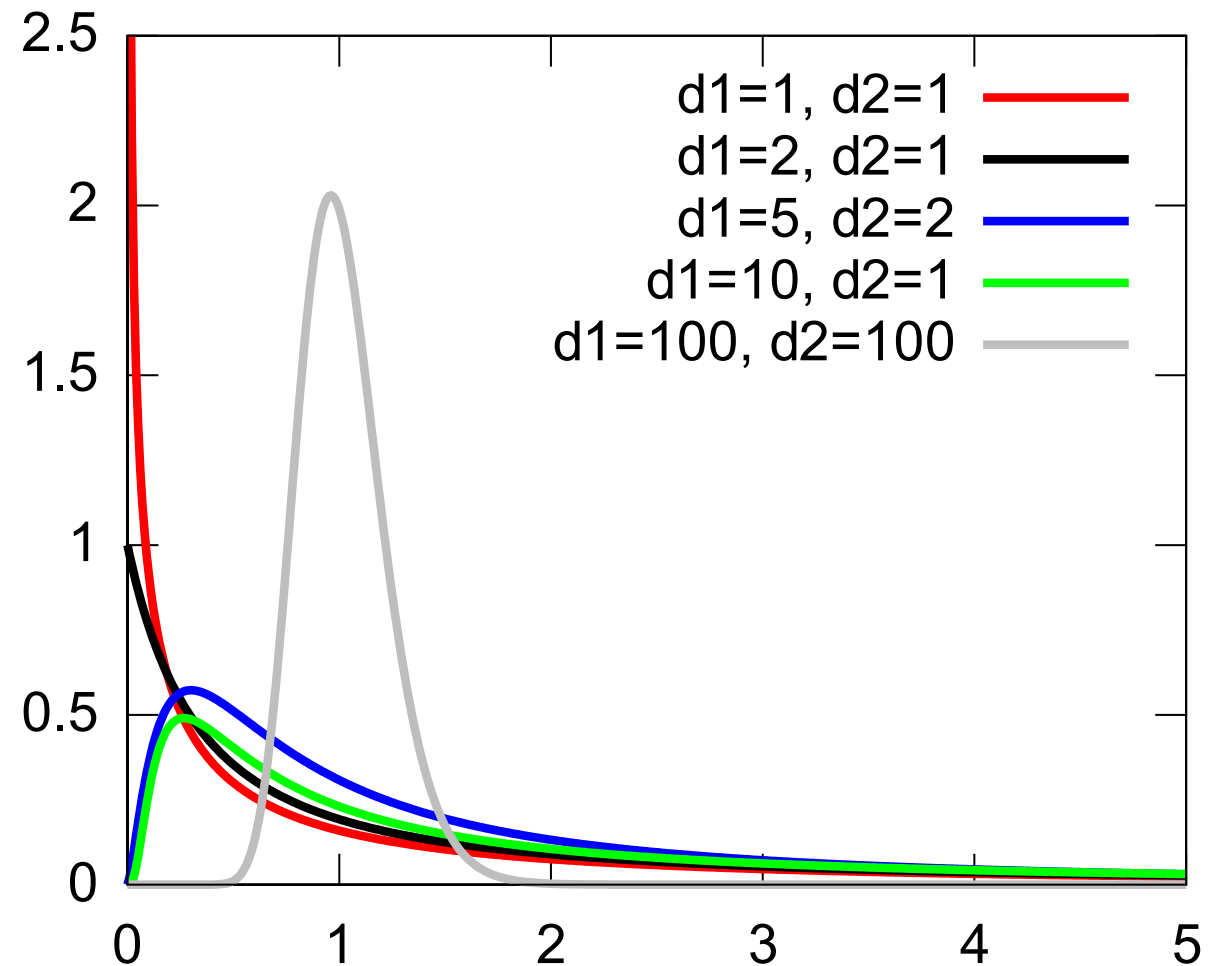
Source: https://en.wikipedia.org/wiki/Chi-squared_distribution

Một Số Phân Phối Thường Gặp (4)

- **Phân phối F (the F distribution)**
 $F(d_1, d_2)$ với d_1 và d_2 bậc tự do
là phân phối có công thức:

$$F = \frac{Q_1/d_1}{Q_2/d_2}$$

với Q_1, Q_2 là các biến ngẫu nhiên có phân phối $\chi^2(d_1)$ và $\chi^2(d_2)$.



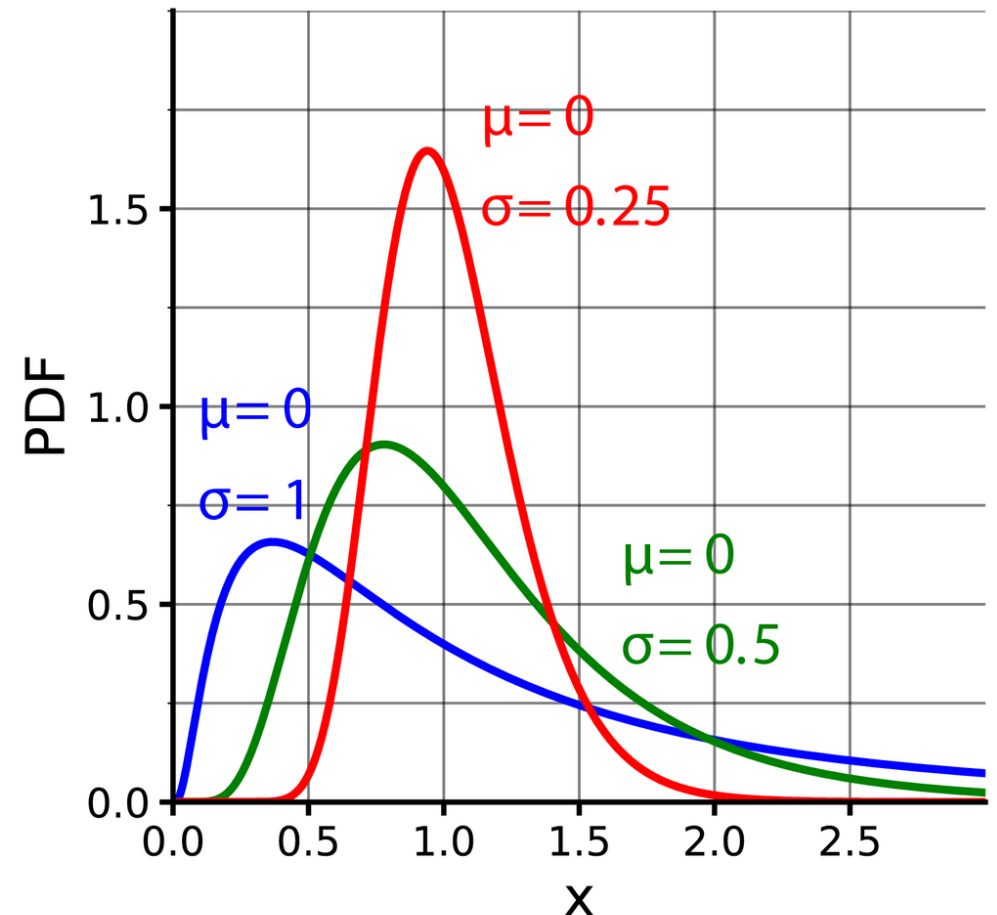
Source: <https://en.wikipedia.org/wiki/F-distribution>

Một Số Phân Phối Thường Gặp (5)

- **Phân phối loga chuẩn** (the log-normal distribution) là biến ngẫu nhiên khi lấy hàm log của chúng thì chúng ta được hàm phân phối chuẩn.

$$X \sim \text{LogNormal}(\mu_x, \sigma_x^2) \\ \Leftrightarrow \log(X) \sim N(\mu, \sigma^2)$$

- Ví dụ: số lượng chữ của comment trên mạng xã hội, tỷ suất lợi nhuận giá cổ phiếu, thời gian của một trận cờ vua.



Source: https://en.wikipedia.org/wiki/Log-normal_distribution