# PHÂN TÍCH DỮ LIỆU NGHIÊN CỨU

# Data Analysis

Lê Kim Long
Phạm Thành Thái

# Nội Dung Bài Giảng

- Chủ đề 1:Các quy luật phân phối xác suất cơ bản và suy diễn thống kê.
- Chủ đề 2:Mô hình hồi quy đơn.
- Chủ đề 3: Mô hình hồi quy bội.
- Chủ đề 4: Biến giả.
- Chủ đề 5: Đa cộng tuyến- Phương sai thay đổi – Tự tương quan.

# Chủ đề 1: CÁC QUY LUẬT PHÂN PHỐI XÁC SUẤT THÔNG DỤNG VÀ SUY DIỄN THỐNG KÊ

# PHẦN I

- ĐẠI LƯỢNG NGẪU NHIÊN(ĐLNN)
- BẢNG PHÂN PHỐI XÁC SUẤT
- HÀM MẬT ĐỘ XÁC SUẤT
- CÁC ĐẶC TRƯNG CỦA ĐLNN X
- MỘT SỐ PHÂN PHỐI XÁC SUẤT QUAN TRỌNG

# ĐẠI LƯỢNG NGẪU NHIÊN

- Khái niệm: Một biến mà giá trị của nó được xác định bởi một phép thử ngẫu nhiên được gọi là một biến ngẫu nhiên hay đại lượng ngẫu nhiên, thường viết tắc là ĐLNN (Random Variable).
- Phân loại:

  - Đại lượng ngẫu nhiên rời rạc
  - Đại lượng ngẫu nhiên liên tục

# ĐẠI LƯỢNG NGẪU NHIÊN

- Ví dụ 1: Gọi X là số chấm xuất hiện khi tung một con súc sắc. X là một biến ngẫu nhiên rời rạc vì nó chỉ có thể nhận các kết quả 1,2,3,4,5 và 6.
- Ví dụ 2: Gọi Y là chiều cao của một người được chọn ngẫu nhiên trong một nhóm người. Y cũng là một biến ngẫu nhiên vì chúng ta chỉ có nhận được sau khi đo đạc chiều cao của người đó. Trên một người cụ thể chúng ta đo được chiều cao 167 cm. Con số này tạo cho chúng ta cảm giác chiều cao là một biến ngẫu nhiên rời rạc, nhưng không phải thế, Y thực sự có thể nhận được bất cứ giá trị nào trong khoảng cho trước thí dụ từ 160 cm đến 170 cm tuỳ thuộc vào độ chính xác của phép đo. Y là một biến ngẫu nhiên liên tục.

# BẢNG PHÂN PHỐI XÁC SUẤT

- ĐLNN X (hữu hạn) được biểu diễn thông qua bảng phân phối xác suất sau:

| X | $x_1$ | $x_2$ | …. | $x_i$ | … | $x_n$ |
|---|---|---|---|---|---|---|
| $P(X=x_i)$ | $p_1$ | $p_2$ | …. | $p_i$ | … | $p_n$ |

- Trong đó: $x_i$ (i=1,2,…,n) là các giá trị khác nhau có thể có của X với xác suất tương ứng là $p_i$.

  Kí hiệu: $p_i = P(X=x_i)$.

- Tính chất của xác suất: $0 \leq p_i \leq 1, \sum_i p_i = 1$

# HÀM MẬT ĐỘ XÁC SUẤT

- Hàm mật độ xác suất của ĐLNN liên tục X cho phép đo lường xác suất mà biến ngẫu nhiên X nhận giá trị trong một khoảng nào đó.

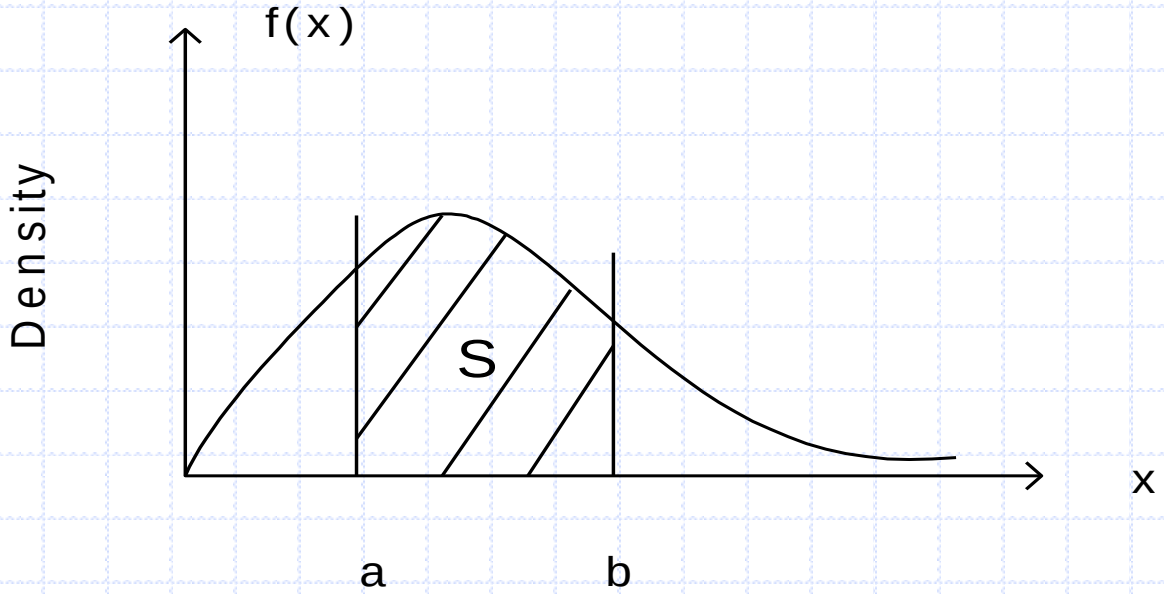
- Hàm mật độ xác suất f(x) có các tính chất sau:

(1) $f(x) \geq 0, \forall x$

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

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

# HÀM MẬT ĐỘ XÁC SUẤT




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

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Giá trị kỳ vọng hay trung bình (Mean) tổng thể:

$$E(X) = \mu_X = \sum_{i=1}^{N} x_i p_i$$ (Nếu X là ĐLNN rời rạc)

$$E(X) = \mu_X = \sum_{x} = xf(x)dx$$ (Nếu X là ĐLNN liên tục)

- Lưu ý: N là số quan sát hay quy mô tổng thể.

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Kỳ vọng có các tính chất sau:

  + E(a) = a với a là hằng số

  + E(a+bX) = a + bE(X) với a và b là hằng số

  + Nếu X và Y là độc lập thì E(XY)= E(X)E(Y)

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

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

- Người ta thường ký hiệu kỳ vọng là $\mu$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Trung bình mẫu (Mean or Average):

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

$$\overline{X} = \sum_{i=1}^{n} X_i P_i$$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Phương sai tổng thể:

$$Var(X) = \sigma_x^2 = E\left[X-\mu_x\right]^2$$

Nếu X là ĐLNN rời rạc thì: $Var(X) = \sigma_X^2 = \sum_x (X-\mu)^2 p_i$

Nếu X là ĐLNN liên tục thì: $Var(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} (X-\mu)^2 f(x)dx$

- Trong tính toán chúng ta sử dụng công thức sau:

**$var(X)=E(X^2)-[E(X)]^2$**

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Các tính chất của phương sai:

+ Var(a) = 0 với a là hằng số.

+ $Var(a+bX) = b^2Var(X)$ với a và b là hằng số.

+ Nếu X và Y là các biến ngẫu nhiên độc lập thì:

$$Var(X+Y) = var(X) + var(Y)$$

$$Var(X-Y) = var(X) + var(Y)$$

+ Nếu X và Y là các biến độc lập, a và b là hằng số thì:

$$Var(aX+bY) = a^2Var(X) + b^2Var(Y)$$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Phương sai mẫu:

$$Var(X) = S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

$$Var(X) = S_X^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2}{n-1}$$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Độ lệch chuẩn tổng thể:(Standard Deviation):

$$SD = \sigma_X = \sqrt{\sigma_X^2}$$

- Độ lệch chuẩn mẫu:(Standard Deviation):

$$SD = S_X = \sqrt{S_X^2}$$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Độ lệch chuẩn của trung bình mẫu ( $\overline{x}$ )

$$S_{\overline{X}} = \frac{S_X}{\sqrt{n}} = \frac{SD}{\sqrt{n}}$$

- Hay còn gọi là sai số chuẩn (Standard Error).

$$SE = S_{\overline{X}} = \frac{S_X}{\sqrt{n}} = \frac{SD}{\sqrt{n}}$$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Hiệp phương sai tổng thể:

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= E(XY) - \mu_X \mu_Y$$

- Hiệp phương sai mẫu:

$$Cov(X,Y) = S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Hệ số tương quan tổng thể:

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

- Hệ số tương quan mẫu:

$$r_{XY} = \frac{S_{XY}}{(S_X)(S_Y)} = \frac{\dfrac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}}{S_X S_Y}$$

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Độ nghiêng tổng thể:

$$Skewness = \frac{\sum_{i=1}^{N} (X_i - \mu_X)^3 / N}{\sigma_X^3} = \frac{E(X_i - \mu_X)^3}{\sigma_X^3}$$

- Độ nghiêng mẫu:

$$Skewness = \frac{n}{(n-1)(n-2)} * \frac{\sum_{i=1}^{n} (X_i - \overline{X})^3}{S_X^3}$$

- Lưu ý: N là số quan sát hay quy mô tổng thể.

# CÁC ĐẶC TRƯNG CỦA ĐLNN X

- Độ nhọn tổng thể:

$$Kurtosis = \frac{\sum_{i=1}^{N} (X_i - \mu_X)^4 / N}{\sigma_X^4} - 3$$

- Lưu ý: N là số quan sát hay quy mô tổng thể.
- Độ nhọn mẫu:

$$Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} * \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{S_X^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

# MỘT SỐ PHÂN PHỐI XÁC SUẤT QUAN TRỌNG

- Phân phối chuẩn.
- Phân phối chuẩn hóa.
- Phân phối t (Student).
- Phân phối chi bình phương.
- Phân phối F (Fisher).

# Phân phối chuẩn

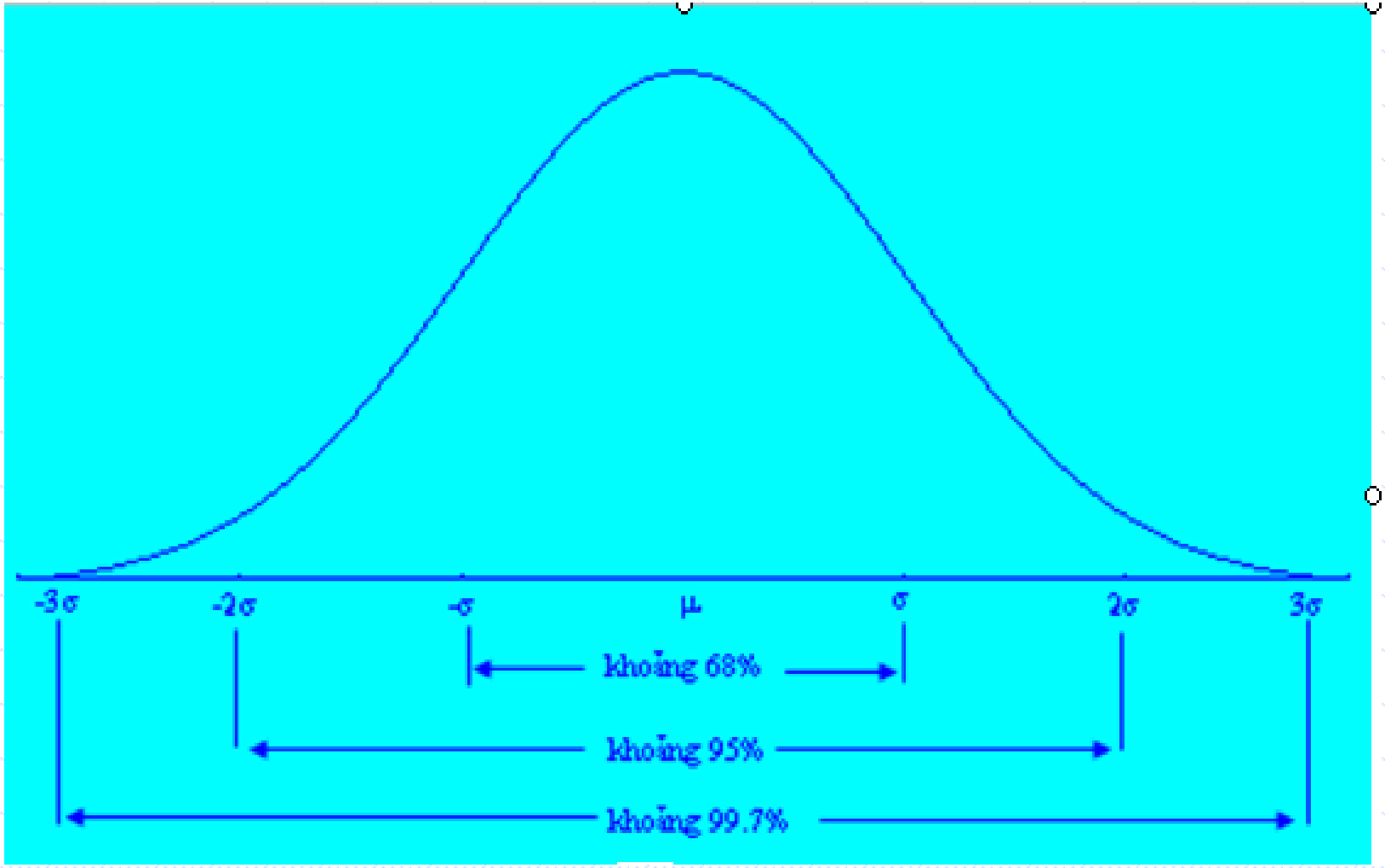
- Khái niệm: Biến ngẫu nhiên X có kỳ vọng là μ, phương sai là $\sigma^2$. Nếu X có phân phối chuẩn thì nó được ký hiệu như sau:

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

- Dạng hàm mật độ xác xuất của phân phối chuẩn như sau:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

# Phân phối chuẩn



# Phân phối chuẩn

- Một kết hợp (hay một hàm) tuyến tính của hai hay nhiều biến ngẫu nhiên theo phân phối chuẩn sẽ theo phân phối chuẩn – đây là một tính chất đặc biệt quan trọng của phân phối chuẩn trong kinh tế lượng.
- Đối với phân phối chuẩn, thì độ nghiên Sk là 0 và độ nhọn K là 3

# Phân phối chuẩn hóa

- Khái niệm: Giả sử đại lượng ngẫu nhiên X phân phối theo quy luật chuẩn với kỳ vọng toán là $\mu$ và phương sai là $\sigma^2$. Xét đại lượng ngẫu nhiên:

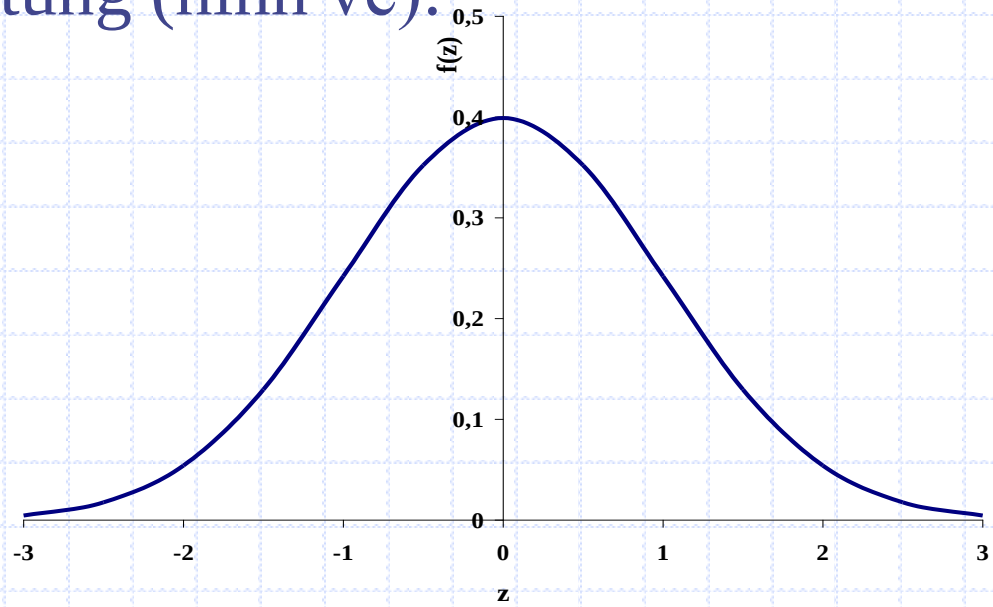
$$Z = \frac{X - \mu}{\sigma}$$

- Đại lượng ngẫu nhiên Z nhận giá trị trong khoảng $(-\infty,+\infty)$ được gọi là phân phối theo quy luật chuẩn hóa nếu hàm mật độ xác suất của Z có dạng:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

# Phân phối chuẩn hóa

- Nếu Z có phân phối chuẩn hóa thì nó được ký hiệu như sau: $Z \sim N(0,1)$
- Đồ thị của hàm f(z) cũng có dạng hình chuông, đối xứng qua trục tung (hình vẽ):



# Phân phối chi bình phương.

- Khái niệm: Nếu $Z_1, Z_2,..., Z_n$ là các biến ngẫu nhiên độc lập có phân phối chuẩn hoá thì:

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2$$

tuân theo phân phối Chi-bình phương với n bậc tự do. Ký hiệu là:

$$\chi^2 \sim \chi^2(n)$$

# Phân phối chi bình phương.

- Tính chất:

  - Phân phối $\chi^2$ là phân phối lệch về bên trái, khi bậc tự do tăng dần thì phân phối $\chi^2$ tiến gần đến phân phối chuẩn.

  - $\chi^2(n_1) + \chi^2(n_2) = \chi^2(n_1+n_2)$, hay tổng của hai biến có phân phối $\chi^2$ cũng có phân phối $\chi^2$ với số bậc tự do bằng tổng các bậc tự do.

# Phân phối t - Student

- Khái niệm: Nếu Z~N(0,1) và $\chi^2(n)$ là độc lập thống kê thì:

$$t_{(n)} = \frac{Z}{\sqrt{\chi^2_n / n}}$$

tuân theo phân phối Student hay nói gọn là phân phối t với n bậc tự do, và được viết dưới dạng t ~ t(n).

- ***Tính chất:*** Phân phối t với n bậc tự do d.f. có những tính chất sau:

  - Phân phối t là đối xứng qua gốc tọa độ và có hình dạng tương tự như trong phân phối chuẩn hóa.

  - Đối với ni lớn, phân phối t tuân theo một cách gần đúng với phân phối N(0,1). Sự gần đúng là tương đối tốt ngay cả với n =30.

# Phân phối F (Fisher).

- Khái niệm: Nếu $\chi^2(n_1)$ và $\chi^2(n_2)$ là độc lập thống kê thì:

$$F_{(n1,n2)} = \frac{\chi^2_{n1} / n_1}{\chi^2_{n2} / n_2}$$

tuân theo phân phối F với $(n_1,n_2)$ bậc tự do, và được viết dưới dạng $F \sim F(n_1,n_2)$. Với $n_1$ là bậc tự do của tử số, $n_2$ là bậc tự do của mẫu số.

- ***Tính chất :*** Phân phối F với $n_1$ và $n_2$ bậc tự do d.f. có những tính chất sau:
  - Phân phối F có hình dạng tương tự như trong phân phối chi bình phương.
  - Nếu biến ngẫu nhiên t có phân phối Student với bậc tự do d.f. n thì $t^2$ có phân phối F với bậc tự do d.f. là 1 và n. Do vậy, $t^2(n) \sim F(1,n)$.