Giới thiệu môn học

- Tên môn học: Lý thuyết Thống kê (Mathematical Statistics).
- Thời lượng: 45 tiết.
- Thời gian: sáng thứ hai (tiết 1 - 3), P. D209.
- Giảng viên: TS. Hoàng Văn Hà.
  Email: hvha@hcmus.edu.vn.
  Web: https://sites.google.com/view/hoangvanha.
- Nội dung môn học: xem đề cương chi tiết.
- Đánh giá:
  - Giữa kỳ và cuối kỳ: 30% GK + 50% CK.
  - Bài tập: 15 % Quiz bài tập.
  - 5% chuyên cần.

Lý thuyết thống kê
Chương 1: Thống kê mô tả

Hoàng Văn Hà
University of Science, VNU - HCM
hvha@hcmus.edu.vn

Mục lục

Một số khái niệm cơ bản

- **Tổng thể (population):** tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- **Mẫu (sample):** là một tập con được chọn ra từ tổng thể.
- **Tham số (parameter):** là một đặc trưng cụ thể của một tổng thể. Ví dụ: trung bình (kỳ vọng), phương sai, trung vị, ...
- **Thống kê (statistic):** là một đặc trưng cụ thể của một mẫu. Ví dụ: trung bình mẫu, phương sai mẫu, trung vị mẫu, ...

Ví dụ về tổng thể:

- Số cử tri đăng ký đi bầu cử
- Thu nhập của các hộ gia đình trong thành phố
- Điểm trung bình của tất cả các sinh viên trong một trường đại học
- Trọng lượng của các sản phẩm trong một nhà máy

Thông thường, ta không thể chọn hết được tất cả các phần tử của tổng thể để nghiên cứu bởi vì:

- số phần tử của tổng thể rất lớn,
- thời gian và kinh phí không cho phép,
- có thể làm hư hại các phần tử của tổng thể.

Do đó, ta chỉ thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ một tổng thể phải thỏa các điều kiện sau:

- Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).
- Mọi mẫu cỡ n cũng có cùng khả năng được chọn từ tổng thể.

Phương pháp chọn mẫu ngẫu nhiên đơn giản (simple random sampling):

- Đánh số các phần tử của tổng thể từ 1 đến N. Lập các phiếu cũng đánh số như vậy.
- Trộn đều các phiếu, sau đó chọn có hoàn lại n phiếu. Các phần tử của tổng thể có số thứ tự trong phiếu lấy ra sẽ được chọn làm mẫu.

Mô tả phân phối của dữ liệu

Các dạng đồ thị:

- Histogram (đồ thị tổ chức tần số)
- Khi cỡ mẫu nhỏ:
  - Stem-and-Leaf (đồ thị thân và lá)
  - Dotplot

Histogram

- Histogram được xây dựng dựa trên bảng phân bố tần số (frequency distribution).
- Một bảng phân bố tần số bao gồm:
  - các khoảng được phân nhóm theo dữ liệu quan trắc (observations),
  - và các tần số tương ứng của dữ liệu nằm bên trong từng khoảng.
- Histogram cho phép:
  - mô tả phân phối của dữ liệu,
  - nhận dạng phân phối chuẩn (bell-shaped),
  - xem xét tính đối xứng/bất đối xứng, tập trung/phân tán của dữ liệu,
  - xác định mode (unimodal, bimodal),
  - ...

Lập một bảng phân bố tần số

Trong một bảng phân bố tần số:

- mỗi nhóm có bề rộng bằng nhau,
- bề rộng của mỗi nhóm được xác định bởi

$$\frac{\text{Giá trị lớn nhất} - \text{Giá trị bé nhất}}{\text{Số khoảng cần chia}},$$

- các khoảng không trùng nhau,
- nên chọn số khoảng tối thiểu ≥ 5.

Lập một bảng phân bố tần số

Ví dụ 1

Chọn ngẫu nhiên 20 ngày mùa đông có nhiệt độ cao và đo nhiệt độ (Đv: độ F) được số liệu như sau

24	*35*	*17*	*21*	*24*	*37*	*26*	*46*	*58*	*30*
32	*13*	*12*	*38*	*41*	*43*	*44*	*27*	*53*	*27*

Hãy lập bảng phân bố tần số và vẽ histogram cho tập dữ liệu này.

Lập một bảng phân bố tần số

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần

$$12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$$

- Xác định khoảng biến thiên của dữ liệu (range): $58 - 12 = 46$.
- Chọn số khoảng cần chia: 5.
- Xác định độ rộng của khoảng: 10 (làm tròn 46/5).
- Xác định biên của các khoảng: từ 10 đến dưới 20, từ 20 đến dưới 30, . . . , từ 50 đến dưới 60.
- Đếm số giá trị dữ liệu nằm trong mỗi khoảng.

Lập một bảng phân bố tần số

Dữ liệu được sắp xếp theo thứ tự tăng dần:

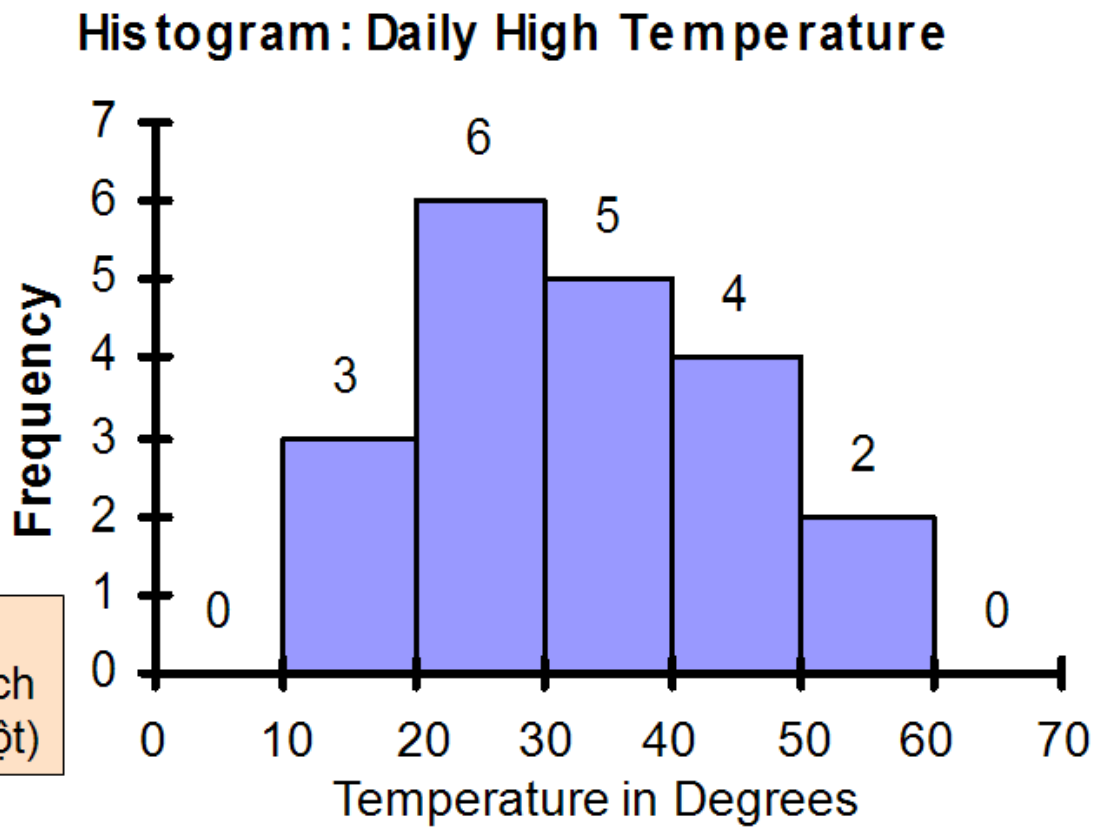
$$12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$$

Khoảng	Tần số	Tần suất	Phần trăm
[10,20)	3	0.15	15
[20,30)	6	0.30	30
[30,40)	5	0.25	25
[40,50)	4	0.20	20
[50,60)	2	0.10	10
Tổng	20	1.00	100

Vẽ histogram

Khoảng	Tần số
[10, 20)	3
[20, 30)	6
[30, 40)	5
[40, 50)	4
[50, 60)	2

(Không có khoảng cách giữa các cột)

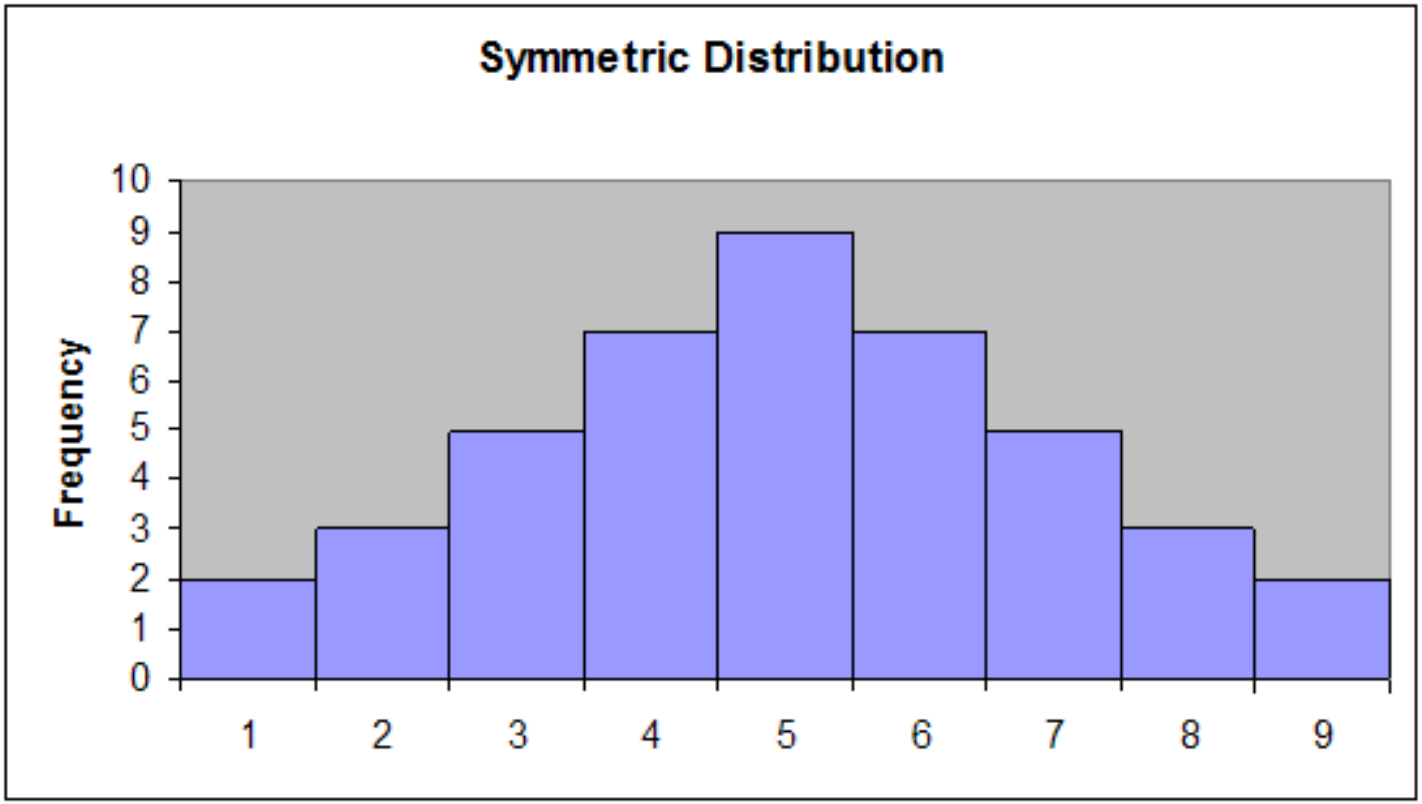


Cần chọn bao nhiêu khoảng khi vẽ histogram

- Không có câu trả lời cụ thể. Thông thường số khoảng cần chia sẽ phụ thuộc vào cỡ mẫu.
- Một số quy tắc:
  - Quy tắc của Sturge: số khoảng $= 1 + \log_2(n)$.
  - Quy tắc của Rice: số khoảng $= 2\sqrt[3]{n}$.

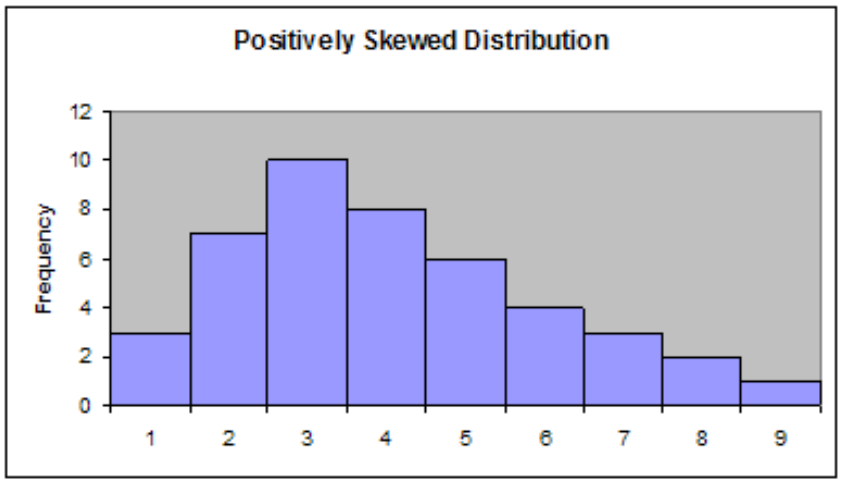
Nhận dạng phân phối của dữ liệu

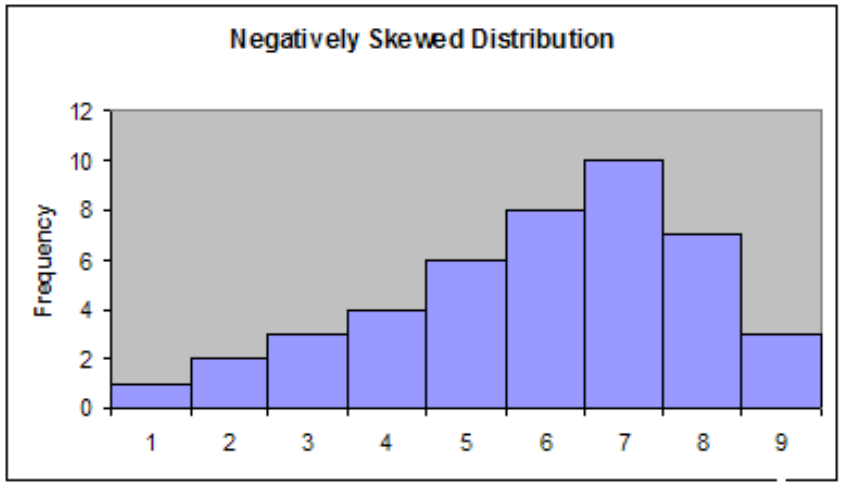
- Đối xứng:



Nhận dạng phân phối của dữ liệu

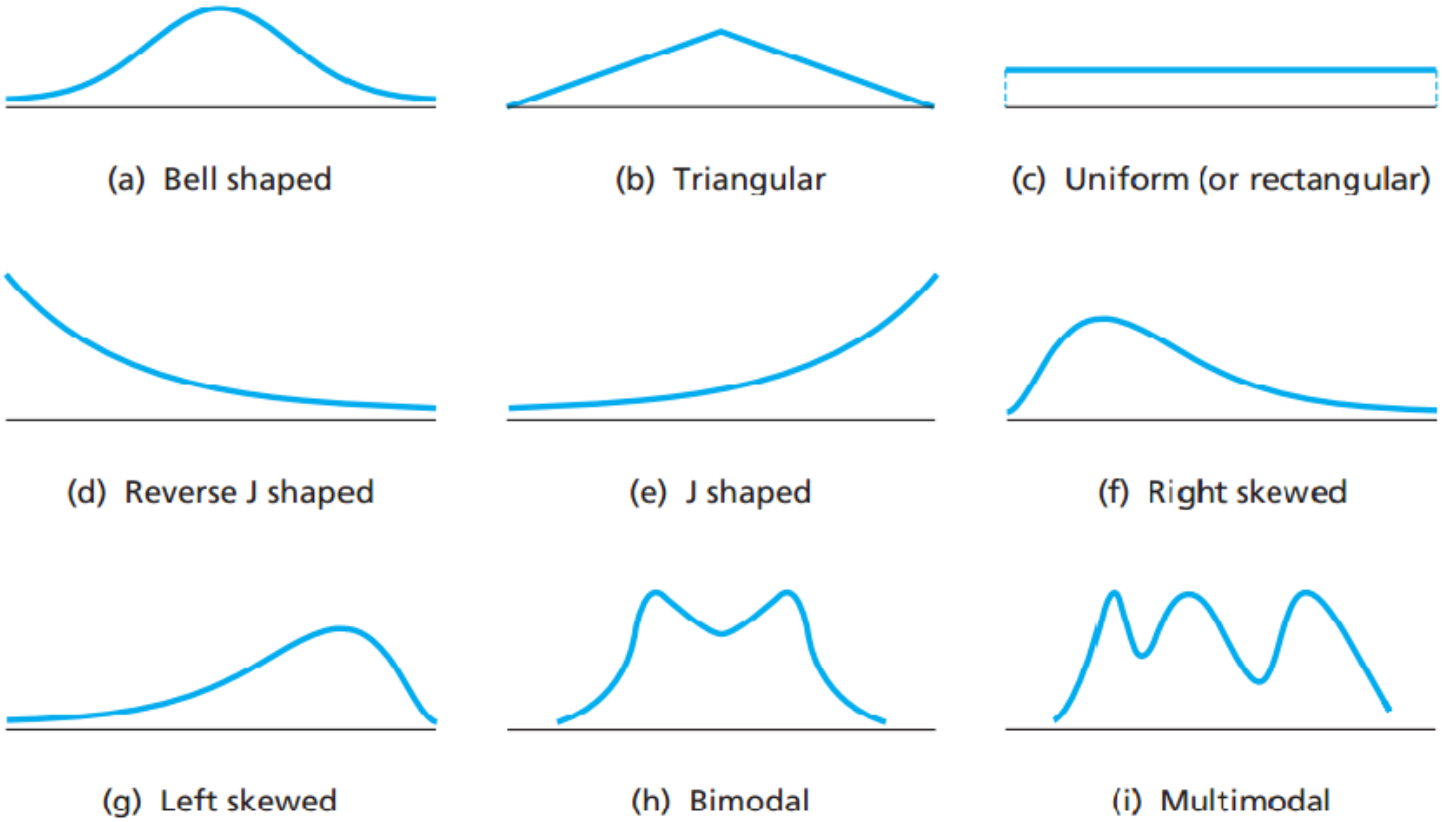
- Bất đối xứng (lệch trái và lệch phải):





Nhận dạng phân phối của dữ liệu

- Các dạng khác:



Đồ thị thân và lá

Ví dụ 2

Bộ dữ liệu sau mô tả kết quả thi môn Toán (thang điểm 100) của 20 sinh viên trong một lớp học.

72	49	62	58	73	55	78	83	57	63
73	73	75	85	85	64	61	67	75	91

Vẽ đồ thị thân và lá cho bộ dữ liệu trên.

Ví dụ 3

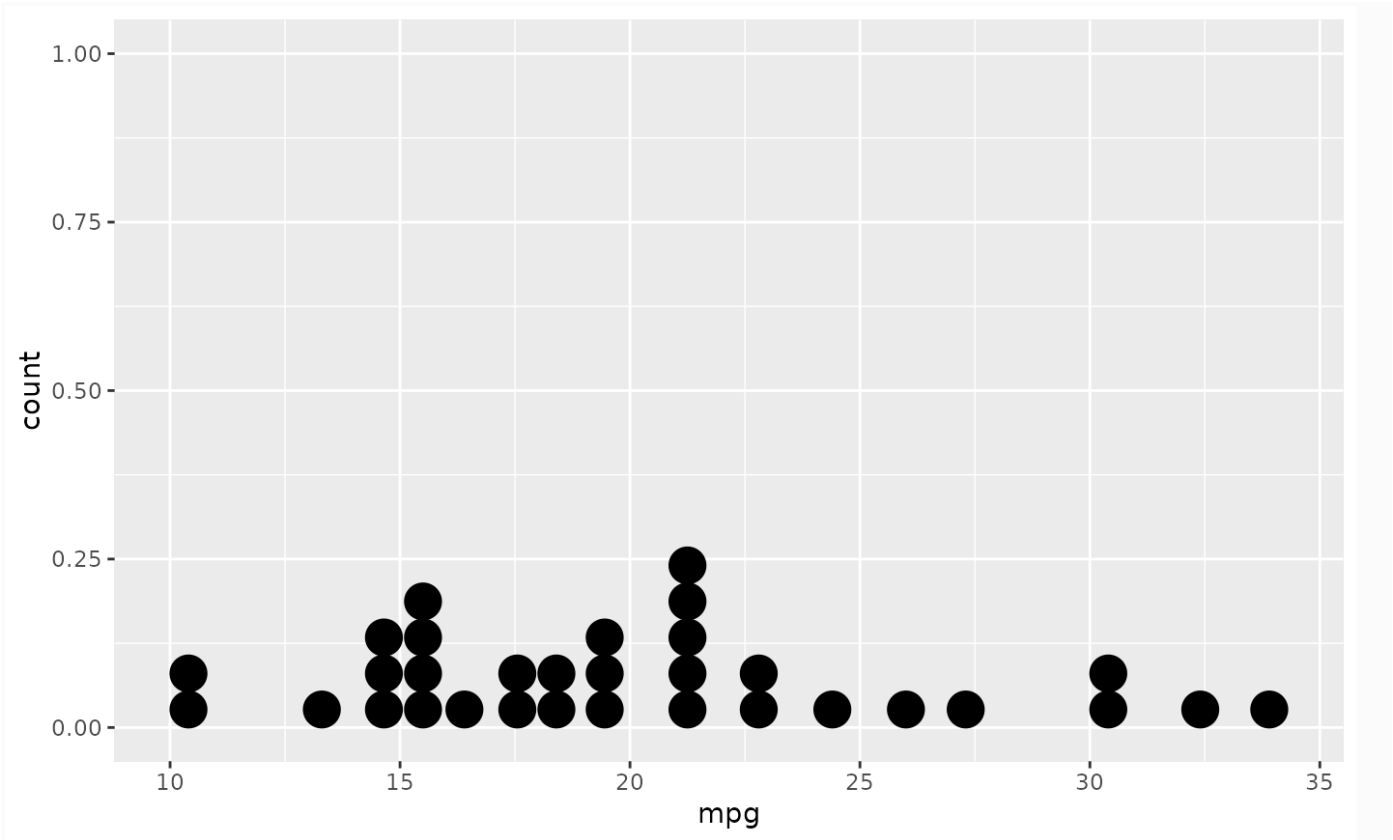
Bộ dữ liệu dưới đây cho biết kết quả của thi môn bật xa (Đv: m) của 10 sinh viên trong môn học giáo dục thể chất:

2.3	2.5	2.5	2.7	2.8	3.2	3.6	3.6	4.5	5.0

Vẽ đồ thị thân và lá cho bộ dữ liệu trên.

Dotplot

```
ggplot(mtcars, aes(x = mpg)) + geom_dotplot()
```



Độ đo xu hướng trung tâm (central tendency)

Gồm các đại lượng sau:

- Trung bình (mean/average)
- Trung vị (median)
- Yếu vị (mode)

Ta cũng có thể có:

- Trimean
- Truncated mean (trimmed mean)

Trung bình

- Trung bình là đại lượng thường được sử dụng nhất để đo giá trị trung tâm của dữ liệu.
- Với một tổng thể có N phần tử (thông thường, N rất lớn), trung bình tổng thể (population mean) được tính bởi

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}.$$

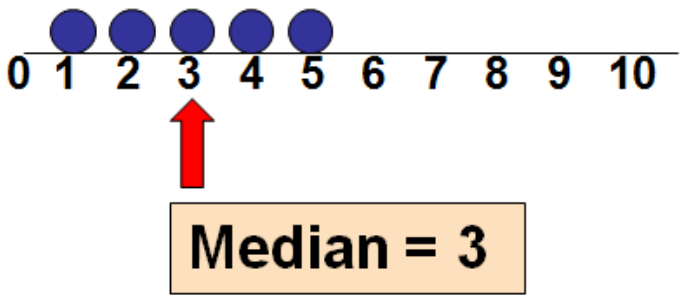
- Với một mẫu cỡ n được chọn từ tổng thể, trung bình mẫu (sample mean) được tính bởi

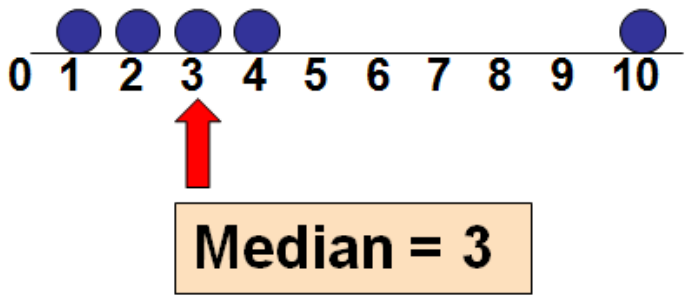
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Chú ý: trung bình rất nhạy cảm với các giá trị ngoại lai (outlier).

Trung vị

- Trong một tập dữ liệu được sắp xếp theo thứ tự tăng dần, trung vị (median) là giá trị "chính giữa" của dữ liệu (50% bên trái, 50% bên phải).
- Trung vị không bị ảnh hưởng bởi các giá trị ngoại lai.





Trung vị

Xác định trung vị:

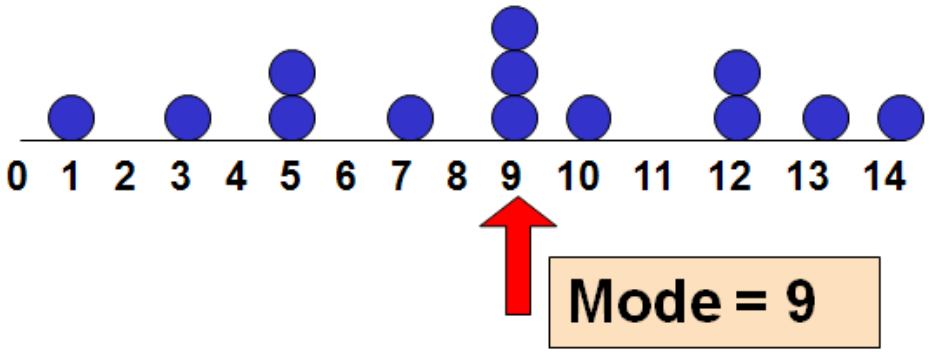
- Ký hiệu $\mathrm{med}(x)$ là trung vị của véc-tơ $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.
- Trung vị được xác định bởi

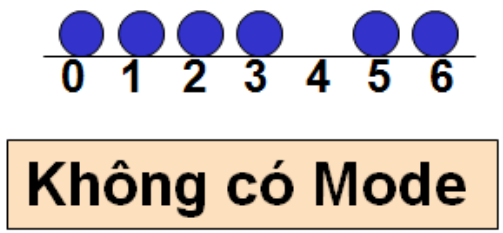
$$\mathrm{med}(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{nếu } n \text{ lẻ} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{nếu } n \text{ chẵn} \end{cases}$$

với $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ là thống kê thứ tự (order statistic).

Mode (Yếu vị)

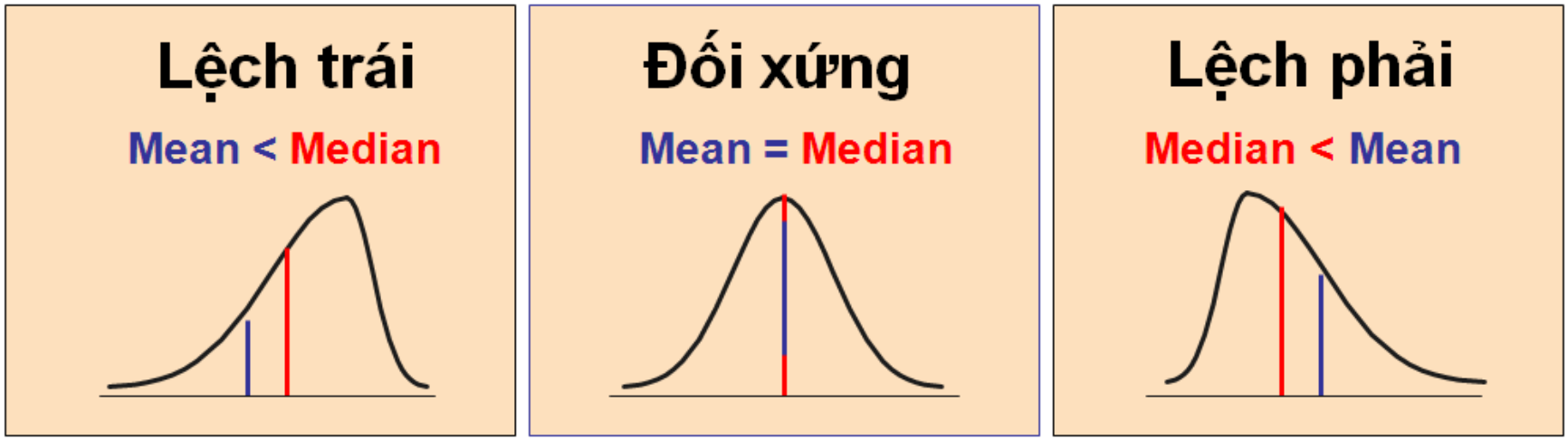
- Là giá trị thường xảy ra nhất,
- Không bị ảnh hưởng bởi các điểm ngoại lai,
- Có thể sử dụng cho cả dữ liệu định tính và dữ liệu định lượng,
- Có thể có nhiều mode hoặc không tồn tại mode.



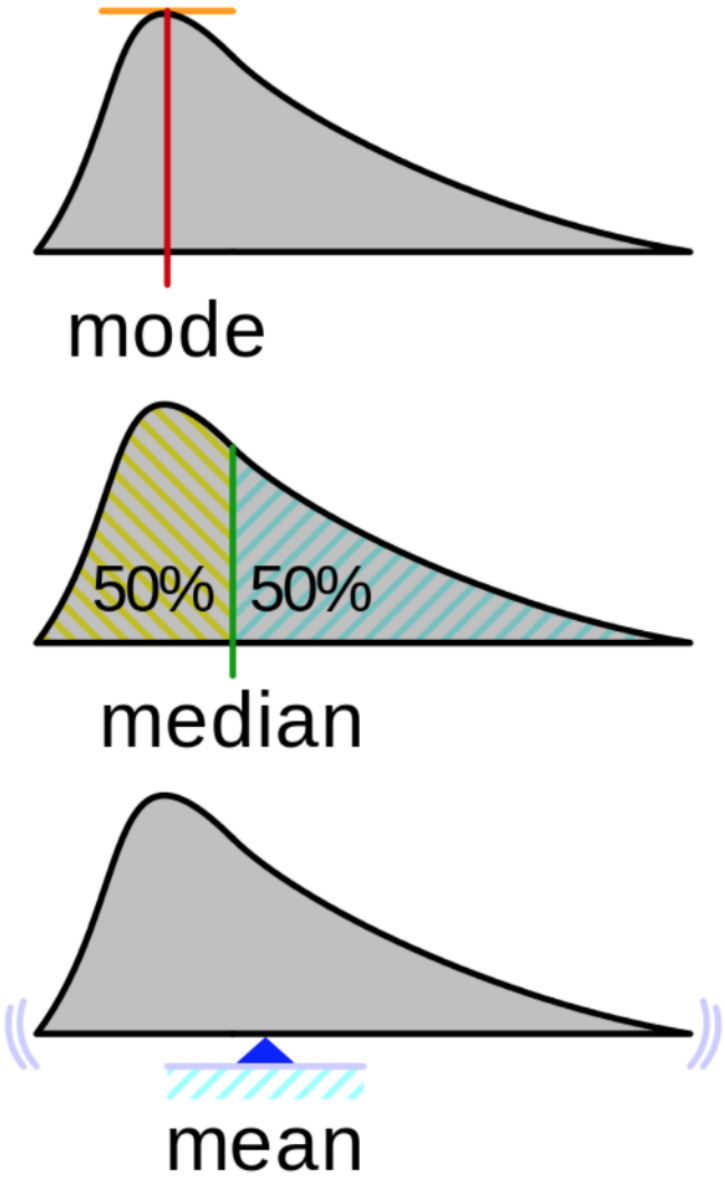


Sử dụng các độ đo xu hướng trung tâm

- Trung bình luôn luôn được sử dụng, nếu các điểm ngoại lai (outliers) không tồn tại hoặc sau khi loại bỏ các điểm ngoại lai.
- Trung vị thường được dùng nếu bộ dữ liệu có các điểm ngoại lai hoặc rất bất đối xứng.
- Mode thường dùng để mô tả các biến định tính.
- Vị trí của trung bình và trung vị bị ảnh hưởng bởi phân phối của dữ liệu:



Trung bình, trung vị và mode



Độ đo sự biến thiên (variability)

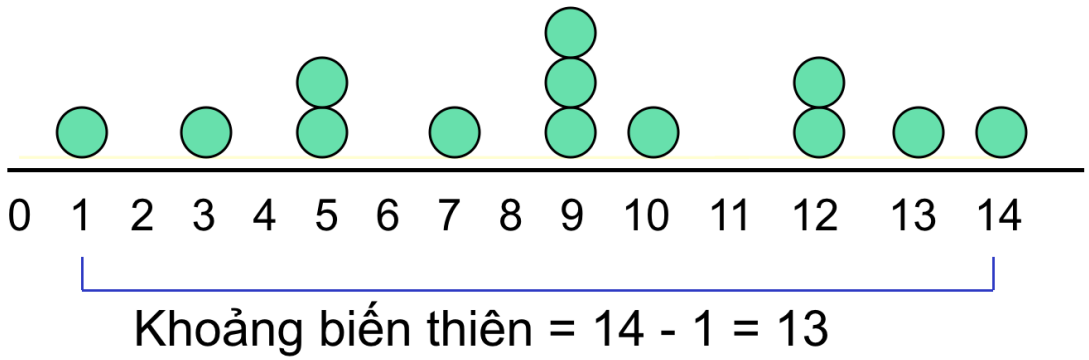
Gồm các độ đo sau:

- Khoảng biến thiên (range)
- Khoảng tứ phân vị (interquartile range)
- Phương sai (variance)
- Độ lệch tiêu chuẩn (Standard deviation)

Khoảng biến thiên

- Khoảng biến thiên (range) là độ đo sự biến thiên đơn giản nhất.
- Là độ chênh lệch giữa giá trị lớn nhất và bé nhất của dữ liệu quan trắc

$$\text{Khoảng biến thiên} = X_{Max} - X_{Min}.$$



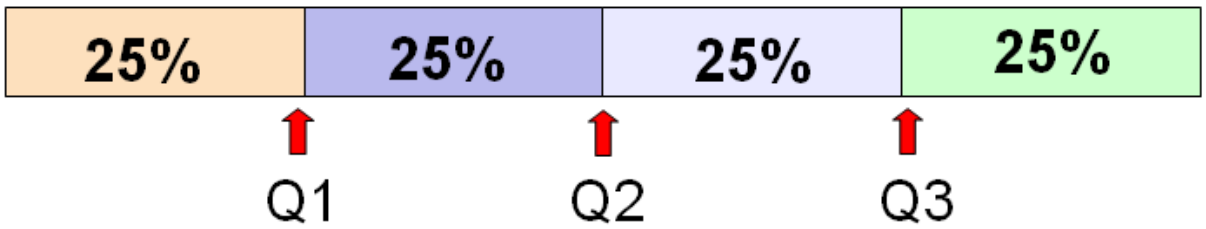
- Hạn chế:
  - Bỏ qua sự phân bố của dữ liệu.
  - Dễ bị ảnh hưởng bởi các điểm ngoại lai (outlier).

Khoảng tứ phân vị

- Khoảng tứ phân vị (interquartile range):

$$\mathrm{IQR} = Q_3 - Q_1,$$

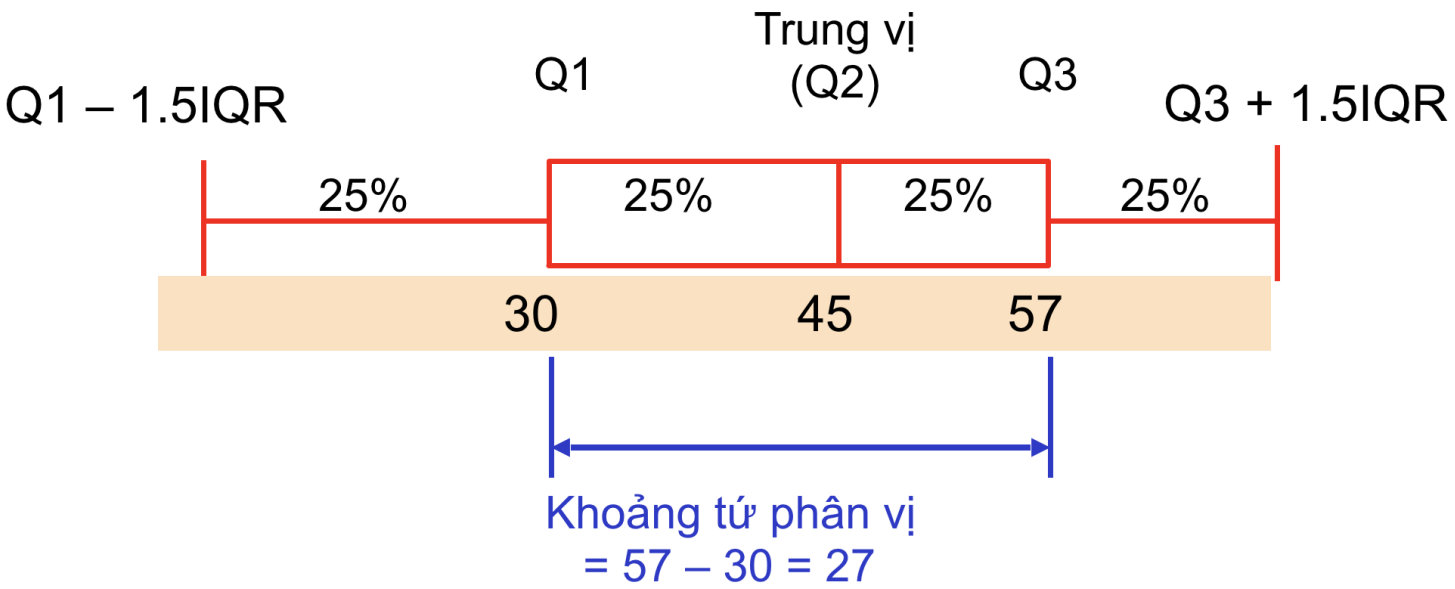
  với Q_1 là phân vị thứ 1 (mức 25%) và Q_3 là phân vị thứ 3 (mức 75%) của dữ liệu.
- Các điểm Q_1, Q_2, và Q_3 được gọi là các điểm tứ phân vị:



- Cách tìm Q_1 và Q_3: tương tự Q_2 (trung vị).

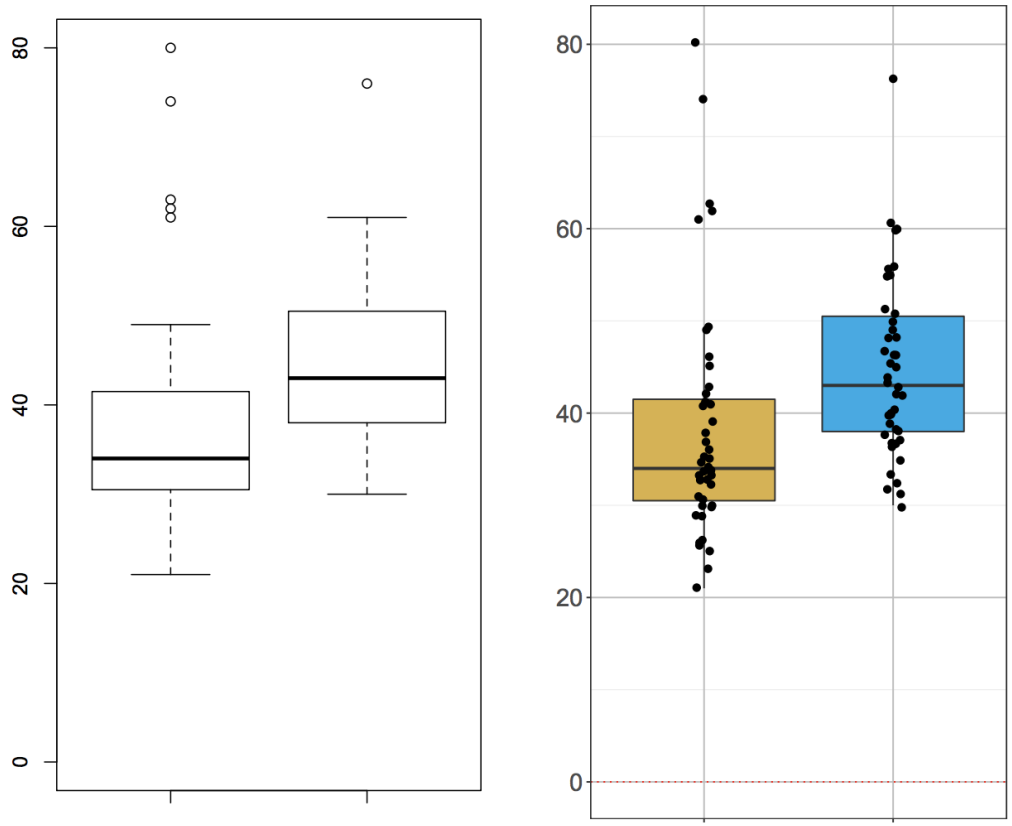
Boxplot

- Để biểu diễn khoảng tứ phân vị và các điểm ngoại lai : sử dụng boxplot.



Boxplot

- Khi vẽ nhiều đồ thị boxplot của nhiều tập dữ liệu khác nhau bên cạnh nhau, ta còn có thể so sánh được độ phân tán và so sánh giá trị trung tâm (trung bình/trung vị) của các tập dữ liệu này.



Phương sai

- Phương sai (variance) là trung bình của bình phương độ lệch các giá trị so với trung bình.
- Phương sai phản ánh độ phân tán hay sự biến thiên của dữ liệu.
- Phương sai tổng thể (popluation variance):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

với N là số phần tử của tổng thể, μ là trung bình tổng thể, x_i là giá trị thứ i của biến x.

- Phương sai mẫu (sample variance):

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

với $\bar{x}$ là trung bình mẫu, n là cỡ mẫu, x_i là giá trị quan trắc thứ i.

Độ lệch tiêu chuẩn

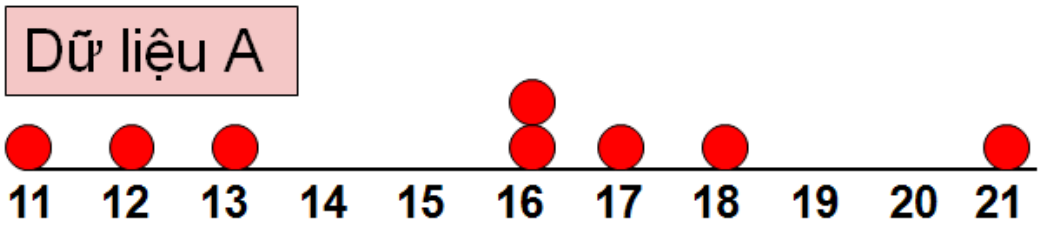
- Độ lệch tiêu chuẩn (standard deviation) được dùng để đo sự biến thiên, biểu diễn sự biến thiên xung quanh trung bình.
- Có cùng đơn vị đo với dữ liệu gốc.
- Độ lệch chuẩn của tổng thể, ký hiệu là σ:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}.$$

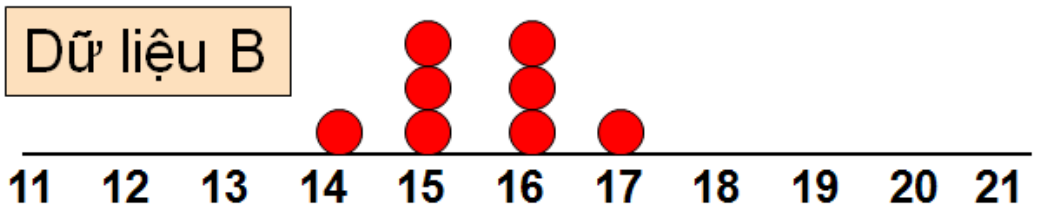
- Độ lệch chuẩn của mẫu,

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

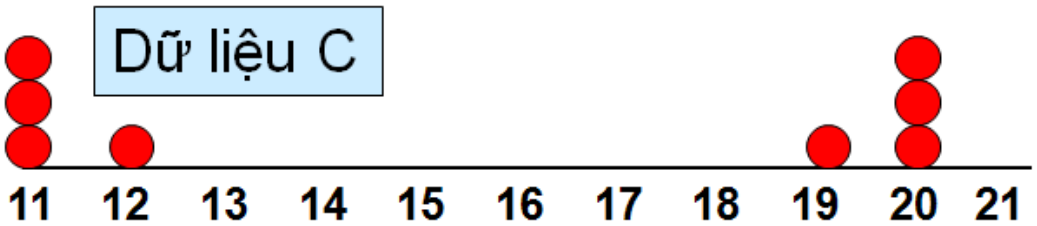
So sánh sự biến thiên của dữ liệu dùng độ lệch chuẩn



Mean = 15.5
s = 3.338



Mean = 15.5
s = 0.926



Mean = 15.5
s = 4.570

Hệ số biến thiên

- Hệ số biến thiên (Coefficient of Variation) được sử dụng để so sánh sự biến thiên của hai hay nhiều tập dữ liệu, có thể đo ở các đơn vị khác nhau.
- Hệ số biến thiên do sự phân tán tương đối của dữ liệu xung quanh giá trị trung bình.
- Đơn vị tính bằng %.
- Công thức

$$CV = \frac{s}{\bar{x}} 100\%.$$

So sánh hệ số biến thiên

- Dữ liệu A:
  - Trung bình $\bar{x}_A = 50$
  - Độ lệch chuẩn $s_A = 5$

$$CV_A = \frac{s_A}{\bar{x}_A}100\% = \frac{5}{50}100\% = 10\%.$$

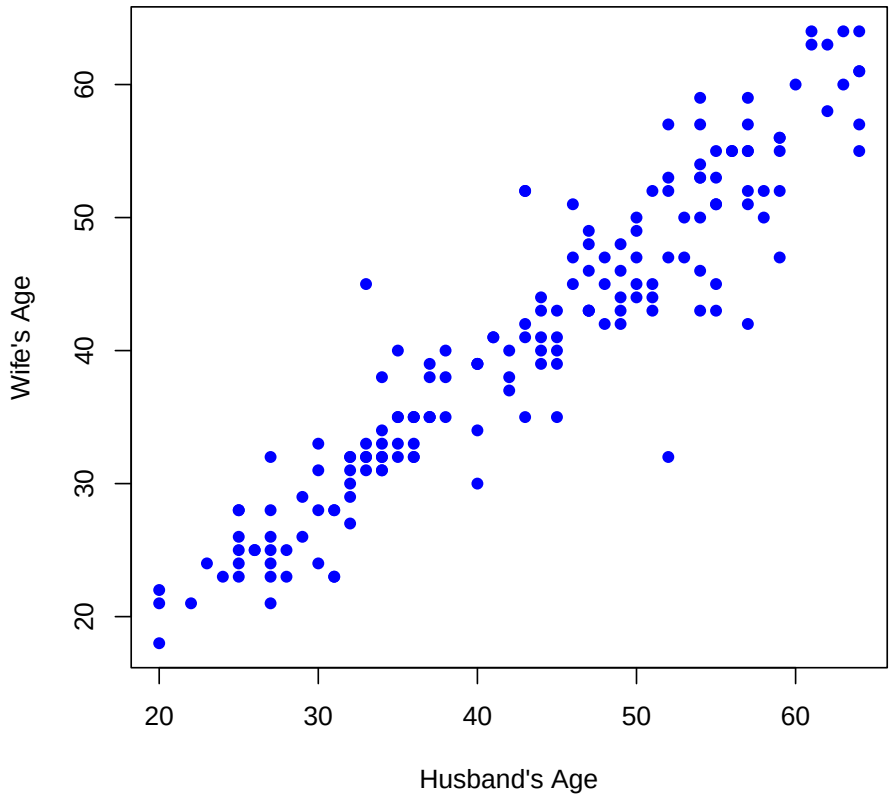
- Dữ liệu B:
  - Trung bình $\bar{x}_B = 100$
  - Độ lệch chuẩn $s_B = 5$

$$CV_B = \frac{s_B}{\bar{x}_B}100\% = \frac{5}{100}100\% = 5\%.$$

- Cả hai tập dữ liệu có cùng độ lệch chuẩn, nhưng dữ liệu B biến thiên ít hơn so với giá trị của nó.

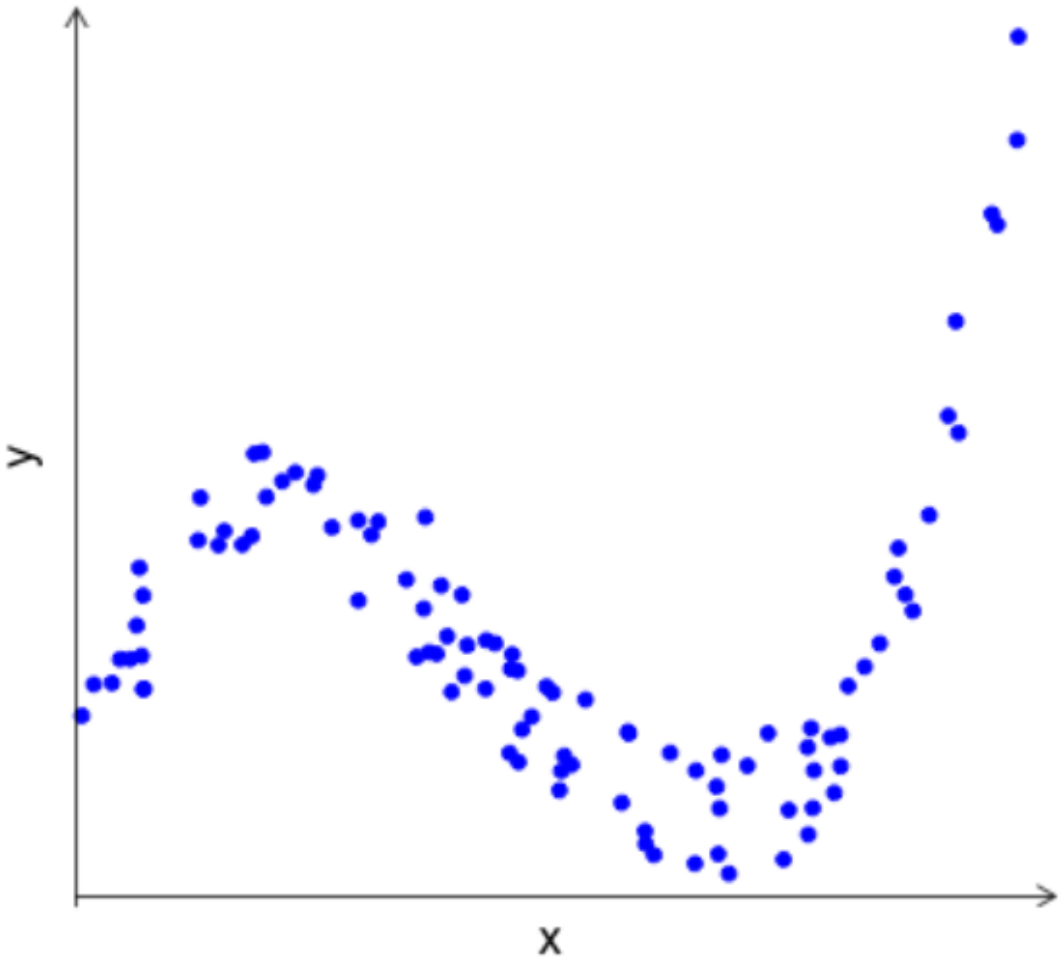
Đồ thị phân tán

- Đồ thị phân tán (scatter plot) dùng để mô tả mối quan hệ giữa hai biến.
- Ví dụ: đồ thị phân tán mô tả tuổi của 199 cặp vợ chồng.



Câu hỏi: người ta có xu hướng kết hôn với những người có cùng độ tuổi hay không?

- Quan hệ phi tuyến (non-linear relationship):



Hệ số tương quan Pearson

- Hệ số tương quan Pearson (Pearson's corelation coefficient) là một độ đo thống kê dùng để đo mối quan hệ tuyến tính giữa hai biến ngẫu nhiên thực.
- Hệ số tương quan tổng thể:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}}.$$

- Hệ số tương quan mẫu:

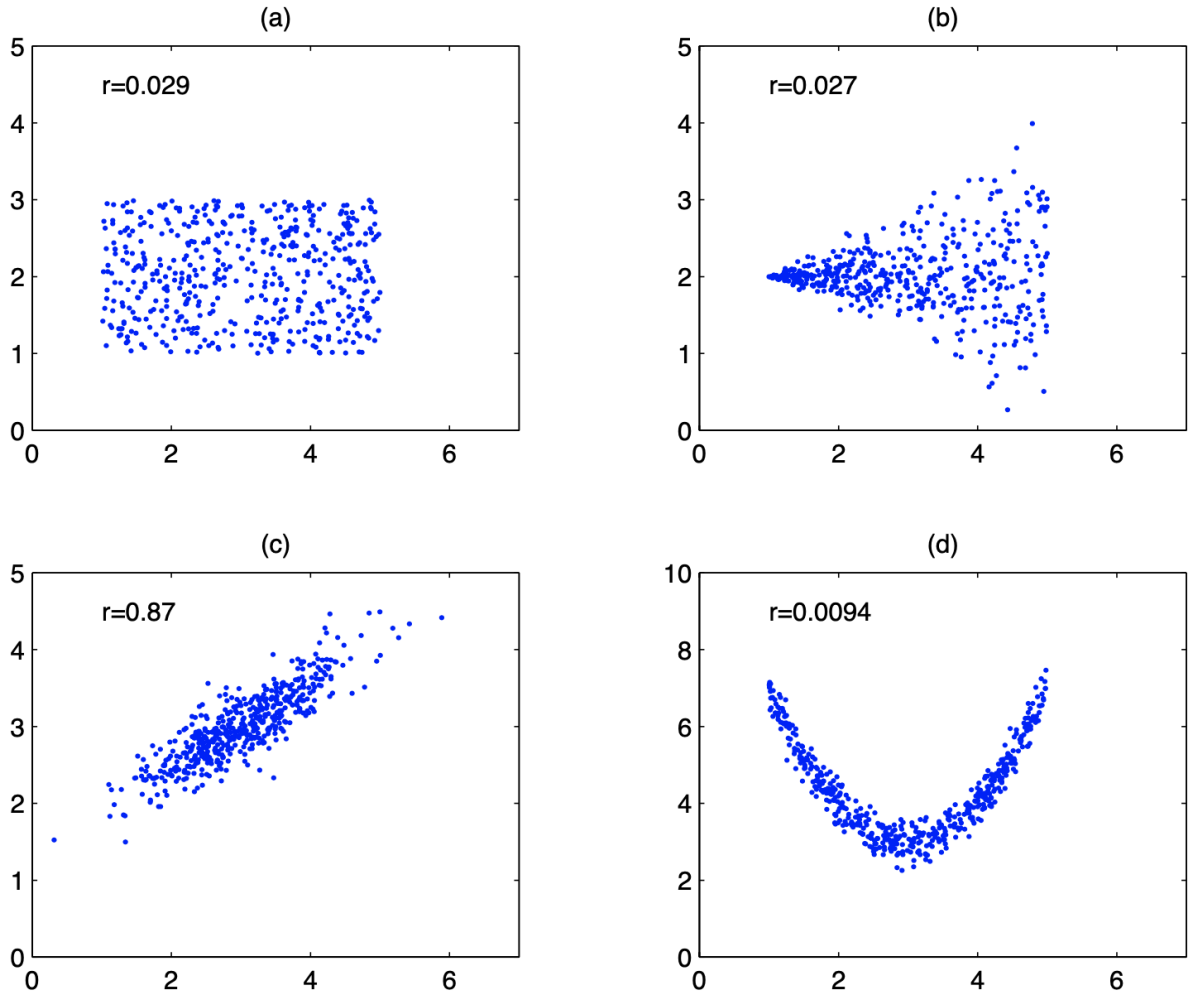
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Hệ số tương quan Pearson

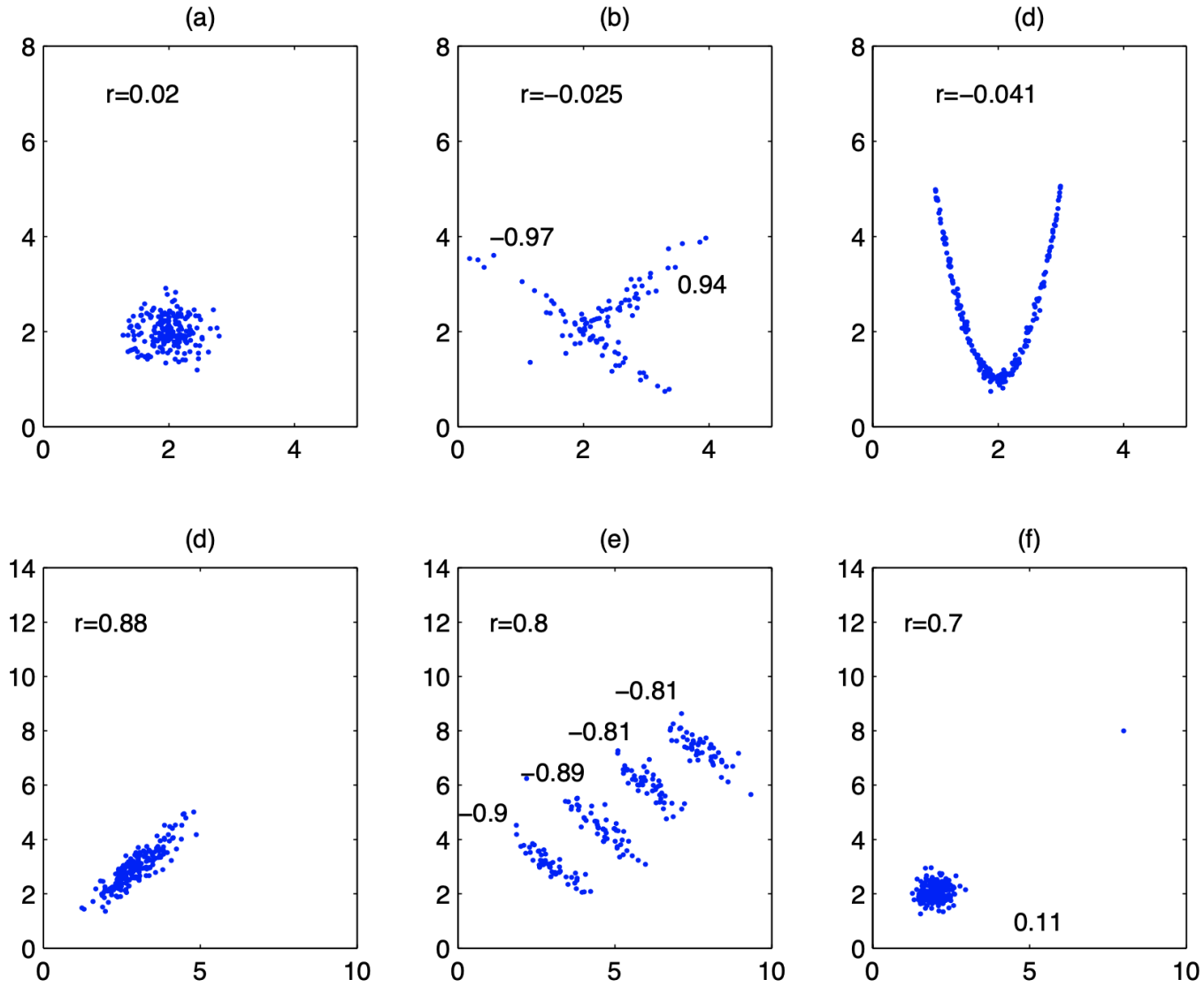
- $-1 \leq r \leq 1$.
- $r > 0$: tương quan tuyến tính thuận.
- $r < 0$: tương quan tuyến tính nghịch.
- $r = 0$: không có tương quan tuyến tính.
- r càng gần 1 hoặc -1, thì mối quan hệ tuyến tính càng mạnh.

Chú ý: $r = 0$ (hoặc $\rho = 0$) suy ra X và Y không có mối quan hệ tuyến tính nhưng không có nghĩa là X và Y độc lập, có thể tồn tại mối quan hệ phi tuyến giữa hai biến.

Hệ số tương quan Pearson

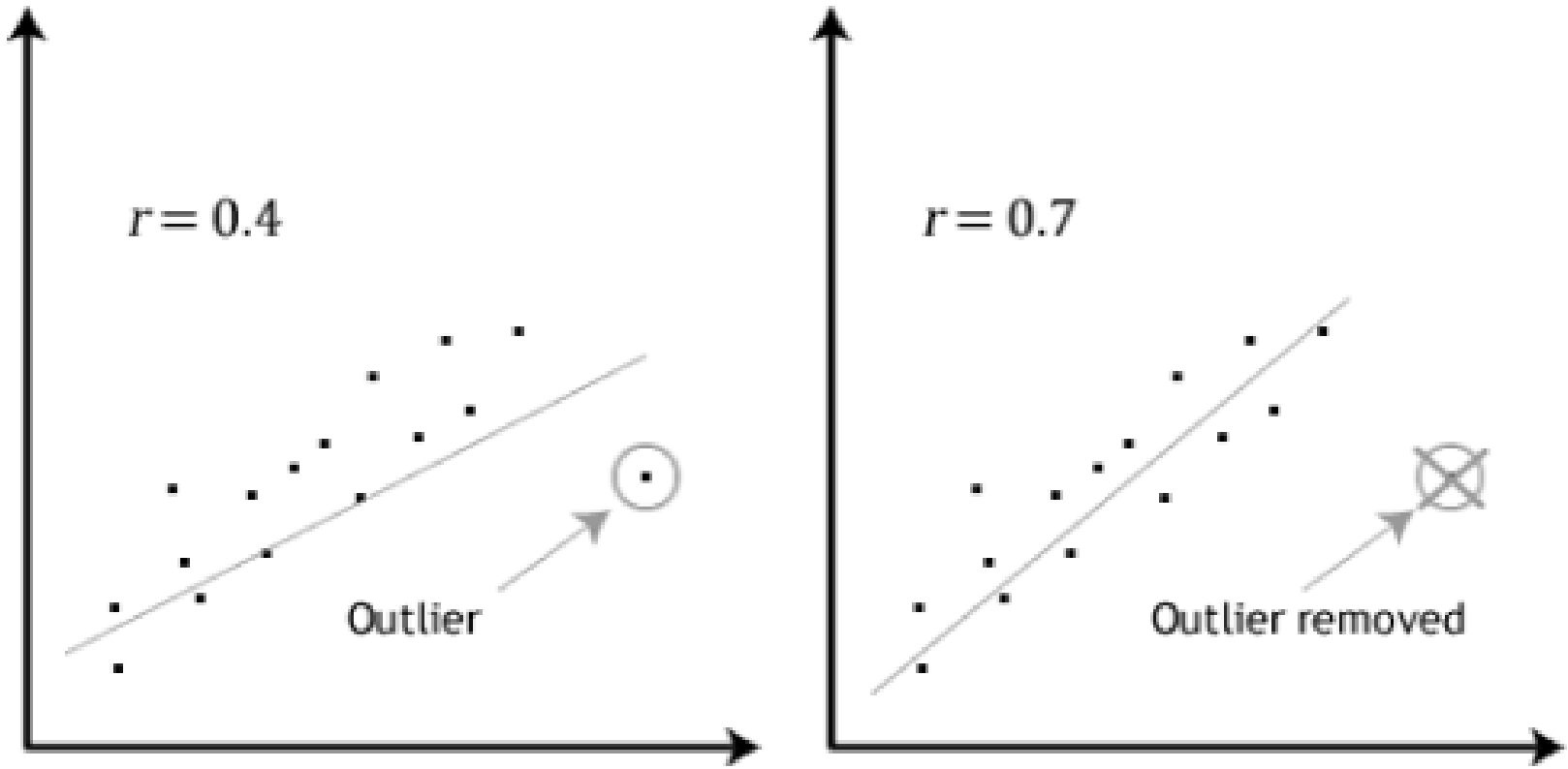


Hệ số tương quan Pearson



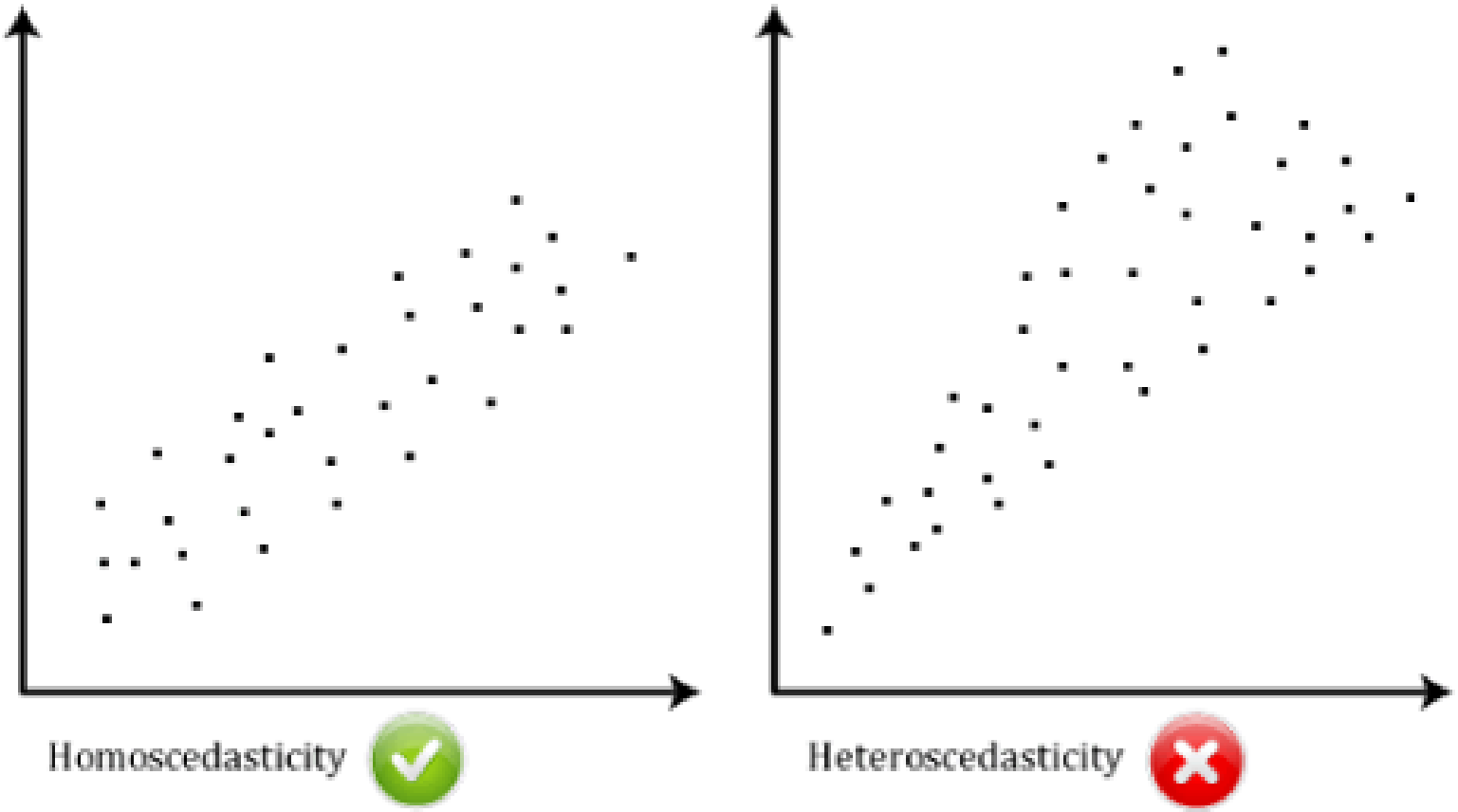
Các yếu tố ảnh hưởng đến hệ số tương quan

- Các điểm ngoại lai (outliers):



Các yếu tố ảnh hưởng đến hệ số tương quan

- Tính đồng nhất (homoscedasticity) và không đồng nhất (heteroscedasticity) của dữ liệu:



Chú ý!

- Sự tồn tại mối tương quan mạnh không có nghĩa là có một liên hệ nhân quả (causal link) giữa các biến.
- Ta cần thực hiện một kiểm định có ý nghĩa (significance test) để quyết định xem liệu với một mẫu cho trước, có đủ bằng chứng để kết luận rằng có mối tương quan tuyến tính hiện diện trong tổng thể hay không?

Phân phối chuẩn

Định nghĩa 1 (Normal distribution)

Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(-\infty, +\infty)$ được gọi là có phân phối chuẩn tham số μ, σ nếu hàm mật độ xác suất có dạng

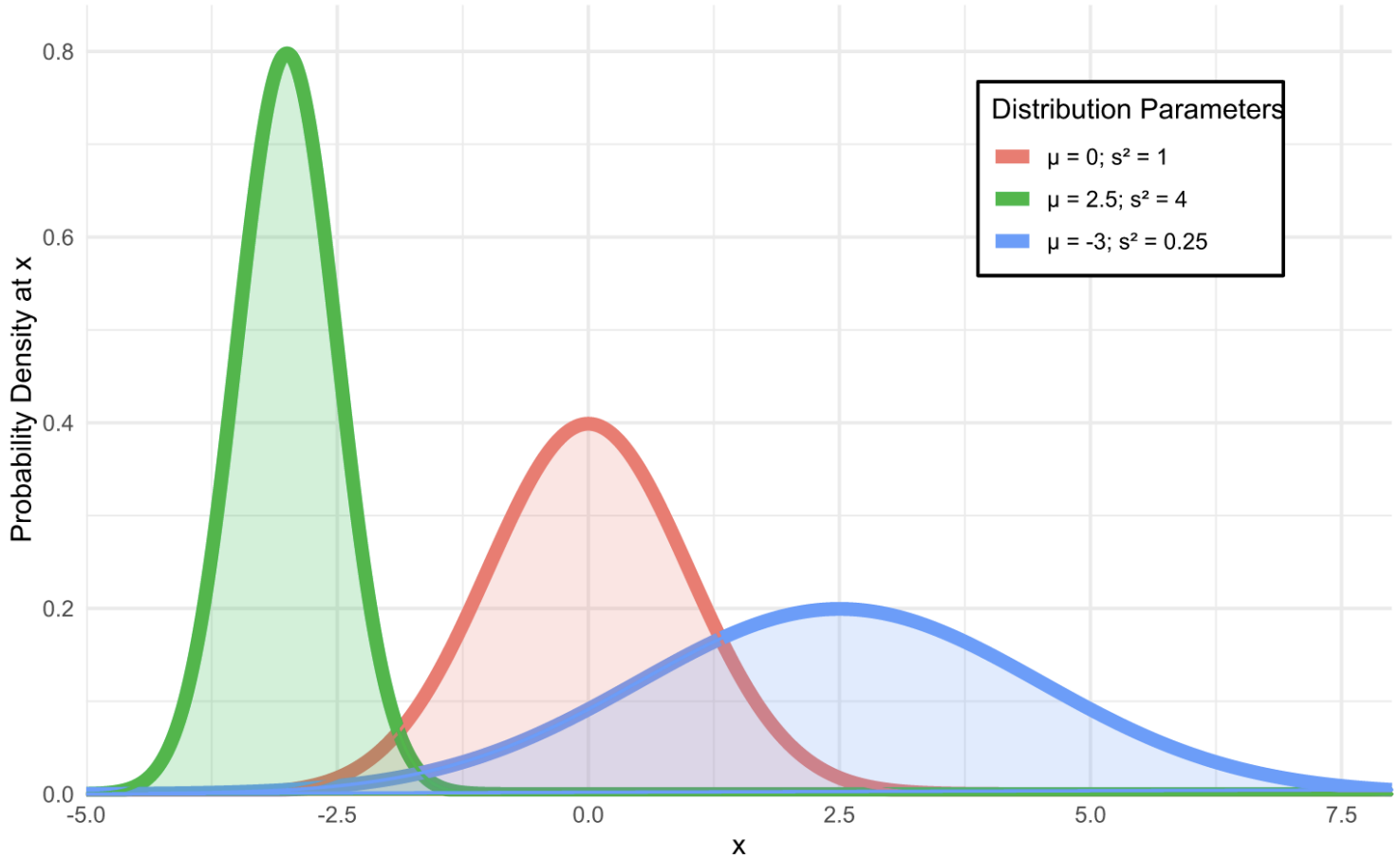
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad -\infty < x < +\infty$$

trong đó μ, σ là hằng số và $\sigma > 0$, $-\infty < \mu < +\infty$, ký hiệu $X \sim \mathcal{N}\left(\mu; \sigma^2\right)$.

Nếu $X \sim \mathcal{N}(\mu, \sigma^2)$, ta có

$$\mathbb{E}(X) = \mu$$
$$\mathbb{V}ar(X) = \sigma^2$$

Phân phối chuẩn



Phân phối chuẩn - Tính chất

- Phân phối chuẩn là một trong những phân phối quan trọng nhất, được dùng để mô tả phân phối của nhiều biến ngẫu nhiên trong thực tế, như chiều cao/cân nặng của một người, tổng doanh thu của một công ty, điểm thi của sinh viên, sai số của một phép đo, v.v. Bên cạnh đó, định lý giới hạn trung tâm (Central Limit Theorem) đã chứng tỏ rằng, phân phối chuẩn là phân phối xấp xỉ của nhiều phân phối khác như nhị thức, tổng các biến ngẫu nhiên độc lập, v.v.
- Một số tính chất của phân phối chuẩn:
  - Đồ thị có dạng chuông (bell-shaped)
  - Phân phối đối xứng
  - Trung bình $=$ trung vị (median) $=$ yếu vị (mode)
  - Vị trí của phân phối được xác định bởi kỳ vọng μ
  - Độ phân tán được xác định bởi độ lệch tiêu chuẩn σ
  - Xác định trên $\mathbb{R}$

Phân phối chuẩn tắc

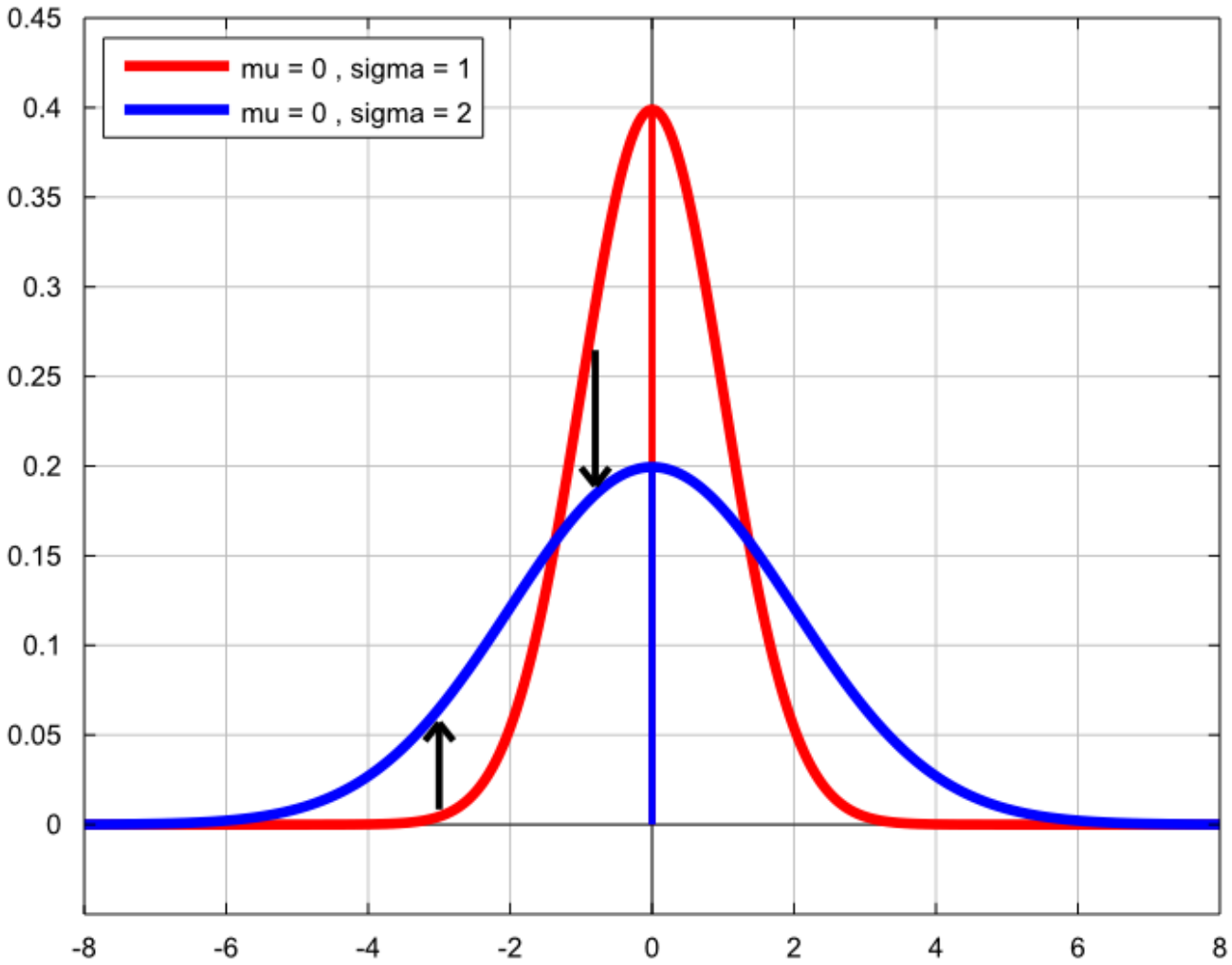
Định nghĩa 2 (Standard normal distribution)

Biến ngẫu nhiên Z được gọi là có phân phối chuẩn tắc nếu nó có phân phối chuẩn với tham số $\mu = 0$ và $\sigma^2 = 1$, ký hiệu $Z \sim \mathcal{N}(0, 1)$.

Theo quy ước, hàm phân phối của biến ngẫu nhiên chuẩn hóa được ký hiệu là $\Phi(z)$, tức

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$$

Phân phối chuẩn tắc



Phân phối chuẩn tắc

Theo định lý về tính tuyến tính của phân phối chuẩn, nếu $X \sim \mathcal{N}\left(\mu; \sigma^2\right)$ thì $\dfrac{X-\mu}{\sigma}$ có phân phối chuẩn tắc hay

$$\frac{X-\mu}{\sigma} \sim N(0,1).$$

Dựa vào tính chất này ta có thể tính xác suất của biến ngẫu nhiên $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

$$\mathbb{P}(X \leq b) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right).$$

Tương tự, với $a \leq b$ thì

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Phân phối Chi bình phương

Định nghĩa 3 (Chi-squared distribution)

Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(0, +\infty)$ được gọi là có phân phối chi bình phương với n bậc tự do, ký hiệu $X \sim \chi^2(n)$, nếu hàm mật độ xác suất có dạng

$$f(x) = \begin{cases} 0 & \text{với } x \leq 0, \\ \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{với } x > 0. \end{cases}$$

trong đó $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ là hàm Gamma .

Xây dựng phân phối Chi bình phương từ phân phối chuẩn

- Nếu $Z \sim \mathcal{N}(0,1)$, thì $Y = Z^2$ sẽ tuân theo một phân phối được gọi là phân phối Chi bình phương với 1 bậc tự do. Ký hiệu: $Y \sim \chi^2(1)$.
- Xét $Y_1, Y_2, \ldots, Y_n$ là n biến ngẫu nhiên độc lập và có phân phối Chi bình phương với 1 bậc tự do. Đặt $X = Y_1 + Y_2 + \cdots + Y_n = \sum_{i=1}^{n} Y_i$, thì X có phân phối Chi bình phương với n bậc tự do. Ký hiệu: $X \sim \chi^2(n)$.
- Suy ra: nếu $Z_1, Z_2, \ldots, Z_n \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, thì $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$.

Định lý 1 (Các đặc trưng của biến ngẫu nhiên có phân phối Chi bình phương)

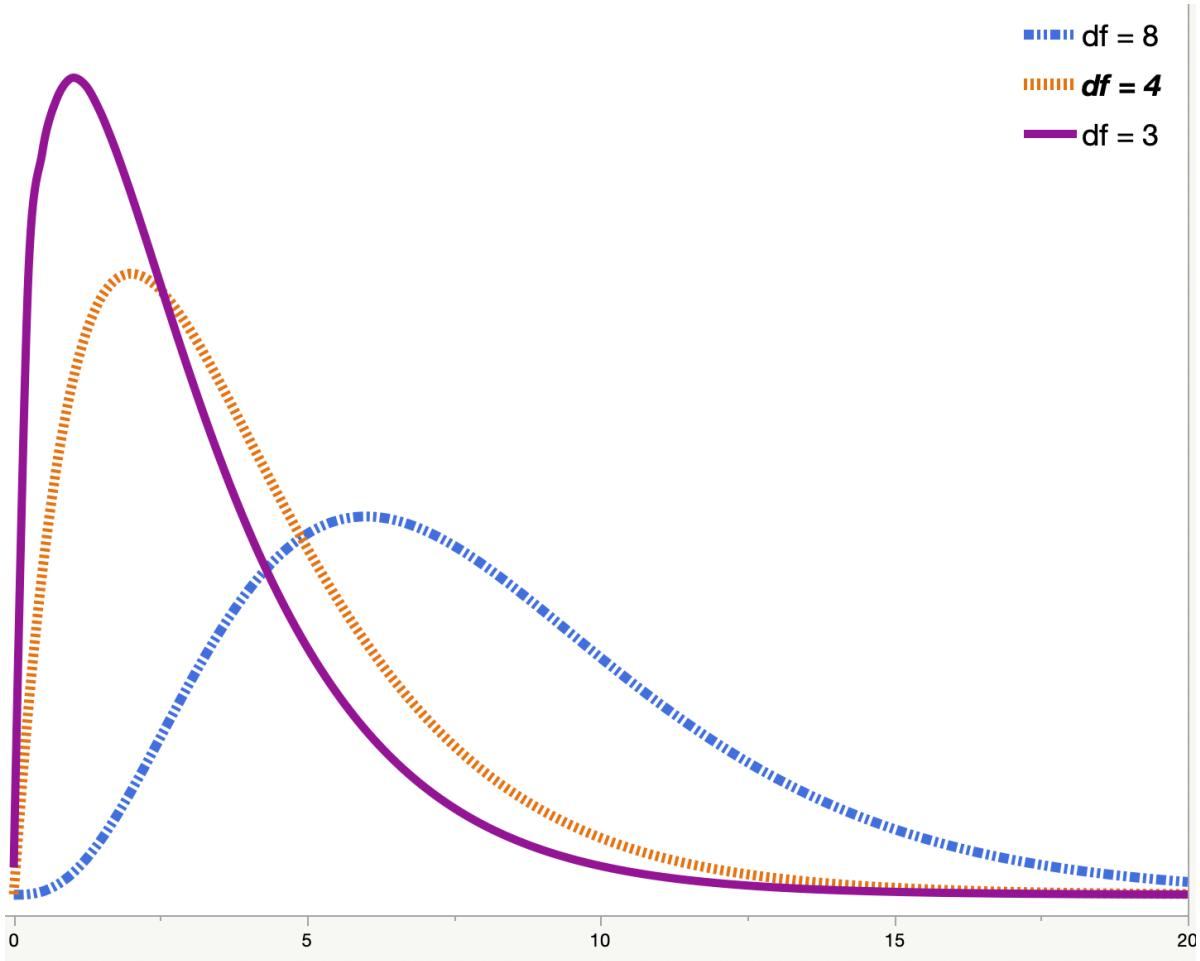
Cho X là biến ngẫu nhiên có phân phối chi bình phương với n bậc tự do thì

i) *Kỳ vọng $\mathbb{E}(X) = n$,*

ii) *Phương sai $\mathbb{V}ar(X) = 2n$,*

iii) *Nếu $X \sim \chi^2(n), Y \sim \chi^2(m)$ và X, Y là hai biến ngẫu nhiên độc lập thì $X + Y \sim \chi^2(m+n)$.*

Phân phối Chi bình phương



Phân phối Student

Định nghĩa 4 (Student distribution)

Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(-\infty, +\infty)$ được gọi là có phân phối Student với n bậc tự do, ký hiệu $X \sim t(n)$, nếu hàm mật độ xác suất có dạng

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

trong đó $\Gamma(x)$ là hàm Gamma.

Xây dựng pp Student từ pp chuẩn và pp Chi bình phương

- Xét $Z \sim \mathcal{N}(0,1)$ và $Y \sim \chi^2(n)$, Z và Y độc lập.
- Đặt:

$$T = \frac{Z}{\sqrt{\frac{Y}{n}}}.$$

- Biến ngẫu nhiên T được định nghĩa như trên sẽ tuân theo phân phối Student với n bậc tự do, ký hiệu $T \sim t(n)$.

Định lý 2 (Các đặc trưng của biến ngẫu nhiên có phân phối Student)

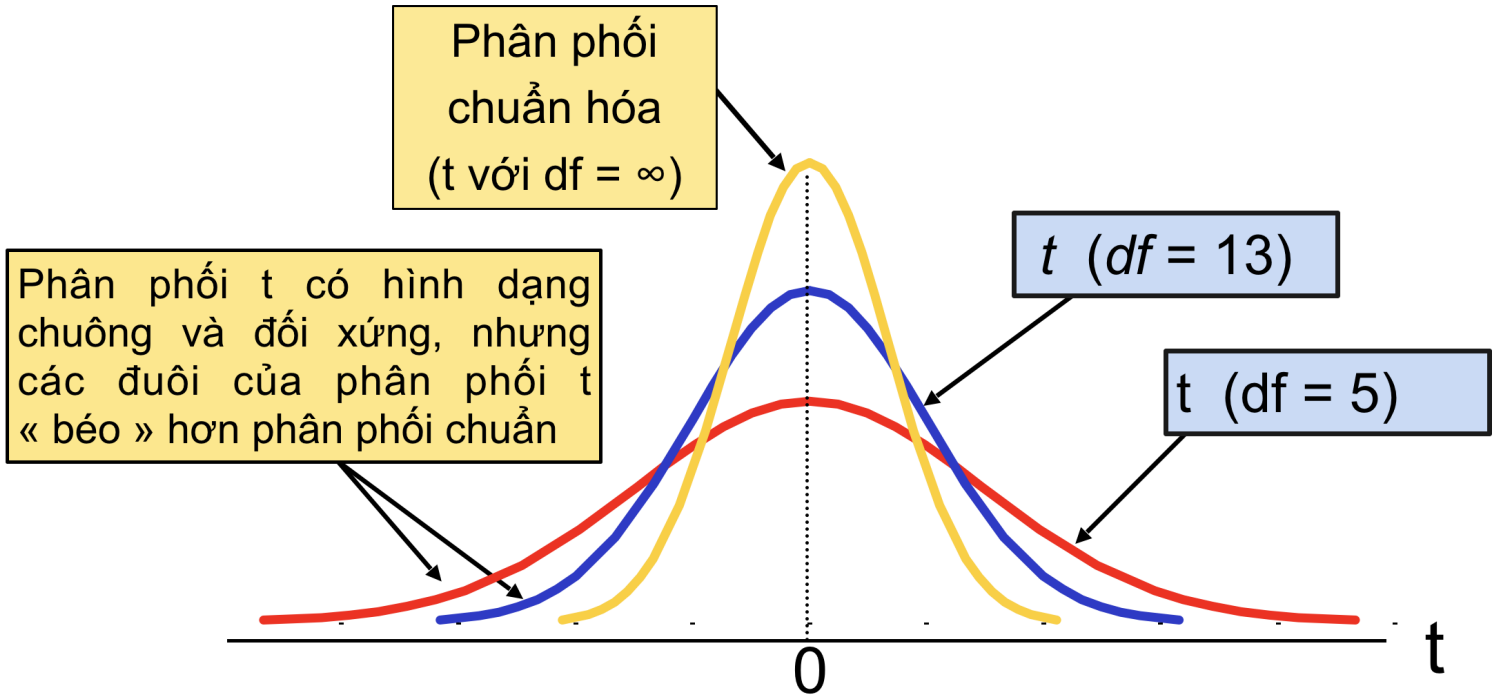
Cho $X \sim t(n)$ thì

i) *Kỳ vọng $\mathbb{E}(X) = 0$ nếu $n > 1$, các trường hợp còn lại $\mathbb{E}(X)$ không được định nghĩa.*

ii) *Phương sai $\mathbb{V}ar(X) = \frac{n}{n-2}$ nếu $n > 2$; $\mathbb{V}ar(X) = \infty$ nếu $1 < n \leq 2$ các trường hợp còn lại $\mathbb{V}ar(X)$ không được định nghĩa.*

Phân phối Student

- Đồ thị của hàm mật độ phân phối Student có dạng hình chuông như đồ thị hàm mật độ của phân phối chuẩn, nhưng có phần đỉnh thấp hơn và hai phần đuôi cao hơn so với phân phối chuẩn.



Phân phối mẫu

Định nghĩa 5

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ một tổng thể và hàm giá trị thực (hay véc-tơ) $T(x_1, x_2, \ldots, x_n)$. Thì biến ngẫu nhiên hay véc-tơ ngẫu nhiên $Y = T(X_1, X_2, \ldots, X_n)$ được coi là một thống kê. Phân phối xác suất của thống kê Y được gọi là phân phối mẫu của Y.

Những phân phối mẫu được khảo sát:

- Phân mối mẫu của trung bình,
- Phân phối mẫu của phương sai,
- Phân phối mẫu của tỷ lệ.

Phân phối mẫu của trung bình và phương sai

Định lý 3

Nếu tổng thể X có phân phối chuẩn $X \sim N(\mu, \sigma^2)$ và $(X_1, ..., X_n)$ là một mẫu ngẫu nhiên từ tổng thể trên. Xét

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{và} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Ta có các kết quả sau:

1. $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
2. $\frac{(n-1)}{\sigma^2}S^2 = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.
3. $\frac{(\bar{X} - \mu)\sqrt{n}}{S} \sim t(n-1)$.
4. $\bar{X}$ *và* S^2 *là hai biến ngẫu nhiên độc lập.*

Phân phối mẫu của trung bình và phương sai

Trong trường hợp tổng thể không có phân phối chuẩn, từ định lý giới hạn trung tâm ta suy ra rằng

$$\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \xrightarrow{D} N(0,1),$$
$$\frac{(\bar{X}-\mu)\sqrt{n}}{S} \xrightarrow{D} N(0,1).$$

Từ kết quả này, trong thực hành, khi mẫu có kích thước, n, đủ lớn ta có các phân phối xấp xỉ chuẩn sau

$$\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \approx N(0,1),$$
$$\frac{(\bar{X}-\mu)\sqrt{n}}{S} \approx N(0,1).$$

Sai số chuẩn của trung bình

Định nghĩa 6

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ một tổng thể có trung bình μ và phương sai $\sigma^2 < \infty$. Sai số chuẩn (Standard Error - SE) của trung bình, ký hiệu $\sigma_{\bar{X}}$ được định nghĩa như sau

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Ý nghĩa:

- $\sigma_{\bar{X}}$ đo độ biến thiên của $\bar{X}$ xung quanh μ,
- Sai số chuẩn càng nhỏ, ước lượng tham số từ tổng thể càng tốt và độ tin cậy cao.

Sai số chuẩn của trung bình

$\sigma_{\bar{X}}$ bị ảnh hưởng bởi hai yếu tố:

(1) Cỡ mẫu n: Cỡ mẫu càng lớn $\Rightarrow$ sai số chuẩn càng nhỏ, chú ý rằng khi $n = 1$ thì $\sigma_{\bar{X}} = \sigma$.

(2) Độ biến thiên của tổng thể σ: σ càng lớn $\Rightarrow$ sai số chuẩn càng lớn.

Phân phối mẫu của tỷ lệ

- Giả sử cần khảo sát đặc trưng $\mathcal{A}$ của một tổng thể, khảo sát n phần tử và đặt

$$X_i = \begin{cases} 1, & \text{nếu thỏa } \mathcal{A} \\ 0, & \text{nếu không thỏa } \mathcal{A} \end{cases}$$

  thu được mẫu ngẫu nhiên $X_1, \ldots, X_n$ với $X_i \sim B(p)$, p là tỷ lệ phần tử thỏa đặc trưng $\mathcal{A}$.
- Đặt $X = \sum_{i=1}^{n}$ là số phần tử thỏa đặc trưng $\mathcal{A}$ trong mẫu khảo sát, thì $X \sim B(n, p)$.
- Tỷ lệ mẫu $\hat{P}$ là một ước lượng của tỷ lệ p xác định bởi

$$\hat{P} = \frac{X}{n}.$$

Phân phối mẫu của tỷ lệ

- Kỳ vọng và phương sai của $\hat{P}$ bằng

$$\mathbb{E}\left(\hat{P}\right) = p, \quad \mathbb{V}ar\left(\hat{P}\right) = \frac{p(1-p)}{n}.$$

- Theo định lý giới hạn trung tâm ta có

$$\frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{p}}} \rightsquigarrow N(0,1).$$

Vì vậy trong thực hành, khi $np \geq 5, n(1-p) \geq 5$, ta có $\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right)$.