Thống kê mô tả

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

Outline

Tổng thể và mẫu

- **Tổng thể (population):** tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.

Tổng thể và mẫu

- **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.

Tổng thể và mẫu

- **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.

Tổng thể và mẫu

- **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.
- **Thống kê (statistic):** là một đặc trưng cụ thể của một 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

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

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.

Thống kê mô tả

- **Thống kê mô tả (descriptive statistics):** là quá trình thu thập, tổng hợp và xử lý dữ liệu để biến đổi dữ liệu thành thông tin.
  - Thu thập dữ liệu: khảo sát, đo đạc, ...
  - Biểu diễn dữ liệu: dùng bảng và đồ thị
  - Tổng hợp dữ liệu: tính các tham số mẫu như trung bình mẫu (sample mean), phương sai mẫu (sample variance), trung vị (median), ...

Thống kê suy diễn

- **Suy diễn (inference)** là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.

Thống kê suy diễn

- **Suy diễn (inference)** là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- **Thống kê suy diễn (Inferential statistics):** xử lý các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở cho những dự đoán (predictions), dự báo (forecasts) và các ước lượng (estimations).

Thống kê suy diễn

- **Suy diễn (inference)** là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- **Thống kê suy diễn (Inferential statistics):** xử lý các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở cho những dự đoán (predictions), dự báo (forecasts) và các ước lượng (estimations).
  + Ước lượng: ví dụ ước lượng tỷ lệ sản phẩm kém chất lượng trong 1 nhà máy, ước lượng tỷ lệ hành khách đã mua vé nhưng vắng mặt trên một chuyến bay.

Thống kê suy diễn

- **Suy diễn (inference)** là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- **Thống kê suy diễn (Inferential statistics):** xử lý các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở cho những dự đoán (predictions), dự báo (forecasts) và các ước lượng (estimations).
  + Ước lượng: ví dụ ước lượng tỷ lệ sản phẩm kém chất lượng trong 1 nhà máy, ước lượng tỷ lệ hành khách đã mua vé nhưng vắng mặt trên một chuyến bay.
  + Kiểm định giả thuyết: ví dụ cần kiểm định khẳng định rằng lợi nhuận trung bình của một cửa hàng trong một tháng là 300 triệu.

Giới thiệu

- Việc mô tả dữ liệu bằng đồ thị sẽ cho ta một cái nhìn tổng quan về dữ liệu trước khi đi vào phân tích cụ thể.
- Các loại đồ thị được sử dụng sẽ phụ thuộc vào dạng biến cần phân tích.
- Trong phần này, ta sẽ khảo sát chủ yếu về đồ thị tổ chức tần số (histogram).

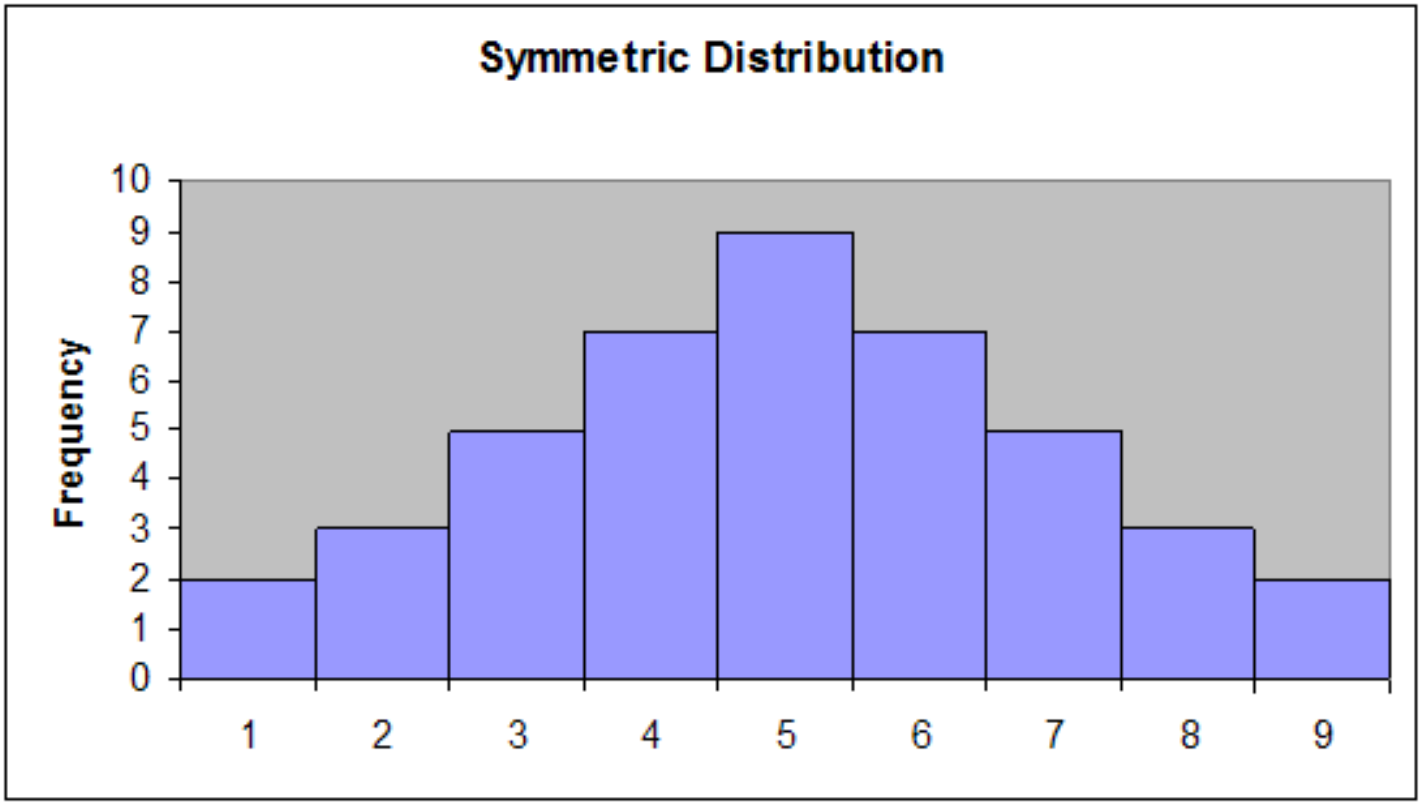
Histogram

Đồ thị tổ chức tần số (histogram) cho phép:

- mô tả phân phối của dữ liệu,
- xem xét tính đối xứng/bất đối xứng, tập trung/phân tán của dữ liệu,
- nhận dạng phân phối chuẩn (bell-shaped),
- xác định mode (unimodal, bimodal).

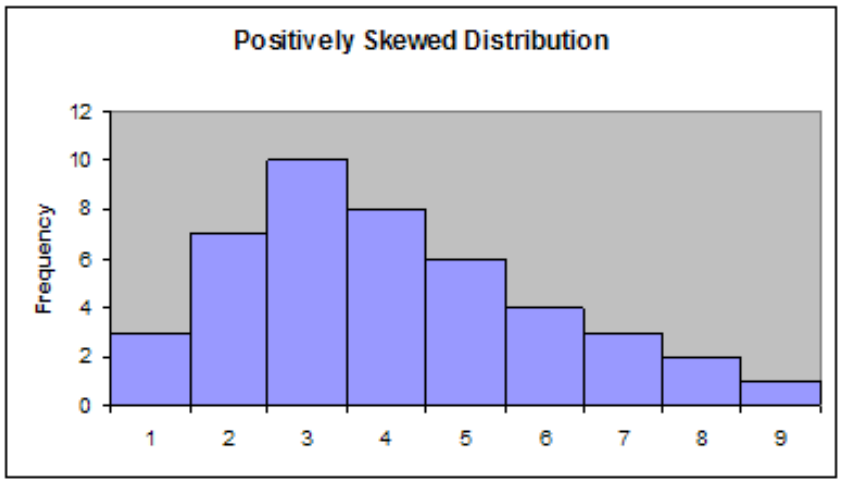
Hình dạng của phân phối

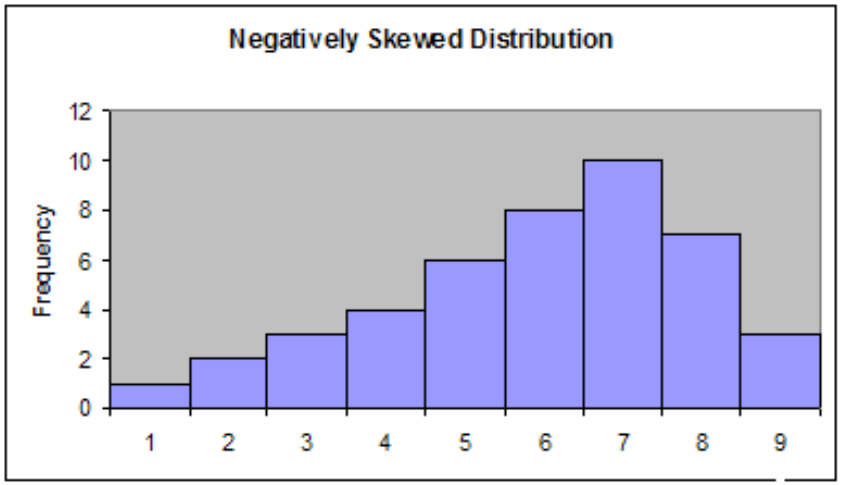
- Hình dạng của phân phối (shape of the distribution) gọi là đối xứng (symmetric) nếu các giá trị quan trắc cân bằng xung quanh trung tâm.



Hình dạng của phân phối

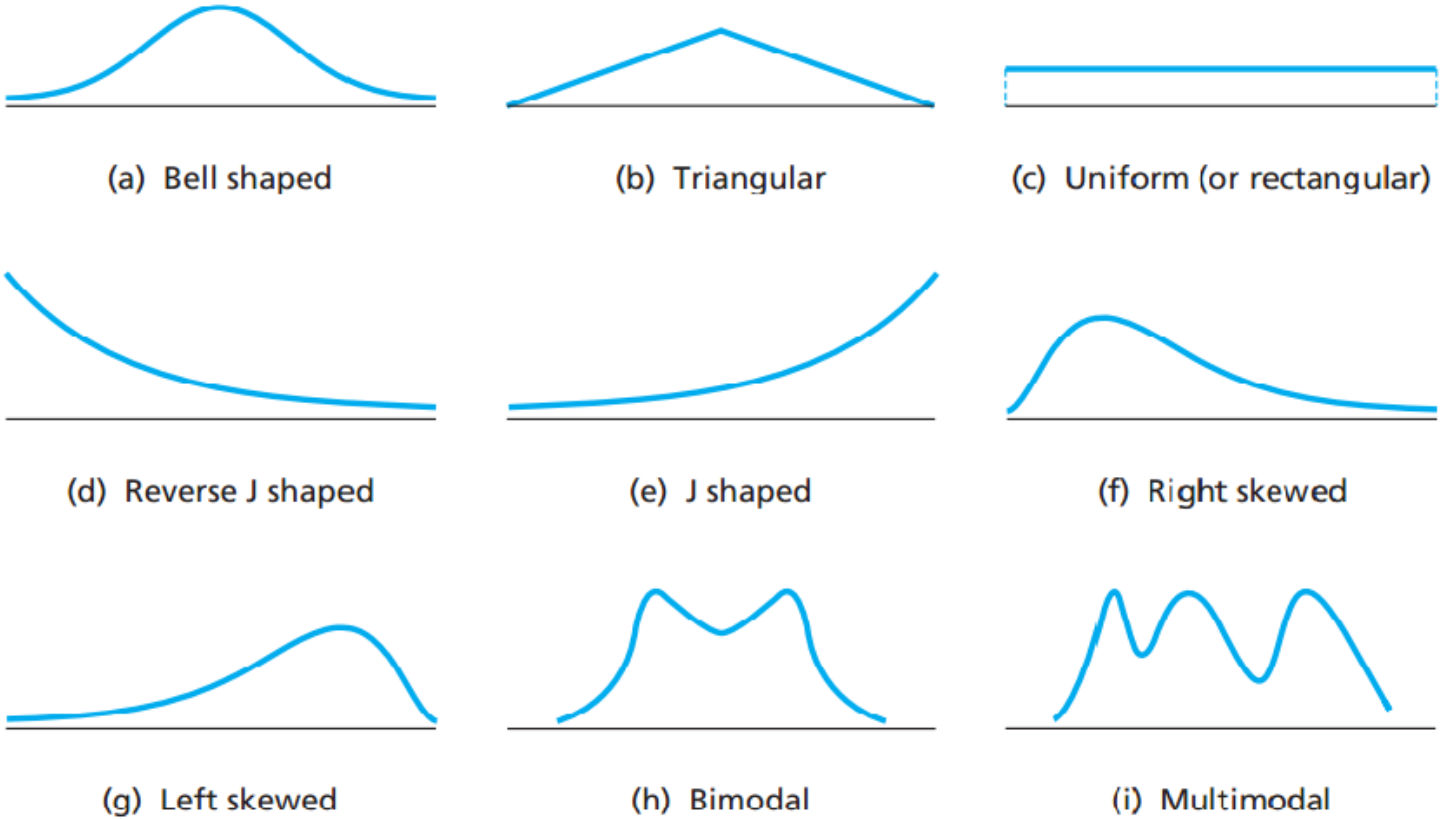
- Hình dạng của phân phối gọi là bất đối xứng (skewed) nếu dữ liệu quan trắc không phân bố đối xứng xung quanh trung tâm.



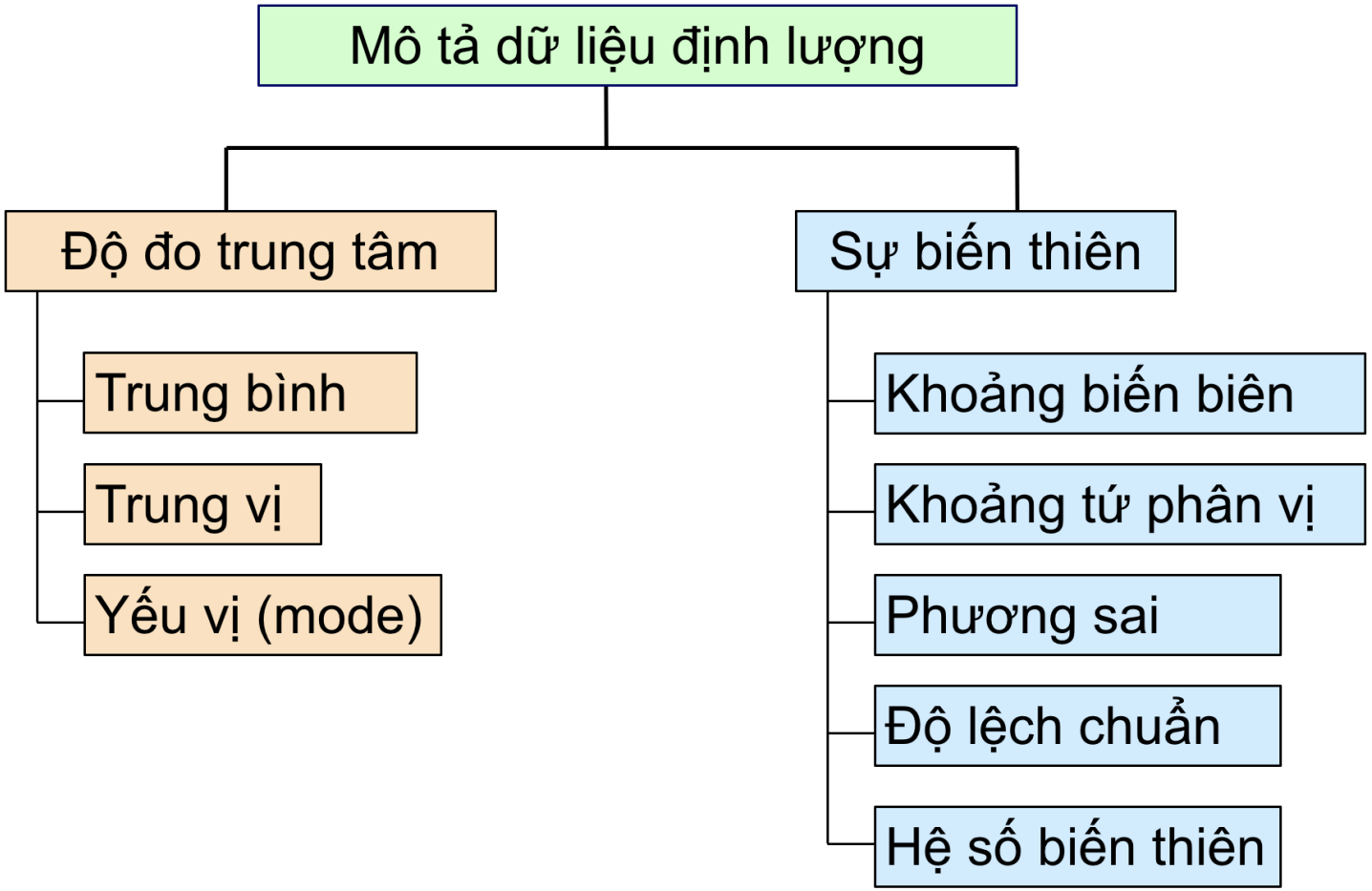


Dáng điệu của phân phối

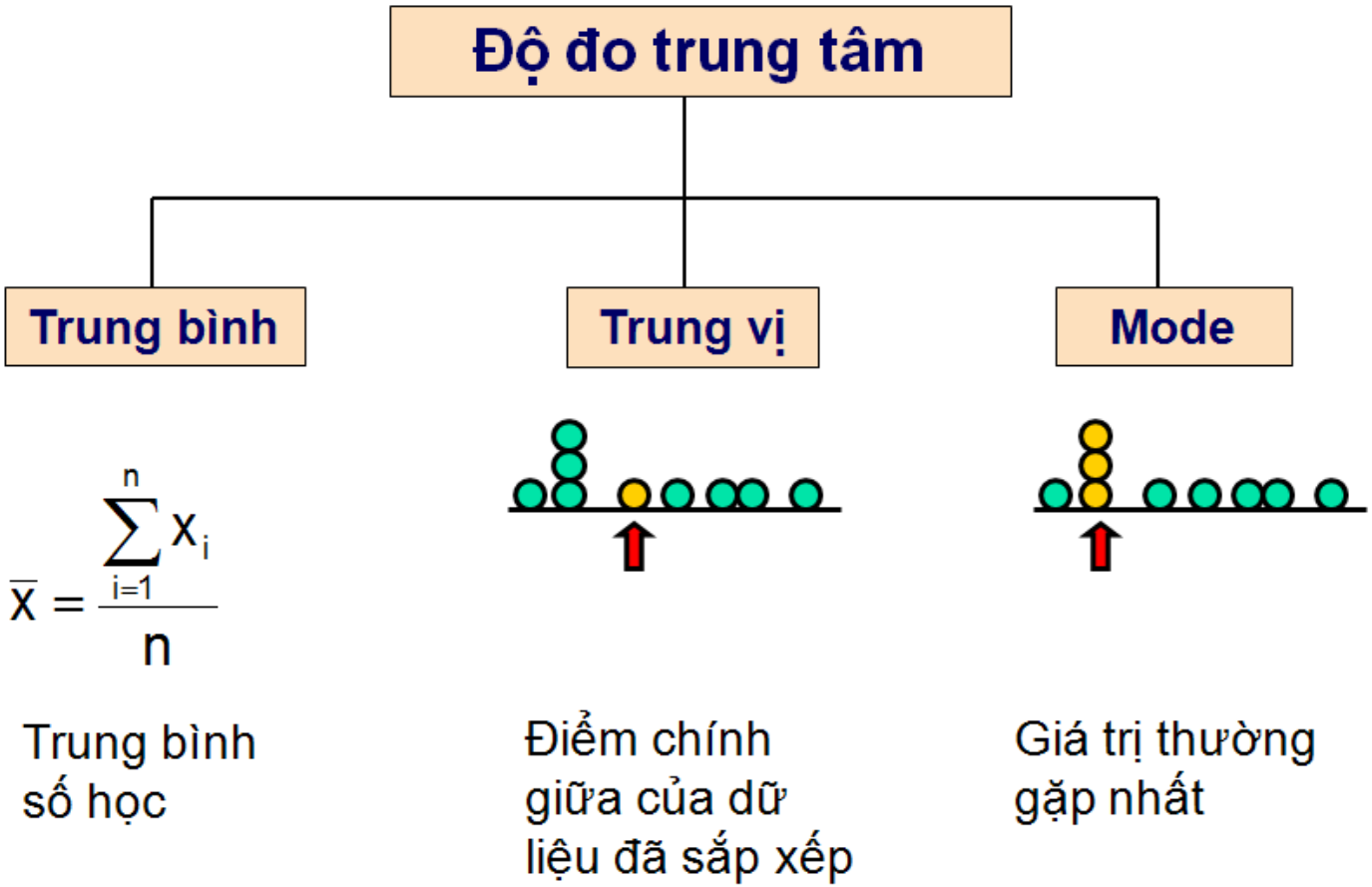
- Sử dụng đồ thị histogram để nhận biết phân phối xác suất của một đại lượng ngẫu nhiên.



Giới thiệu


Mô tả dữ liệu định lượng
Độ đo trung tâm
Trung bình
Trung vị
Yếu vị (mode)
Sự biến thiên
Khoảng biến biên
Khoảng tứ phân vị
Phương sai
Độ lệch chuẩn
Hệ số biến thiên

Các độ đo trung tâm



Trung bình

- Trung bình (mean) là đại lượng thường được sử dụng nhất để đo giá trị trung tâm của dữ liệu.

Trung bình

- Trung bình (mean) 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, trung bình tổng thể tính bởi

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

Trung bình

- Trung bình (mean) 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, trung bình tổng thể tính bởi

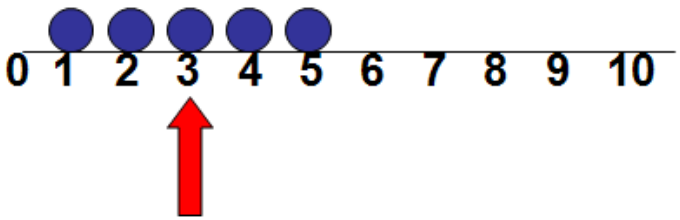
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}.$$

- Với một mẫu cỡ n, trung bình mẫu

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

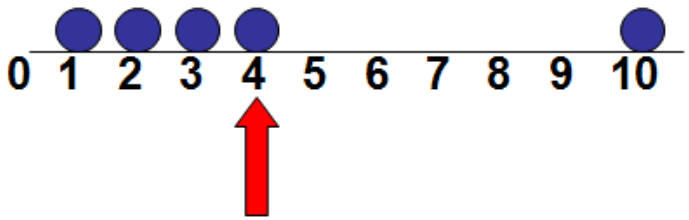
Trung bình

- Trung bình bị ảnh hưởng bởi các giá trị ngoại lai (outliers).



Mean = 3

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

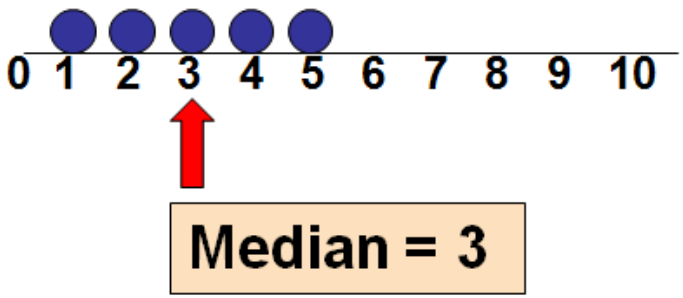


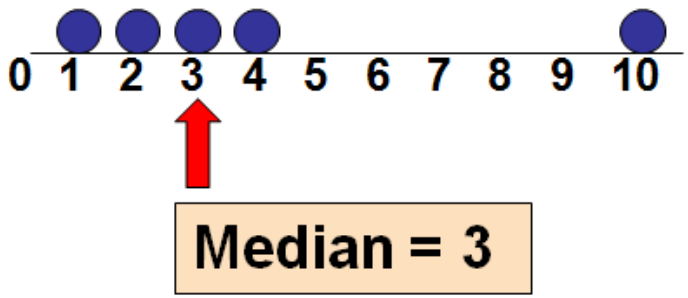
Mean = 4

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

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 trên, 50% bên dưới).
- Trung vị không bị ảnh hưởng bởi các điểm ngoại lai (outliers).





Trung vị

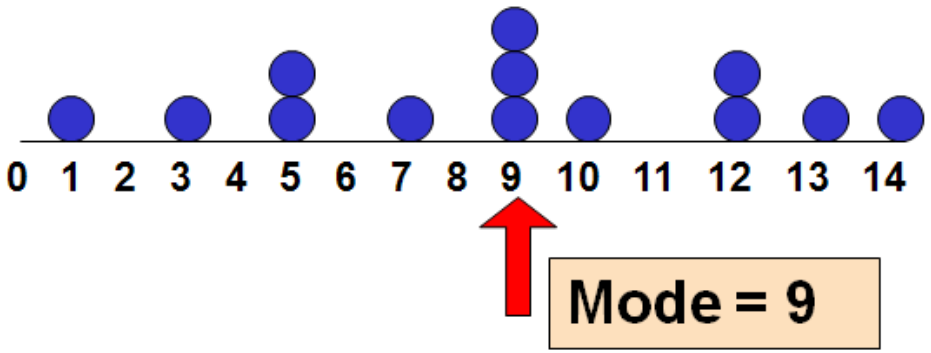
- Vị trí của trung vị: sắp xếp dữ liệu theo thứ tự tăng dần, gọi i là vị trí của trung vị

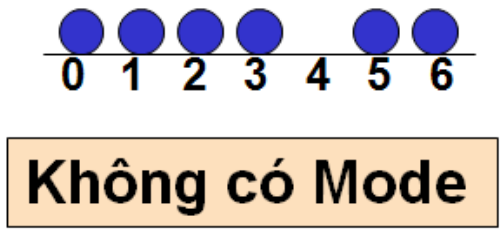
$$i = \frac{n+1}{2}$$

  + Nếu i chẵn, trung vị $= X_i$,
  + Nếu i lẻ, trung vị $= \frac{X_{[i]} + X_{[i]+1}}{2}$, với $[i]$ là phần nguyên của i.

Mode (Yếu vị)

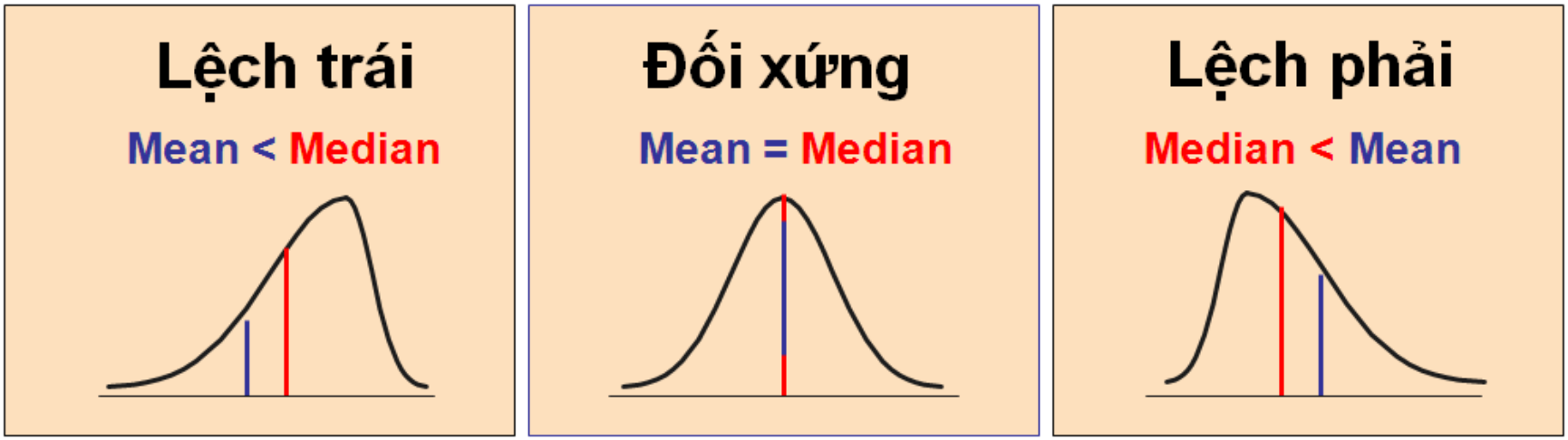
- Mode (yếu vị) là một đại lượng để đo xu hướng trung tâm của dữ liệu,
- là giá trị thường xảy ra nhất,
- không bị ảnh hưởng bởi các điểm ngoại lai,
- có thể sử dụng cho cả dữ liệu định tính và dữ liệu định lượng,
- có thể có nhiều mode hoặc không tồn tại mode.





Độ đo nào là tốt nhất?

- 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.
- Yếu vị (mode) thường dùng để mô tả các biến định tính.
- Vị trí của trung vị và trung bình ảnh hưởng bởi hình dạng của phân phối:



Các độ đo trung tâm - ví dụ

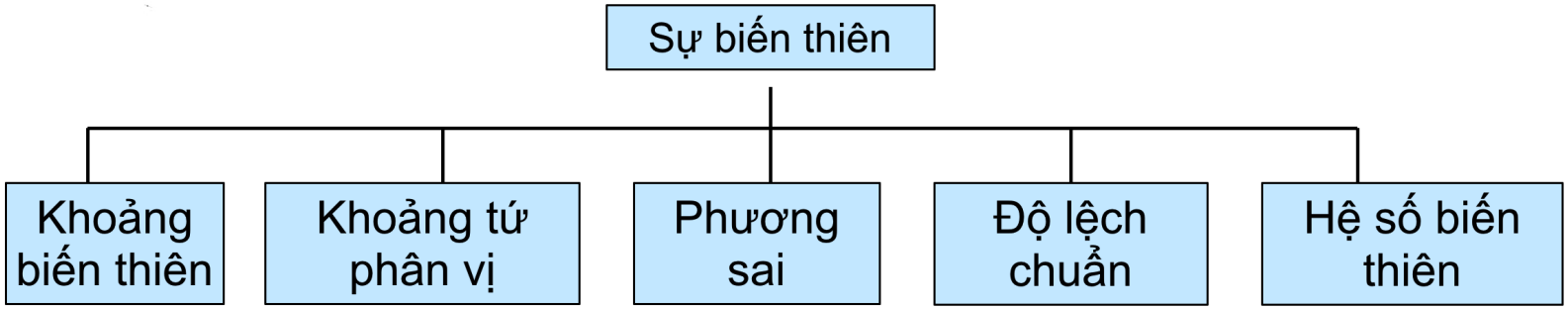
Ví dụ 1

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

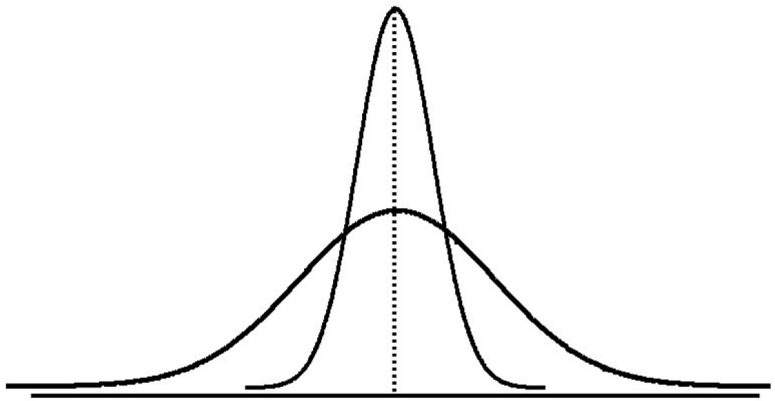
72	*49*	*62*	*58*	*73*	*55*	*78*	*83*	*57*	*63*
73	*73*	*75*	*85*	*85*	*64*	*61*	*67*	*75*	*91*

Tìm các điểm điểm trung bình, trung vị và yếu vị.

Độ đo sự biến thiên



Độ đo về sự biến thiên cho biết thông tin về sự phân tán hay sự biến thiên của dữ liệu

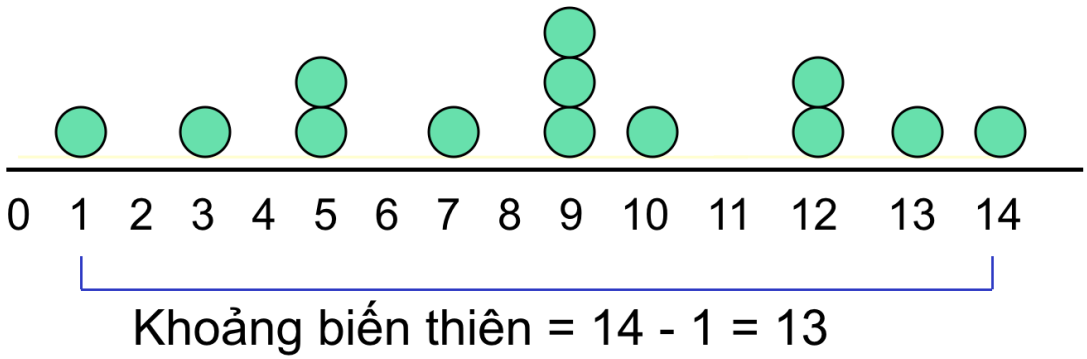


cùng trung tâm, sự biến thiên khác nhau

Khoảng biến thiên

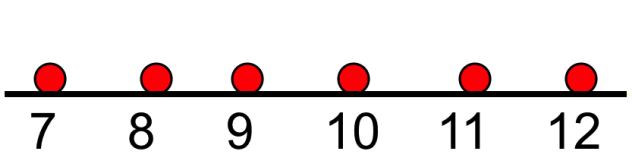
- Khoảng biến thiên (range) là độ đo sự biến thiên đơn giản nhất,
- Là độ chênh lệch giữa giá trị lớn nhất và bé nhất của dữ liệu quan trắc

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

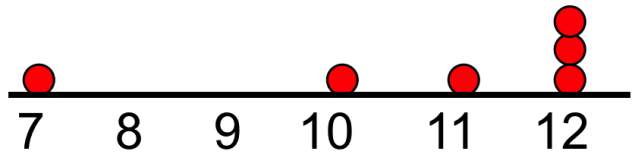


Nhược điểm của khoảng biến thiên

- Bỏ qua phân bố của dữ liệu



Khoảng biến thiên = 12 - 7 = 5



Khoảng biến thiên = 12 - 7 = 5

- Bị ảnh hưởng bởi các điểm outlier

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Khoảng biến thiên = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Khoảng biến thiên = 120 - 1 = 119

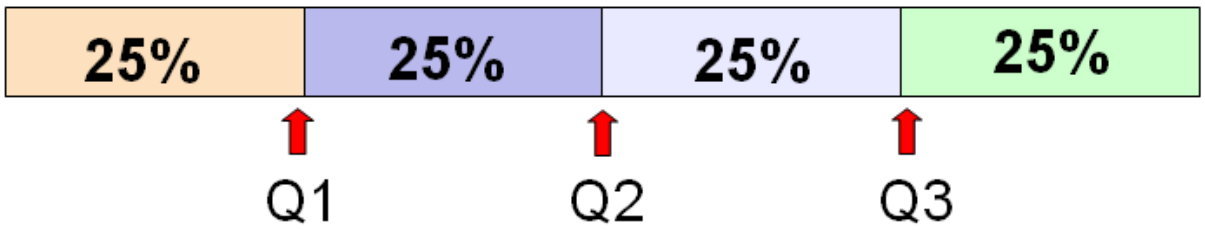
Khoảng tứ phân vị

- Ta có thể loại bỏ các điểm ngoại bằng cách sử dụng khoảng tứ phân vị (Inter-Quartile Range - IQR) (hay còn được gọi là độ trãi giữa).
- Công thức tính khoảng tứ phân vị:

$$IQR = Q_3 - Q_1$$

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.



Công thức tìm phân vị

Sắp xếp dữ liệu theo thứ tự tăng dần, gọi Q_1, Q_2 (trung vị), Q_3 lần lượt là phân vị thứ 1, 2 và 3 của dữ liệu. Vị trí của Q_1, Q_2 và Q_3 được xác định như sau

$$\text{Vị trí phân vị thứ nhất} = 0.25(n+1)$$

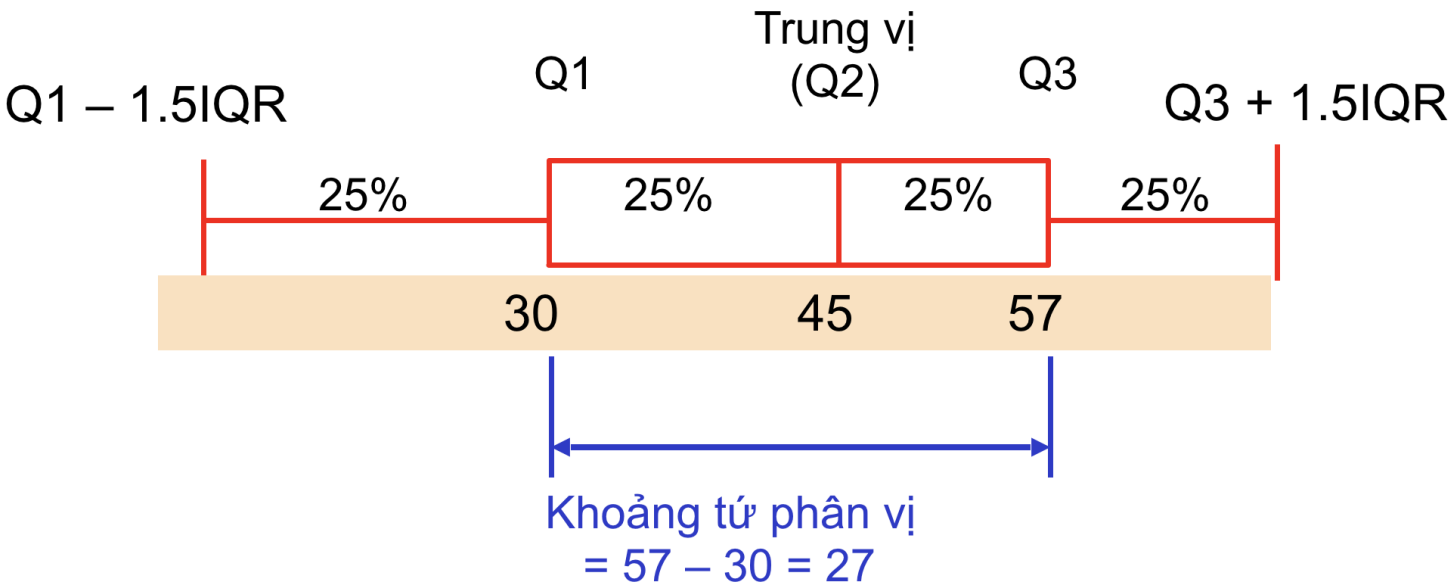
$$\text{Vị trí phân vị thứ hai} = 0.5(n+1)$$

$$\text{Vị trí phân vị thứ ba} = 0.75(n+1)$$

với n là số giá trị quan trắc.

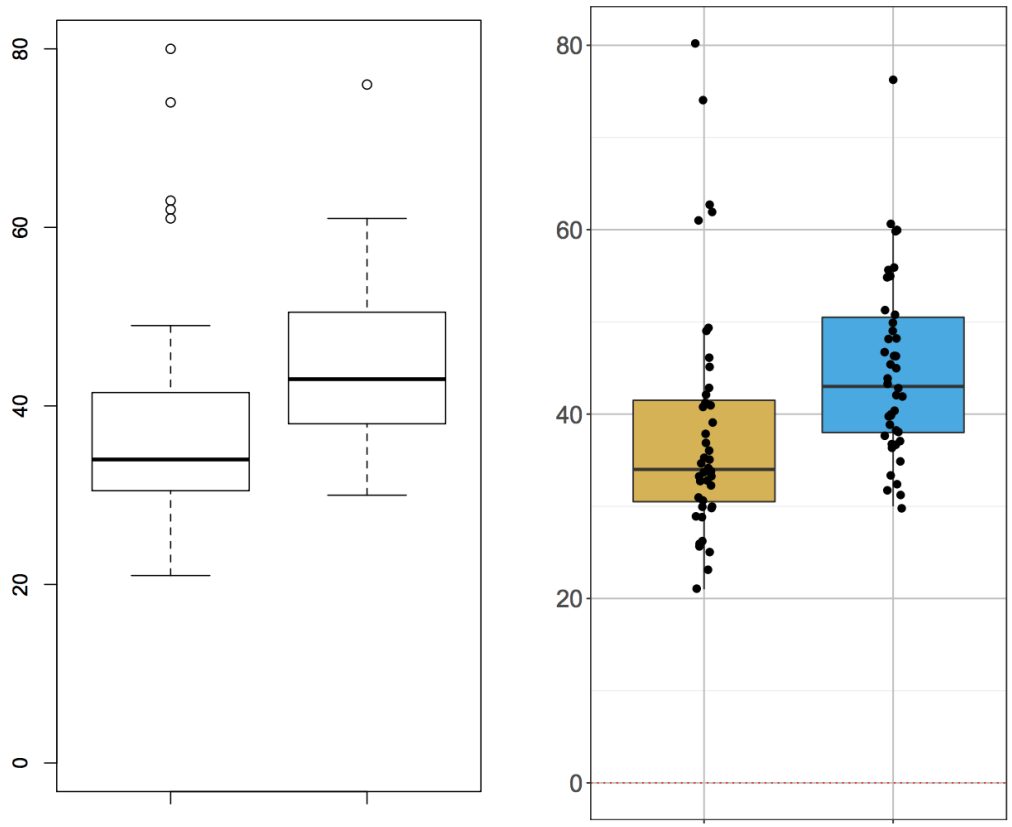
Đồ thị hộp (boxplot)

- Để biểu diễn khoảng tứ phân vị và các điểm ngoại lai (outliers) ta dùng đồ thị hộp (boxplot):



Đồ thị hộp (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

- 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

- 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

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

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

- 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

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

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

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

với $\bar{X}$ là trung bình mẫu, n là cỡ mẫu, X_i là giá trị thứ i của biến X.

Độ lệch tiêu chuẩn

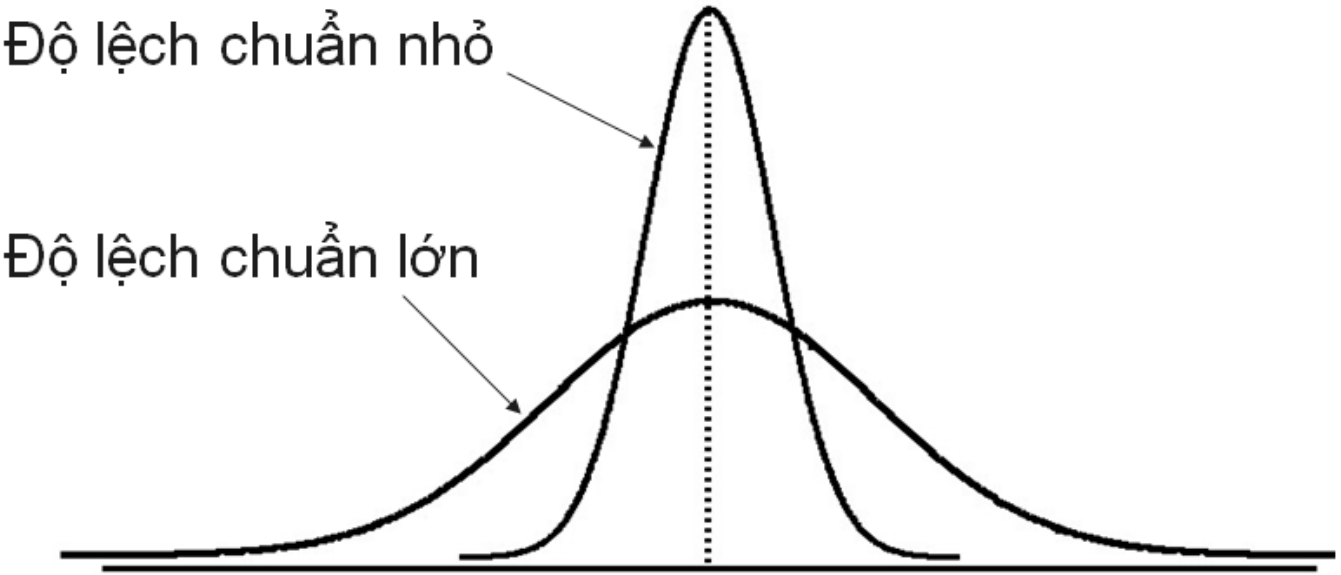
- Độ lệch tiêu chuẩn (Standard deviation) được dùng để đo sự biến thiên, biểu diễn sự biến thiên xung quanh trung bình,
- Có cùng đơn vị đo với dữ liệu gốc.
- Độ lệch chuẩn của tổng thể, ký hiệu là σ:

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

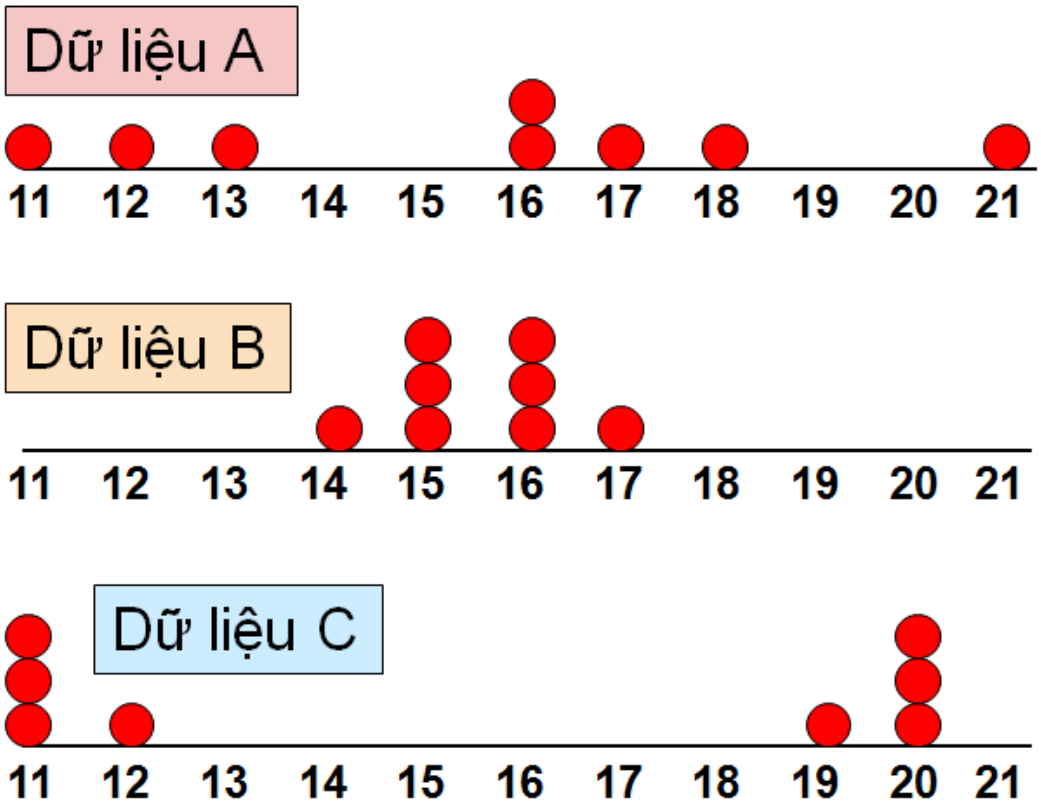
- Độ lệch chuẩn của mẫu,

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

Độ đo sự biến thiên



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



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

- 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.
- Đo mối liên hệ giữa sự biến thiên và trung bình.

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.
- Đo mối liên hệ giữa sự biến thiên và trung bình.
- Đơn vị tính bằng %.

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.
- Đo mối liên hệ giữa sự biến thiên và 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.

Các độ đo sự biến thiên - ví dụ

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

a) *Tìm* Q_1, Q_2, Q_3.

b) *Vẽ đồ thị boxplot cho tập dữ liệu trên.*

c) *Tính phương sai và độ lệch chuẩn.*

Phân phối Chi bình phương

Định nghĩa 4.1 (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: $Y \sim \chi^2(1)$.

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: $X \sim \chi^2(n)$.

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 $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$.

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 $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$.

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)

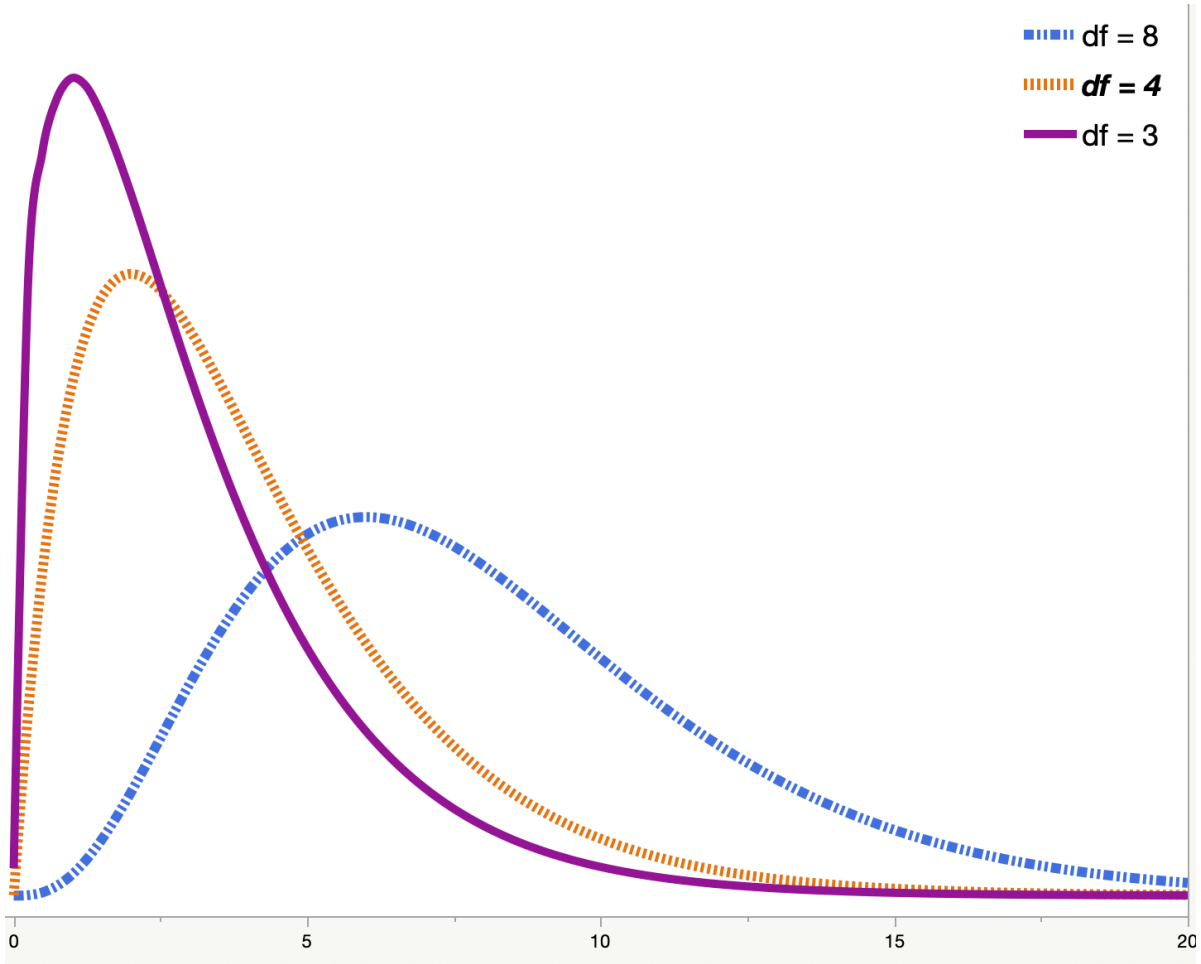
Cho X là biến ngẫu nhiên có phân phối chi bình phương với n bậc tự do thì

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 4.2 (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

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

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 $T \sim t(n)$.

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

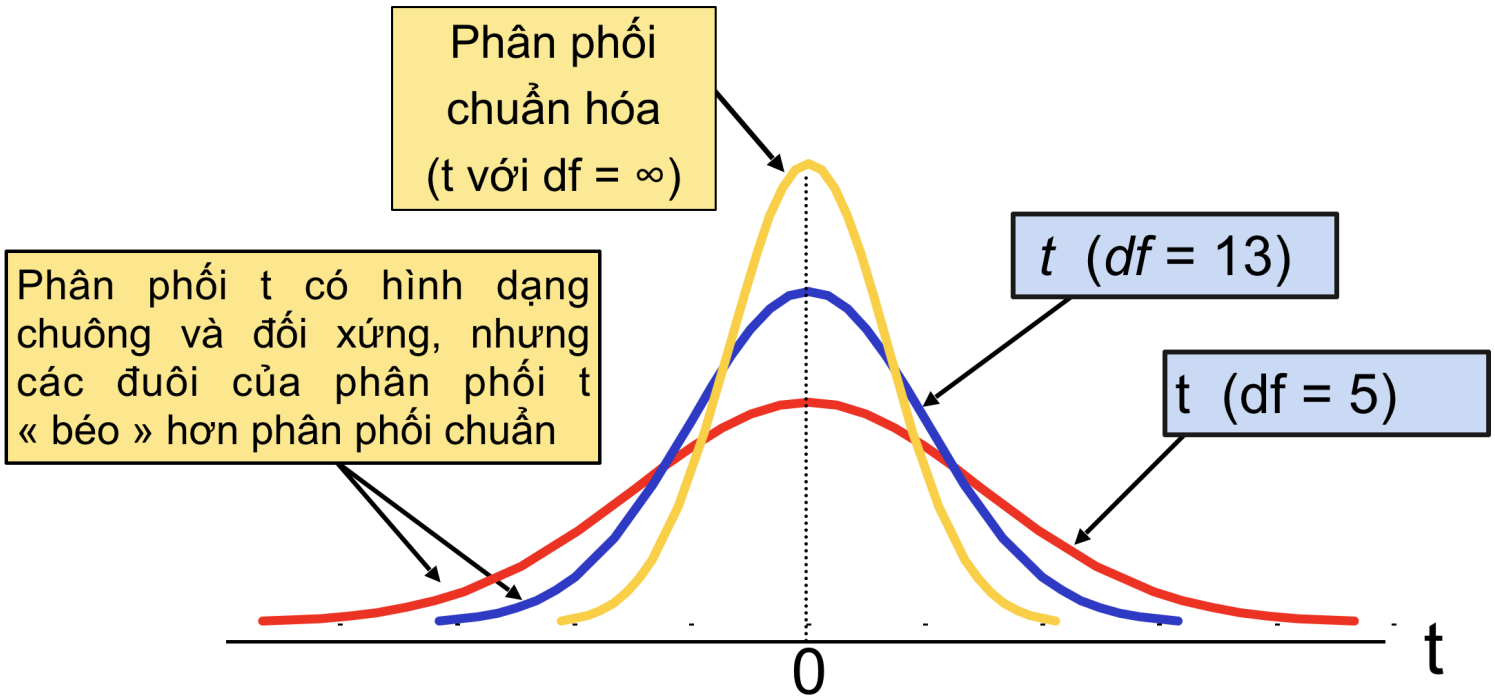
i) *Kỳ vọng $\mathbb{E}(X) = 0$ nếu $n > 1$, các trường hợp còn lại $\mathbb{E}(X)$ không được định nghĩa.*

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

Lưu ý

- 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 5.1

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. $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.
2. $\dfrac{(n-1)}{\sigma^2}S^2 = \sum_{i=1}^{n}\dfrac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.
3. $\dfrac{(\bar{X} - \mu)\sqrt{n}}{S} \sim t(n-1)$
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 5.2

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 $np \geq 10, n(1-p) \geq 10$, ta có $\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right)$.