Chương Thống kê mô tả

Nguyễn Thị Mộng Ngọc
University of Science, VNU - HCM
ngtmngoc@hcmus.edu.vn

Plan

Tổng quan về thống kê

Hai lĩnh vực thống kê:

- Thống kê mô tả
  - Thu thập số liệu
  - Tính toán các đặc trưng đo lường
  - Mô tả, trình bày dữ liệu, ...
    nhằm mô tả đối tượng nghiên cứu.
- Thống kê suy diễn
  - Ước lượng, kiểm định thống kê
  - Phân tích mối liên hệ
  - Dự đoán, ...
    Mô hình hóa trên các dữ liệu quan trắc để đưa ra các suy diễn về đối tượng được nghiên cứu.

Một số khái niệm thường dùng trong thống kê

- Thống kê: khoa học và nghệ thuật thu thập, phân tích, trình bày và diễn giải dữ liệu.
- Tổng thể: tập hợp các đơn vị/phần tử cần phân tích/nghiên cứu.
- Đơn vị tổng thể: phần tử nhỏ nhất tạo thành tổng thể.
- Mẫu: một phần của tổng thể được chọn ra để thu thập thông tin.

Ví dụ: Để tìm hiểu điểm trung bình môn Thống kê của sinh viên trường đại học KHTN tp HCM, người ta xét bảng điểm của 250 sinh viên được chọn ngẫu nhiên. Hãy chỉ ra tổng thể, đơn vị tổng thể và mẫu.

Một số khái niệm thường dùng trong thống kê (tt)

- Dữ liệu: những sự kiện và con số được thu thập, phân tích và tổng hợp để trình bày và giải thích.
- Tập dữ liệu: tất cả các dữ liệu thu thập trong nghiên cứu cụ thể.
- Biến: khái niệm dùng để chỉ các đặc điểm của đơn vị tổng thể mà ta nghiên cứu.

Ví dụ: Để nghiên cứu sinh viên của một trường đại học, ta cần nghiên cứu các biến như: giới tính, tuổi, dân tộc, ngành học, số tiền chi tiêu trung bình hàng tháng, ...

Một số khái niệm thường dùng trong thống kê (tt)

- Biến: đặc điểm của đơn vị tổng thể dùng để quan sát hay thu thập dữ liệu
  - Biến định tính: đặc điểm biểu hiện không phải là số
  - Biến định lượng: đặc điểm biểu hiện là các trị số có thể rời rạc hay liên tục

  Trong ví dụ trước:
  - biến định tính: giới tính, dân tộc , ngành học,
  - biến định lượng: tuổi, số tiền chi tiêu trung bình hàng tháng.

Một số khái niệm thường dùng trong thống kê (tt)

- Quan trắc (quan sát): Tập hợp tất cả các dữ liệu thu thập được của một đơn vị tổng thể hay mẫu.

Ví dụ:

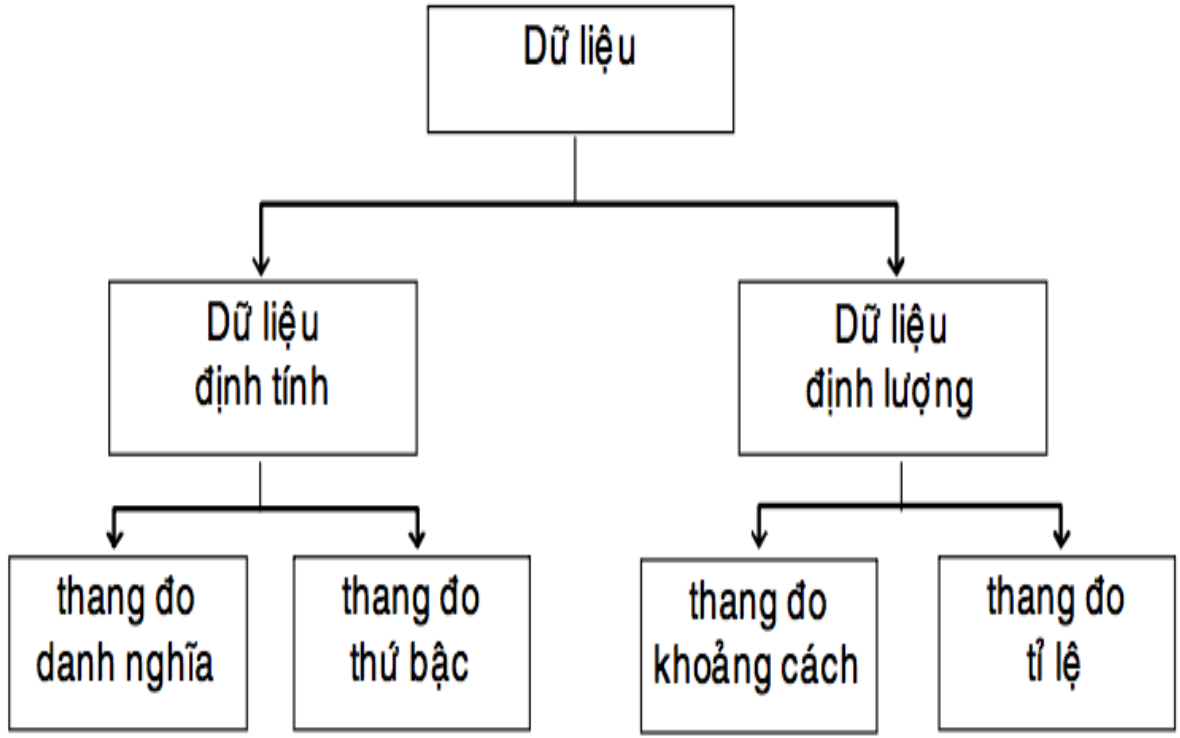
- Quan trắc 1: giới tính: nam, tuổi: 19, dân tộc: Kinh , ngành học: 401, số tiền chi tiêu trung bình hàng tháng: 2.5 triệu đồng.
- Quan trắc 2: giới tính: nữ, tuổi: 21, dân tộc: Tày , ngành học: 402, số tiền chi tiêu trung bình hàng tháng: 2 triệu đồng.

Các loại thang đo

- Thang đo danh nghĩa: dùng để phân loại
- Thang đo thứ bậc: phản ánh sự hơn kém
- Thang đo khoảng cách: phản ánh mức độ hơn kém
- Thang đo tỷ lệ: phản ánh mức độ hơn kém và so sánh tỷ lệ

Phân loại dữ liệu:

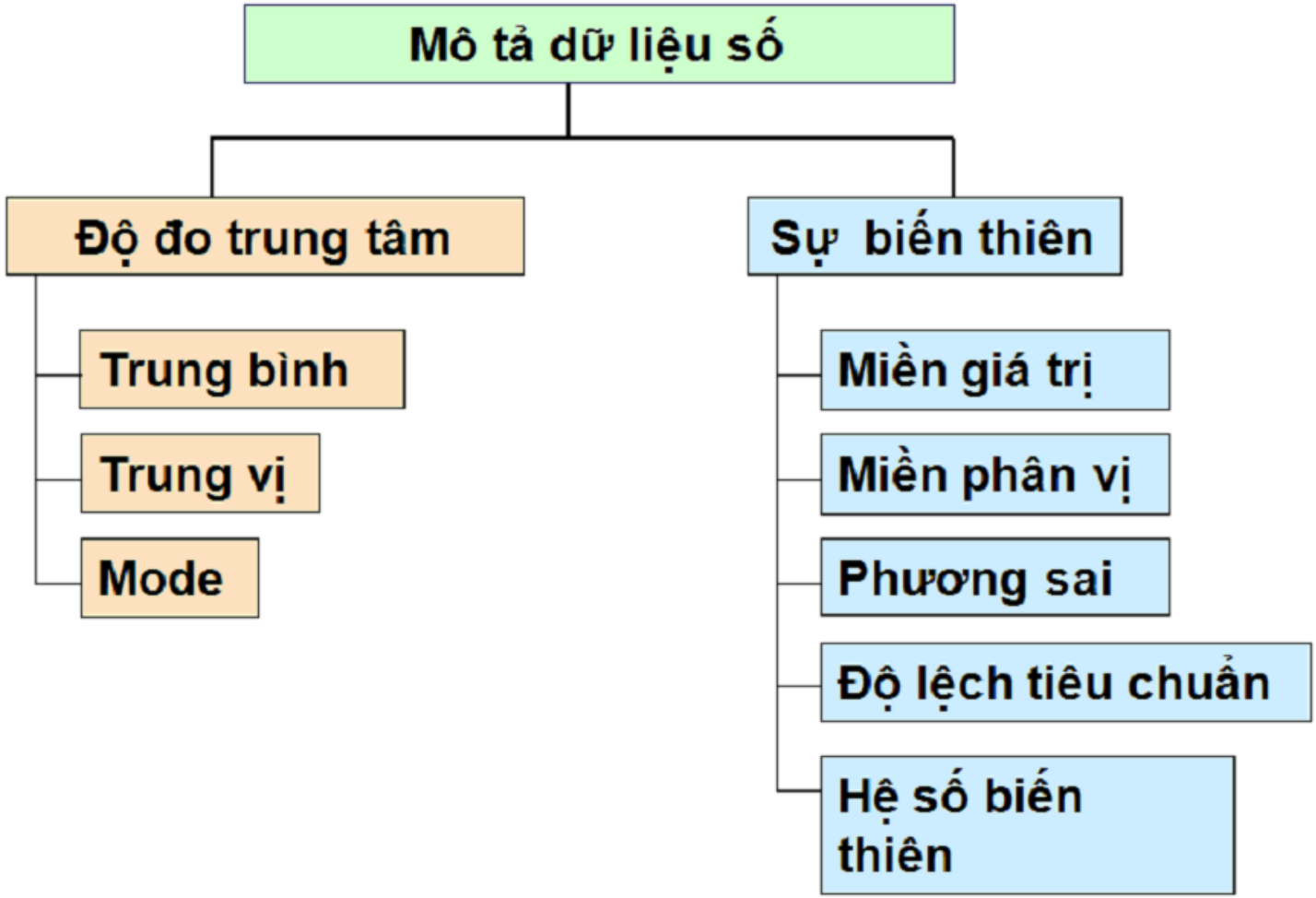
- **Dữ liệu định tính**: thu thập từ thang đo danh nghĩa và thứ bậc $\Rightarrow$ không tính được trị trung bình.
- **Dữ liệu định lượng**: thu thập từ thang đo khoảng cách và tỷ lệ $\Rightarrow$ tính được trị trung bình.



Plan

Mô tả dữ liệu một biến bằng phương pháp số

Giới thiệu


Mô tả dữ liệu số
Độ đo trung tâm
Trung bình
Trung vị
Mode
Sự biến thiên
Miền giá trị
Miền phân vị
Phương sai
Độ lệch tiêu chuẩn
Hệ số biến thiên

Các độ đo hướng tâm

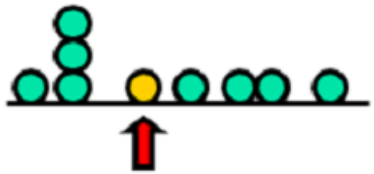
Độ đo trung tâm

Trung bình

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

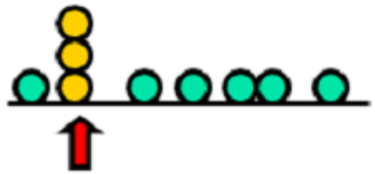
Trung bình số học

Trung vị



Điểm chính giữa của dữ liệu đã sắp xếp

Mode



Giá trị thường gặp nhất

Trung bình

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

Definition 1

Giả sử ta có dữ liệu (của tổng thể hoặc mẫu) là $x_1, x_2, \ldots, x_n$. Khi đó, trung bình (của tổng thể hoặc mẫu) là trung bình cộng của các phần tử trong dữ liệu, tức là

$$\frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Ta sẽ ký hiệu tổng này là μ (tương ứng $\bar{x}$) nếu dữ liệu là của tổng thể (tương ứng, của mẫu).

Nhận xét 1

Trường hợp dữ liệu có tần số như trong bảng sau

Giá trị dữ liệu	x_1	x_2	...	x_k
Tần số tương ứng	n_1	n_2	...	n_k

Trong đó, $n_1 + n_2 + \ldots + n_k = n$.
Khi đó, trung bình (tổng thể hoặc mẫu) được tính theo công thức

$$\frac{\sum_{i=1}^{k} n_i x_i}{n} \tag{2}$$

Nhận xét 2

Khi dữ liệu được trình bày dưới dạng khoảng như sau

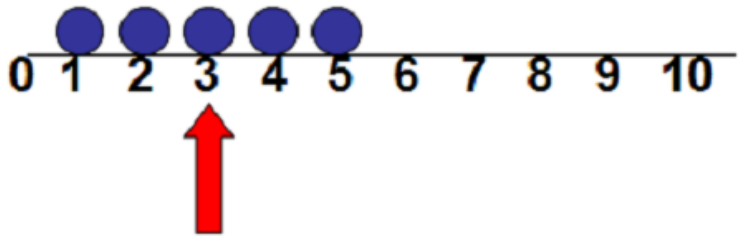
Giá trị dữ liệu	$< a_1$	$[a_1, b_1[$	$\ldots$	$[a_k, b_k[$	$\geq b_k$
Tần số tương ứng	n_1	n_2	$\ldots$	n_{k+1}	n_{k+2}

Bảng 1: Dữ liệu dưới dạng khoảng

Giả sử rằng độ rộng các khoảng là như nhau, tức là $b_i - a_i = c$ với mọi i. Khi đó, mỗi khoảng ta thay bằng điểm chính giữa của khoảng, riêng hai khoảng đầu và cuối ta thay bằng $a_1 - c/2$ và $b_k + c/2$. Sau đó, dùng công thức (2) để tính trung bình.

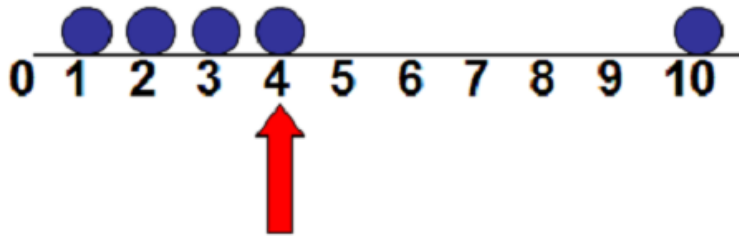
Trung bình

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

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

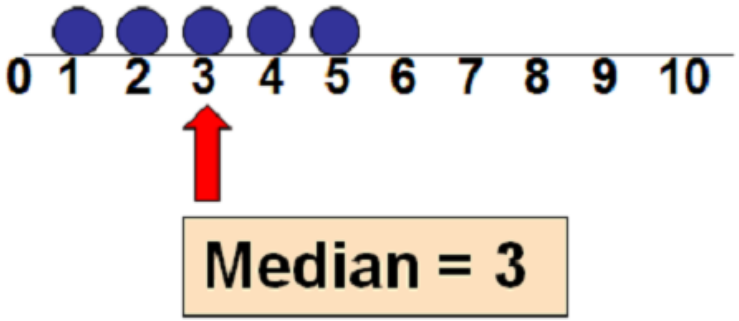
Trung vị mẫu

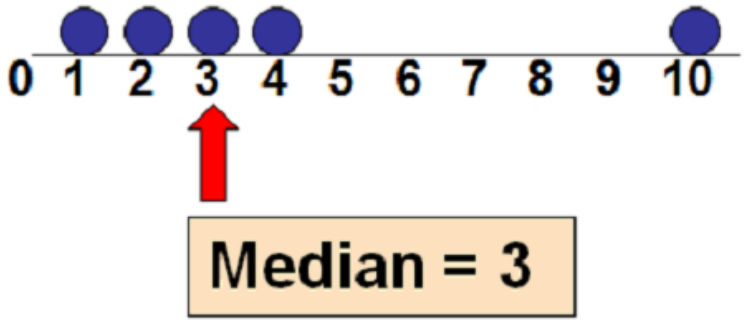
Definition 2

Trung vị mẫu (sample median) là giá trị chia các quan sát thành hai phần bằng nhau. Một phần chứa các quan sát nhỏ hơn trung vị và phần còn lại chứa các quan sát lớn hơn trung vị.

Nhận xét 3

Trung vị không bị ảnh hưởng bởi các điểm outlier.





Trung vị mẫu

Cách tìm trung vị

Sắp xếp mẫu theo thứ tự tăng dần.

- Nếu kích thước mẫu là lẻ thì **trung vị** là giá trị ở vị trí trung tâm của mẫu được sắp
- Nếu kích thước mẫu là chẵn thì **trung vị** là trung bình của hai giá trị ở vị trí trung tâm của mẫu được sắp

Nói cách khác, gọi n là kích thước mẫu và $i = (n+1)/2$, thì

- Nếu n lẻ thì **trung vị** $= x_i$
- Nếu n chẵn thì **trung vị** $= \dfrac{x_{[i]} + x_{[i]+1}}{2}$, với $[i]$ là phần nguyên của i.

Đối với dữ liệu dạng khoảng (xem bảng 1)

Trước hết ta phải xác định khoảng đầu tiên $[a_i, b_i]$ có tần suất tích lũy, F_i, lớn hơn 0.5.
Sau đó, trung vị được tính theo công thức

$$a_i + (0.5 - F_{i-1}) \times \frac{b_i - a_i}{F_i - F_{i-1}}$$

Mode

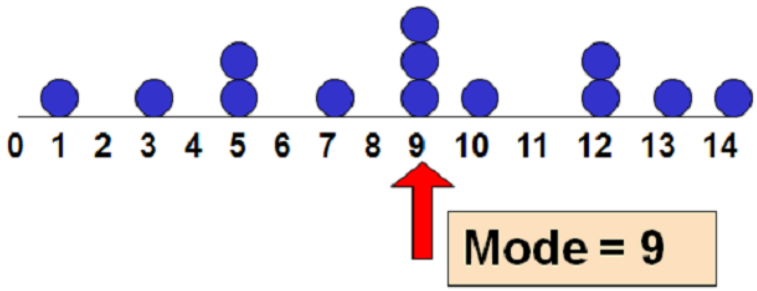
Definition 3

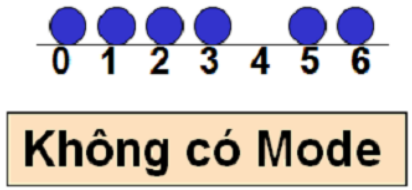
Mode của dữ liệu là giá trị của dữ liệu có tần số xuất hiện lớn nhất. Nếu mọi giá trị dữ liệu đều có cùng tần số, ta nói dữ liệu không có mode.

Nhận xét 4

- Mode không bị ảnh hưởng bởi các điểm outlier
- Mode có thể sử dụng cho cả dữ liệu số và dữ liệu phân loại
- Trường hợp dữ liệu dạng khoảng (xem bảng 1), thì mode của dữ liệu là điểm chính giữa của khoảng có tần số lớn nhất.

Mode

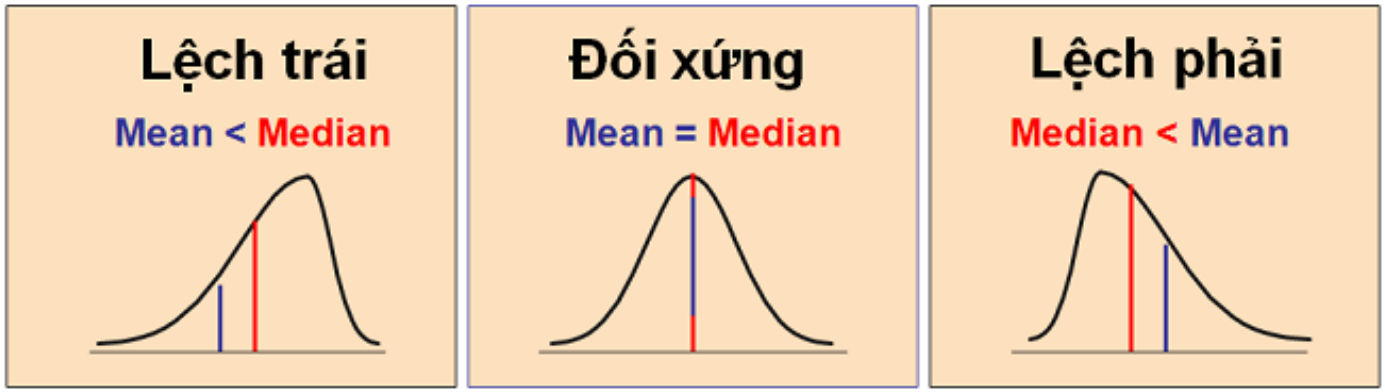




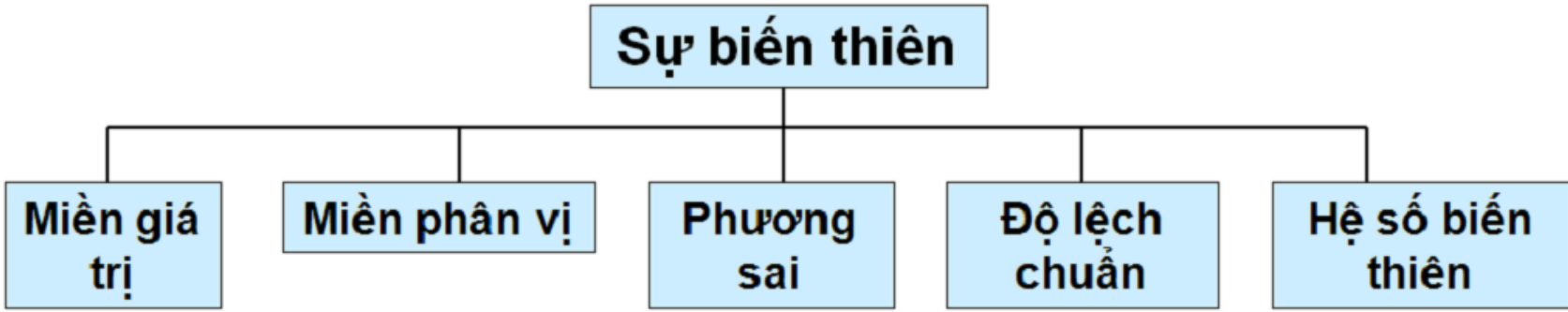
So sánh trung bình, trung vị và mode

- Nếu dữ liệu có phân phối đối xứng, thì trung bình và trung vị sẽ bằng nhau và rơi vào tâm của phân phối.
- Nếu dữ liệu có phân phối bị lệch (skewed) (tức là bất đối xứng, với một đuôi kéo dài về một phía), thì trung bình và trung vị đều bị kéo về phía đuôi dài hơn, nhưng trung bình, thông thường, được kéo xa hơn trung vị.
- Cụ thể, nếu phân phối là lệch phải thì mode $<$ trung vị $<$ trung bình; ngược lại, nếu phân phối là lệch trái thì mode $>$ trung vị $>$ trung bình.

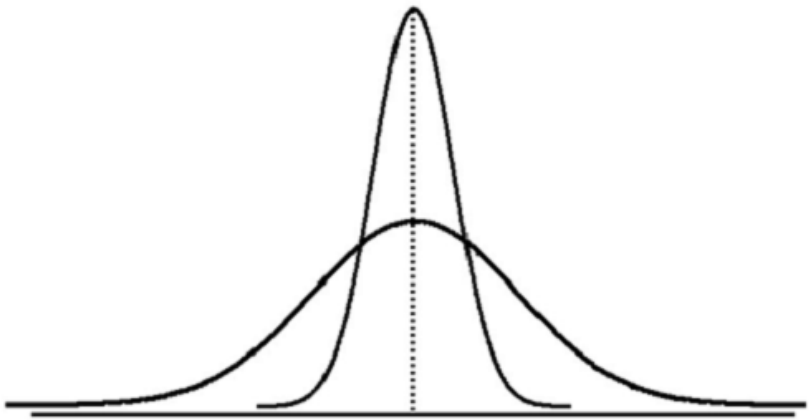
So sánh trung bình, trung vị và mode



Độ đo sự biến thiên của dữ liệu



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



Cùng trung tâm, sự biến thiên khác nhau

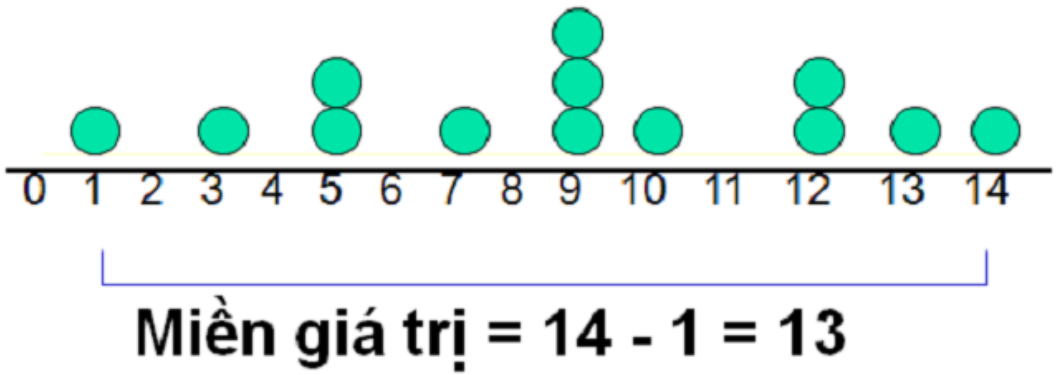
Miền giá trị mẫu (sample range)

Definition 4

Miền giá trị mẫu là khoảng cách giữa giá trị lớn nhất và giá trị nhỏ nhất trong mẫu.

Nếu n quan sát trong một mẫu được kí hiệu là $x_1, x_2, \ldots, x_n$ thì **miền giá trị mẫu** là

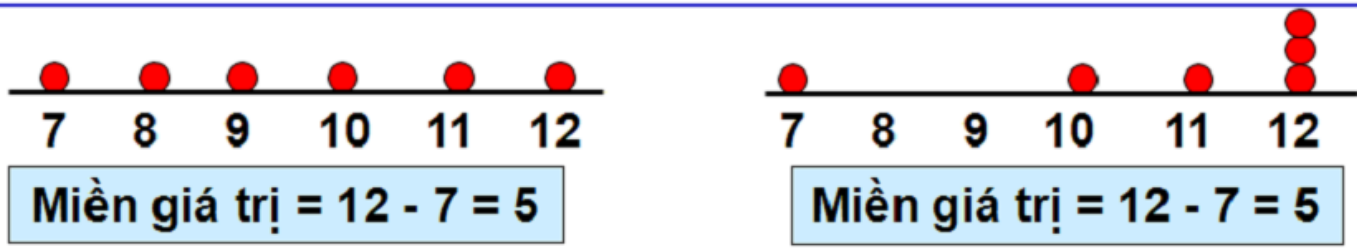
$$r = \max(x_i) - \min(x_i) \tag{3}$$



Miền giá trị mẫu

Nhược điểm

- Bỏ qua phân bố của dữ liệu



- Bị ảnh hưởng bởi các điểm oulier

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

Miền giá trị = 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

Miền giá trị = 120 - 1 = 119

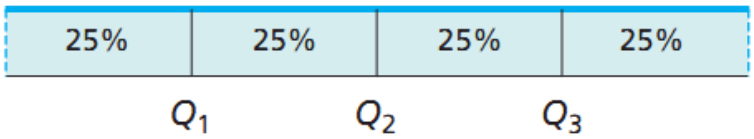
Tứ phân vị

Definition 5

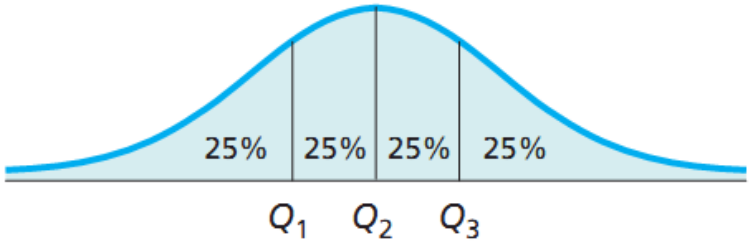
Nếu ta chia dữ liệu thành 4 phần bằng nhau. Các điểm chia này được gọi là **các tứ phân vị** (quartiles).

- Tứ phân vị đầu tiên, Q_1, là giá trị có xấp xỉ 25% số quan sát nằm bên dưới nó và xấp xỉ 75% số quan sát nằm trên nó.
- Tứ phân vị thứ hai, Q_2, có xấp xỉ 50% số quan sát nằm bên dưới nó, tứ phân vị thứ hai chính là trung vị.
- Tứ phân vị thứ ba, Q_3, là giá trị có xấp xỉ 75% số quan sát nằm bên dưới nó.

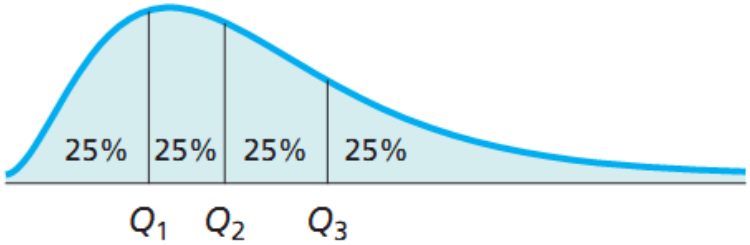
Các tứ phân vị cho một số phân phối



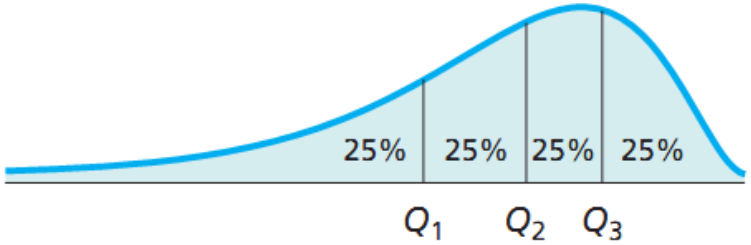
(a) Đều



(b) Dạng chuông



(c) Lệch phải



(d) Lệch trái

Tứ phân vị

Cách tìm tứ phân vị

Sắp xếp dữ liệu (kích thước n) theo thứ tự tăng dần $x_1, x_2, \ldots, x_n$.

Gọi q_1, q_2, q_3 lần lượt là phân vị thứ nhất, thứ hai, thứ ba của dữ liệu và

$$\begin{aligned} k_1 &= 0.25(n+1) \\ k_2 &= 0.5(n+1) \\ k_3 &= 0.75(n+1) \end{aligned}$$

Khi đó,

$$q_i = \begin{cases} x_{k_i} & \text{nếu } k_i \text{ nguyên} \\ \dfrac{x_{[k_i]} + x_{[k_i]+1}}{2} & \text{nếu ngược lại} \end{cases}, \quad i = 1, 2, 3$$

Khoảng tứ phân vị (interquartile range - IQR)

Definition 6

Khoảng tứ phân vị (IQR) là khoảng cách giữa tứ phân vị đầu tiên và tứ phân vị thứ ba; tức là, $IQR = Q_3 - Q_1$.

Nhận xét 5

- Người ta thường sử dụng IQR để đo sự biến thiên của dữ liệu khi trung vị được sử dụng để đo trung tâm của dữ liệu.
- Tương tự trung vị, IQR không bị ảnh hưởng bởi các điểm outlier.

Ví dụ

Example 7

Một công ty truyền thông khảo sát thói quen xem ti vi của một cộng đồng dân cư. 20 người được chọn ngẫu nhiên và có thời gian (giờ) xem ti vi hàng tuần như sau:

25	41	27	32	43
66	35	31	15	5
34	26	32	38	16
30	38	30	20	21

a. Tìm các tứ phân vị của dữ liệu trên?

b. Tìm khoảng tứ phân vị?

Dữ liệu outlier

Definition 8

- Dữ liệu nằm ngoài khoảng $[Q_1 - 1.5IQR; Q_3 + 1.5IQR]$ được gọi là **outlier**.
- Dữ liệu nằm ngoài khoảng $[Q_1 - 3IQR; Q_3 + 3IQR]$ được gọi là **extreme outlier**.

Nguyên nhân xuất hiện dữ liệu outlier

(1) lỗi ghi chép; (2) đo đạc sai; (3) một dữ liệu thuộc tổng thể khác bị trộn lẫn vào; (4) một dữ liệu cực trị (quá lớn hoặc quá nhỏ) bất thường,v.v.

Dữ liệu outlier

Nhận xét 6

- Các dữ liệu cực trị có thể không phải là outlier vì nó có thể là dấu hiệu của tổng thể bị lệch.
- Khi quan sát một giá trị oulier, cố gắng xác định nguyên nhân gây ra nó.
- Nếu giá trị outlier là do sai sót trong đo đạc hoặc lỗi ghi chép, hoặc vì một lí do nào đó mà rõ ràng nó không thuộc vào tập dữ liệu, thì giá trị outlier này có thể được loại bỏ một cách dễ dàng.
- Tuy nhiên, nếu không thể giải thích rõ ràng giá trị outlier này, đôi khi rất khó quyết định có nên giữ lại nó trong tập dữ liệu hay không.

Example 9

Xét dữ liệu về thời gian xem phim hàng tuần trong Ví dụ 7. Xác định các giá trị outlier (nếu có)?

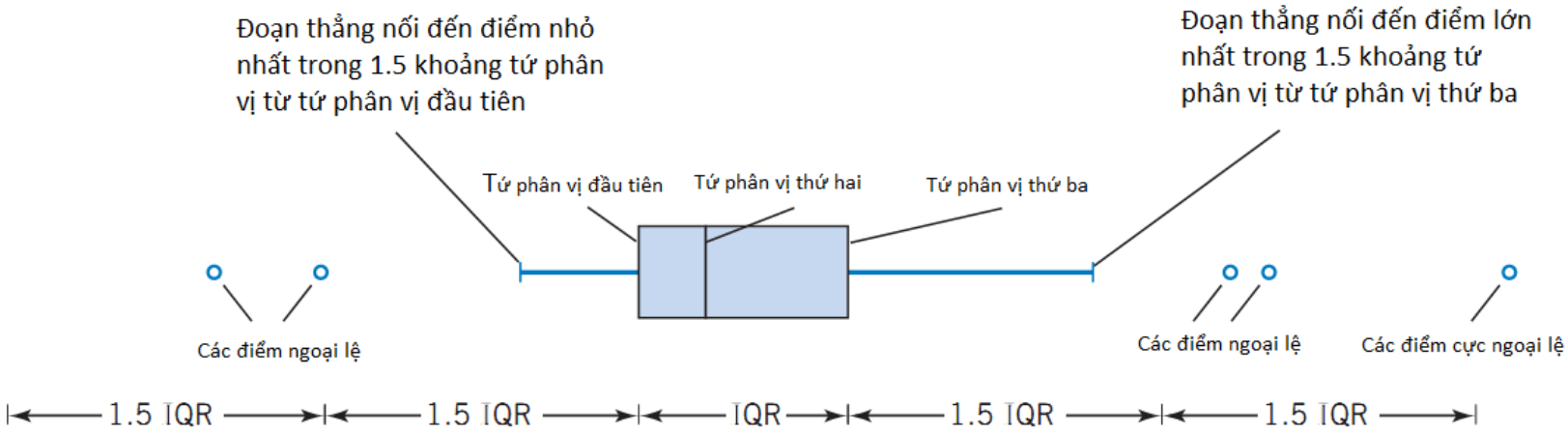
Đồ thị dạng hộp (boxplot)

Đồ thị dạng hộp (**boxplot** hoặc **box-and-whisker diagram**) được sử dụng để mô tả đồng thời, bằng hình ảnh, về trung tâm và sự biến thiên của dữ liệu.

Xây dựng đồ thị dạng hộp

31. Xác định Q_1, Q_2, Q_3 và $IQR = Q_3 - Q_1$
32. Xác định các điểm outlier và extreme outlier (nếu có)
33. Vẽ một trục tọa độ ngang (hoặc dọc), và vẽ các đoạn thẳng tại Q_1, Q_2 và Q_3. Đóng khung các đoạn thẳng này trong một hộp.
34. Vẽ một đoạn thẳng từ Q_1 đến giá trị dữ liệu nhỏ nhất nhưng lớn hơn $Q_1 - 1.5IQR$. Vẽ một đoạn thẳng từ Q_3 đến giá trị dữ liệu lớn nhất nhưng nhỏ hơn $Q_3 + 1.5IQR$.
35. Đánh dấu các điểm outlier và extreme outlier.

Đồ thị dạng hộp (boxplot)



Chú ý

Đôi khi, các kí hiệu khác nhau, chẳng hạn các hình tròn được tô và không tô được dùng để xác định hai loại điểm ngoại lệ này.

Đồ thị dạng hộp (boxplot)

Example 10

Vẽ đồ thị dạng hộp cho dữ liệu thời gian xem ti vi hàng tuần trong Ví dụ 7.

Giải

Sắp xếp dữ liệu theo thứ tự tăng dần

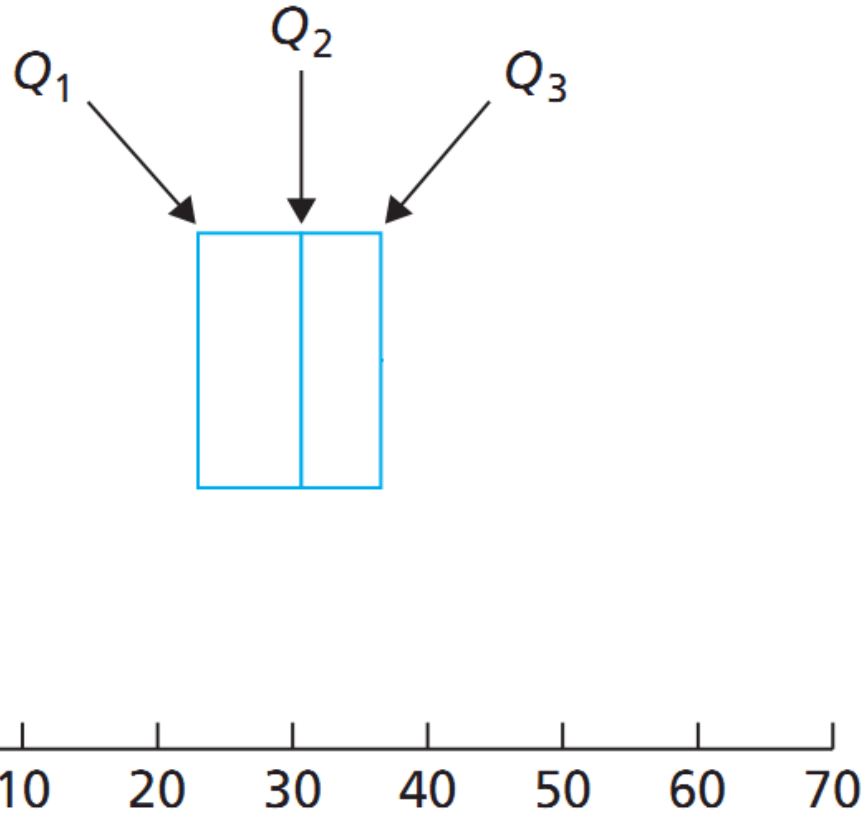
5 15 16 20 21 25 26 27 30 30 31 32 32 34 35 38 38 41 43 66

31. **Xác định** Q_1, Q_2, Q_3 **và** $IQR = Q_3 - Q_1$**.**
    $Q_1 = 23, Q_2 = 30.5, Q_3 = 36.5$, và $IQR = 13.5$
32. **Xác định các điểm outlier và extreme outlier (nếu có)**
    66

Đồ thị dạng hộp (boxplot)

Ví dụ 10 (tt)

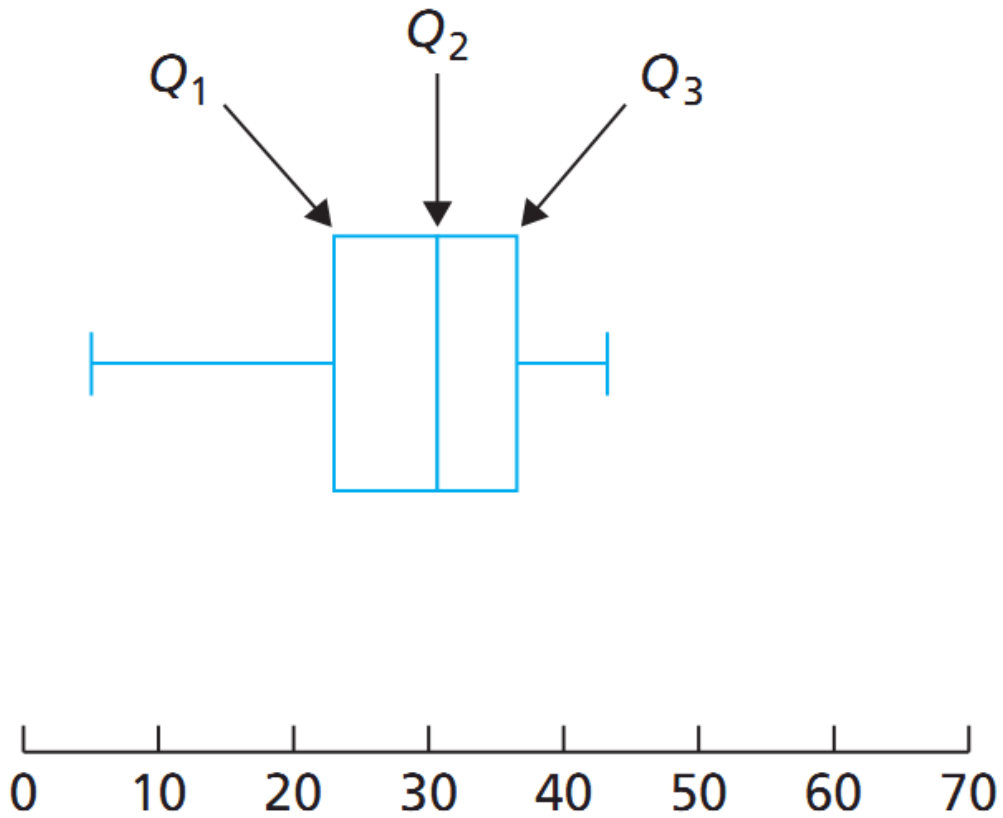
B3. **Vẽ một trục tọa độ ngang (hoặc dọc), và vẽ các đoạn thẳng tại Q_1, Q_2 và Q_3. Đóng khung các đoạn thẳng này trong một hộp.**



Đồ thị dạng hộp (boxplot)

Ví dụ 10 (tt)

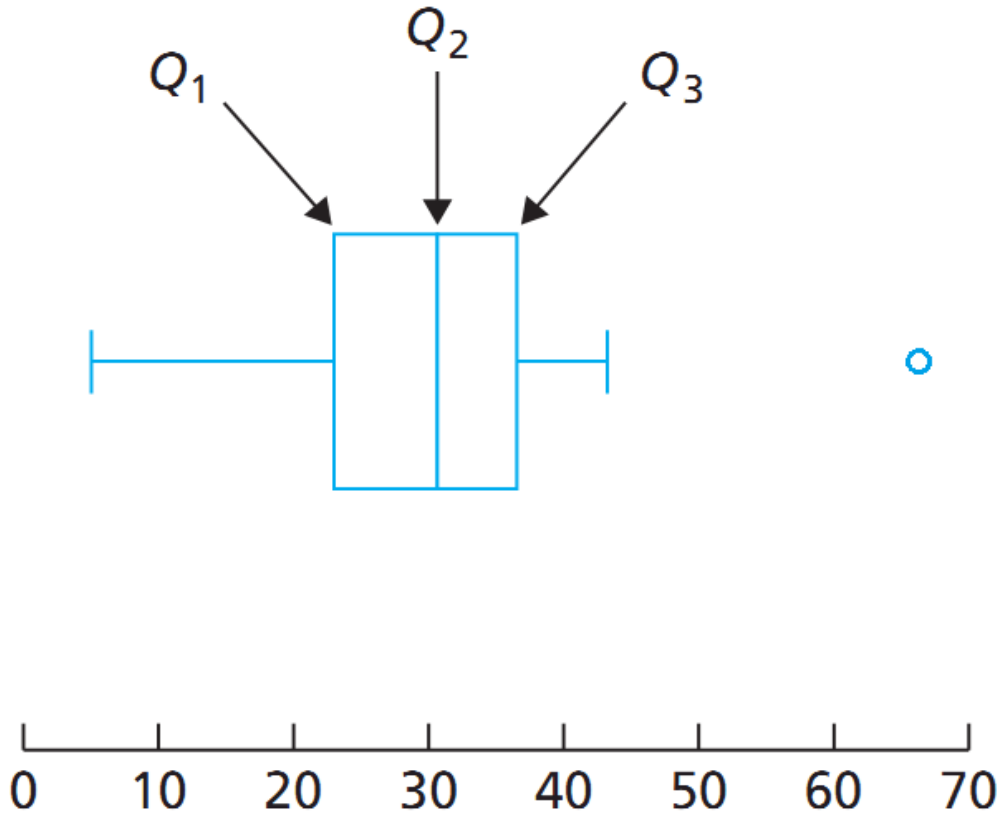
B4. **Vẽ một đoạn thẳng từ Q_1 đến giá trị dữ liệu nhỏ nhất nhưng lớn hơn $Q_1 - 1.5IQR$. Vẽ một đoạn thẳng từ Q_3 đến giá trị dữ liệu lớn nhất nhưng nhỏ hơn $Q_3 + 1.5IQR$.**



Đồ thị dạng hộp (boxplot)

Ví dụ 10 (tt)

B5. **Đánh dấu các điểm outlier và extreme outlier.**



Đồ thị dạng hộp (boxplot)

Nhận xét 7

Người ta thường sử dụng đồ thị dạng hộp để so sánh hai hay nhiều tập dữ liệu. Để so sánh thì tất cả các đồ thị dạng hộp phải sử dụng cùng thang đo.

Runners			Others			
7.3	6.7	8.7	24.0	19.9	7.5	18.4
3.0	5.1	8.8	28.0	29.4	20.3	19.0
7.8	3.8	6.2	9.3	18.1	22.8	24.2
5.4	6.4	6.3	9.6	19.4	16.3	16.3
3.7	7.5	4.6	12.4	5.2	12.2	15.6

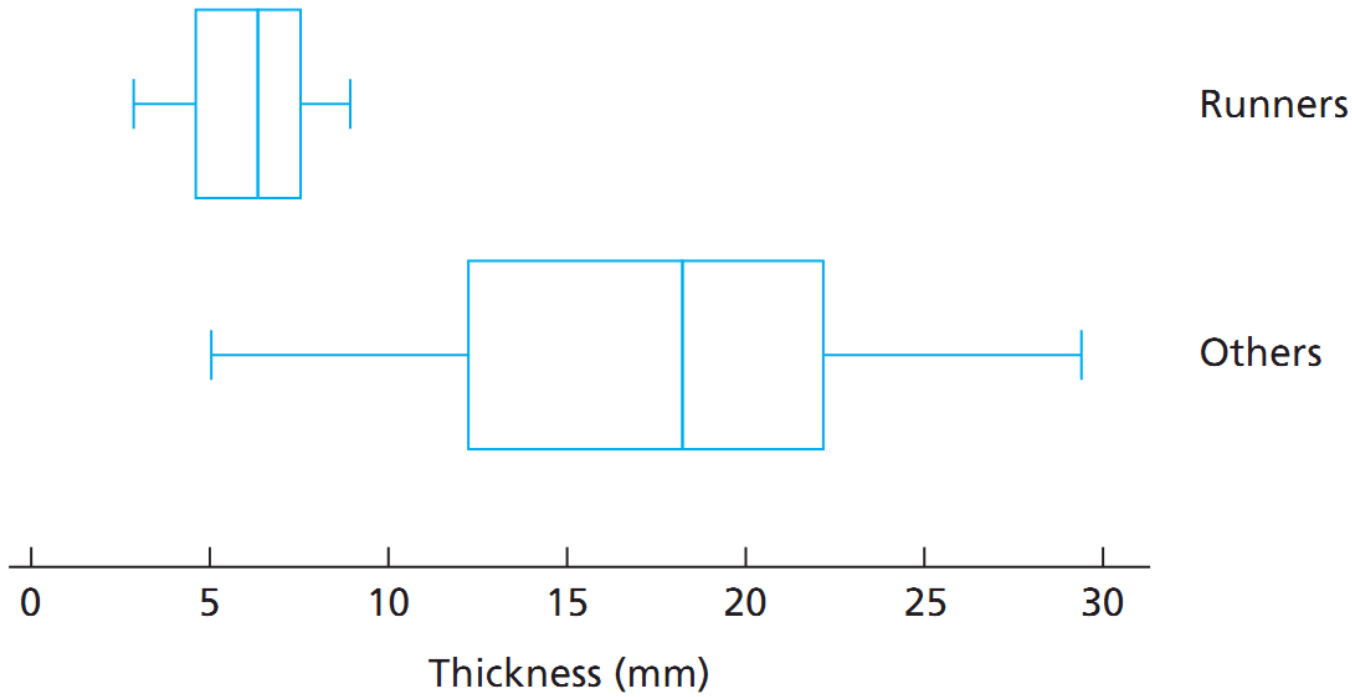
Bảng 2: Độ dày nếp gấp da

Đồ thị dạng hộp (boxplot)

So sánh các tập dữ liệu bằng cách sử dụng đồ thị dạng hộp

Example 11 (Độ dày nếp gấp da (skinfold thickness))

Một nghiên cứu có tiêu đề "Thành phần cơ thể của những vận động viên chạy nước rút" được thực hiện bởi M. Pollock et al. để xác định xem những vận động viên chạy nước rút có thực sự nhẹ cân hơn những người khác hay không. Các kết quả của họ được xuất bản trong *The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies* (P. Milvey (ed.), New York: New York Academy of Sciences, p. 366). Các nhà nghiên cứu đã đo độ dày nếp gấp da, một chỉ số gián tiếp về độ phì cơ thể, của các mẫu những người chạy nước rút và những người khác trong cùng nhóm tuổi. Dữ liệu mẫu, theo mm, được trình bày bên dưới. Sử dụng đồ thị dạng hộp để so sánh hai tập dữ liệu này, tập trung vào trung tâm và sự biến thiên.



Nhận xét 8

- Về mặt trung bình, mẫu những người chạy nước rút có độ dày nếp gấp da nhỏ hơn mẫu những người khác.
- Độ dày nếp gấp da trong mẫu những người chạy nước rút có sự biến thiên nhỏ hơn nhiều so với trong mẫu những người khác.

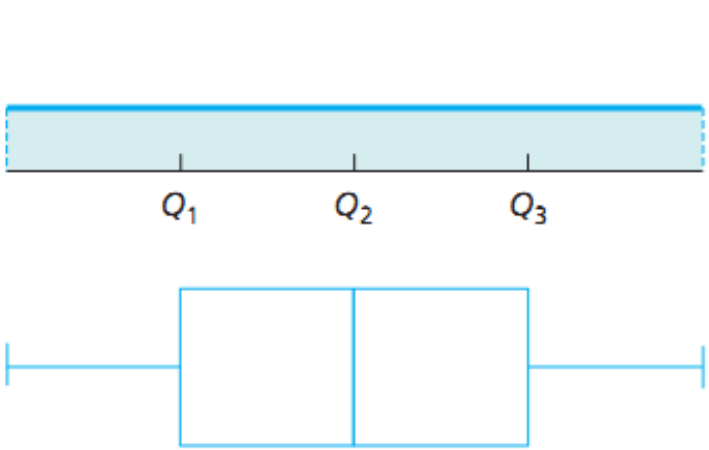
Đồ thị dạng hộp (boxplot)

Hình dạng của phân phối

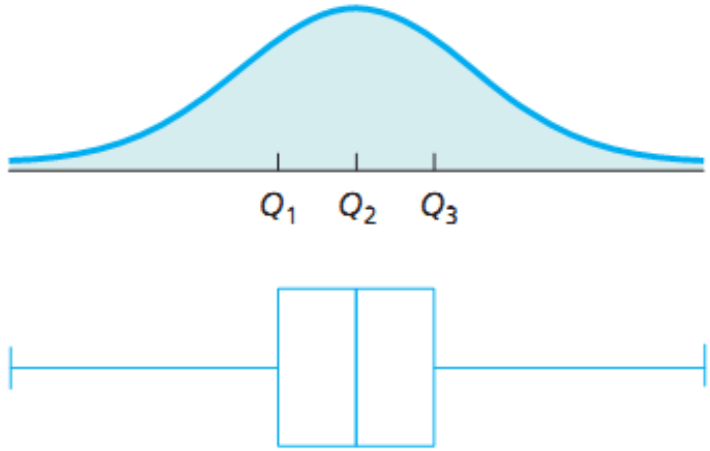
- Đồ thị dạng hộp có thể được dùng để xác định hình dạng xấp xỉ của phân phối của tập dữ liệu.
- Với kích thước mẫu lớn, đồ thị dạng hộp xác định hình dạng của phân phối một cách hiệu quả nhất.
- Với kích thước mẫu nhỏ, đồ thị dạng hộp không đáng tin cậy trong việc xác định hình dạng của phân phối; trường hợp này ta nên sử dụng đồ thị stem-leaf sẽ tốt hơn.

Đồ thị dạng hộp (boxplot)

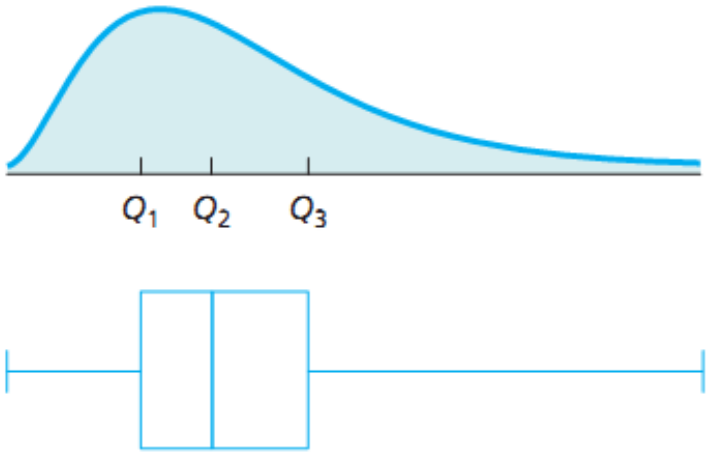
Hình dạng của phân phối



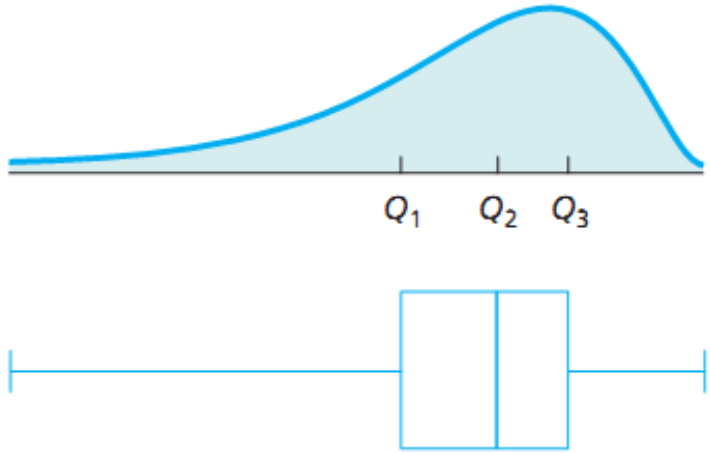
(a) Đều



(b) Dạng chuông



(c) Lệch phải



(d) Lệch trái

Phương sai

Definition 12

Nếu $x_1, x_2, \ldots, x_N$ là các phần tử của tổng thể, thì **phương sai tổng thể** là

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N} \tag{4}$$

Độ lệch chuẩn tổng thể là $\sigma = \sqrt{\sigma^2}$.

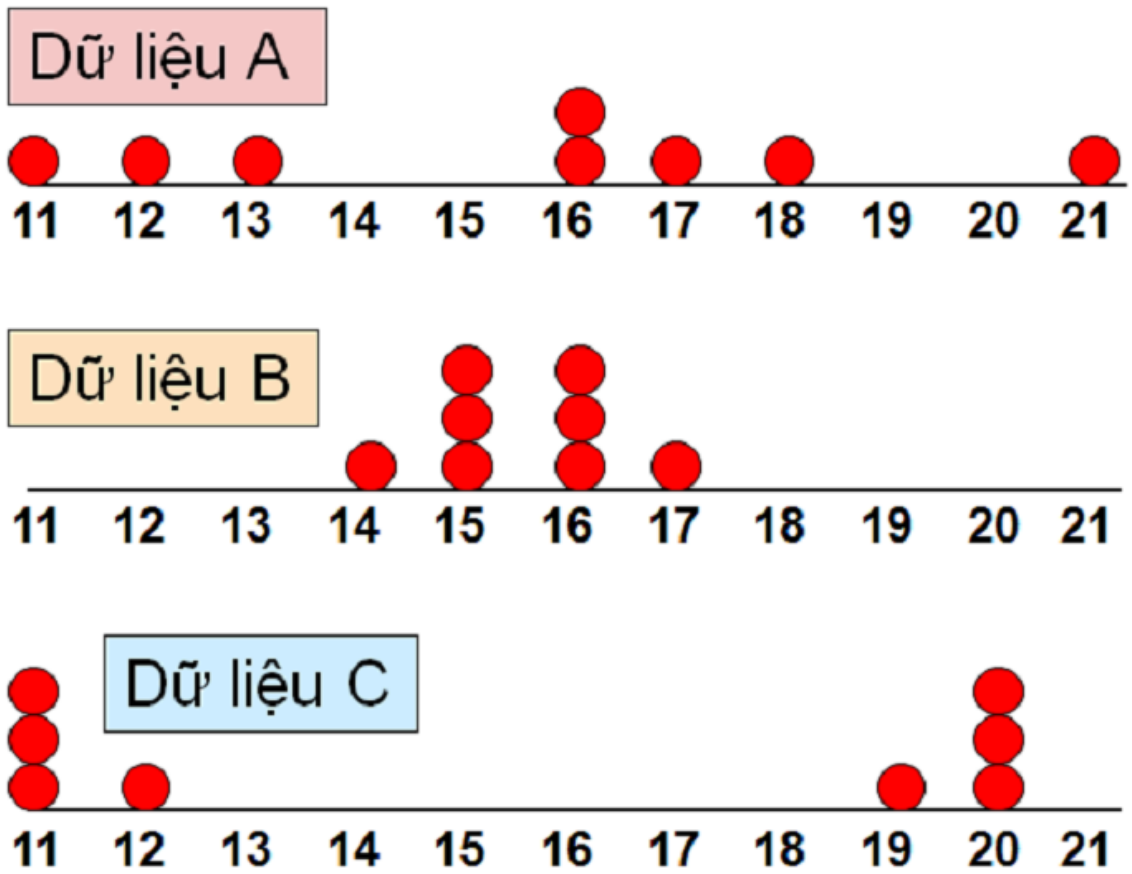
Definition 13

Nếu $x_1, x_2, \ldots, x_n$ là một mẫu có n quan sát, thì **phương sai mẫu** là

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{5}$$

Độ lệch chuẩn mẫu là $s = \sqrt{s^2}$.

So sánh các độ lệch chuẩn



Mean = 15.5
s = 3.338

Mean = 15.5
s = 0.926

Mean = 15.5
s = 4.570