Chương

03

MÔ TẢ, KHÁM PHÁ & SO SÁNH DỮ LIỆU

NỘI DUNG

- Độ đo xu hướng tập trung (central tendency)
- Độ đo sự phân tán (variability)
- Độ đo vị trí tương đối & Đồ thị hộp (boxplots)

NHẮC LẠI KIẾN THỨC

- **Chương 1**: phân biệt giữa quần thể & mẫu, tham số & số liệu thống kê, một số phương pháp lấy mẫu tốt.
- **Chương 2**: đồ thị stem & leaf, phân phối tần số, tổng hợp dữ liệu bằng các đồ thị, mô tả trung tâm, sự phân tán, phân phối, ngoại lệ, và sự thay đổi đặc tính của dữ liệu theo thời gian

MỤC TIÊU

- **Thống kê mô tả** *(Descriptive Statistics)*: trong chương này chúng ta sẽ mô tả những đặc tính quan trọng của tập dữ liệu mẫu bằng các số liệu thống kê như giá trị trung bình, độ lệch chuẩn…
- **Thống kê suy diễn** *(Inferential Statistics)*: ở những chương sau chúng ta sử dụng dữ liệu mẫu để thực hiện suy diễn để biết các thông tin quần thể

NỘI DUNG

- **Độ đo xu hướng tập trung (central tendency)**
- Độ đo sự phân tán (variability)
- Độ đo vị trí tương đối & Đồ thị hộp (boxplots)

- Để đo xu hướng hướng tâm của dữ liệu, người ta thường sử dụng các độ đo sau:
  - trung bình
  - trung vị
  - yếu vị

Trung bình (mean)

- Giá trị trung bình: được tính bằng cách cộng tất cả các giá trị dữ liệu và chia cho số lượng các giá trị đã cộng vào.
- Một số ký hiệu:
  - Σ: tổng tất cả các giá trị
  - x: biến đại diện cho mỗi giá trị dữ liệu
  - n: số lượng phần tử của mẫu
  - N: số lượng phần tử của quần thể

Công thức tính

$\overline{x}$: giá trị trung bình của tập mẫu

$$\overline{x} = \frac{\Sigma x}{n}$$

μ : giá trị trung bình của tất cả các giá trị của quần thể

$$\mu = \frac{\Sigma x}{N}$$

Trung bình (mean)

- Ưu điểm:
  - Tất cả các giá trị của dữ liệu điều được dùng để tính trung bình.
  - Trung bình mẫu có xu hướng chênh lệch không nhiều so với trung bình của quần thể.
- Khuyết điểm:
  - Nhạy cảm với các giá trị ngoại lệ (outliers)

Ví dụ

- Kiểm tra 5 bịch sôcôla khác nhau, người ta thấy số mẫu sôcôla ở mỗi bịch lần lượt là: 22, 22, 26, 24, 23. Tìm số mẫu sôcôla trung bình của các bịch trên.
- *Giải:*

$$\overline{x} = \frac{\Sigma x}{n} = \frac{22+22+26+24+23}{5} = \frac{117}{5}$$

$$= 23.4 \text{ mẫu}$$

Ví dụ: tính trung bình từ bảng phân phối tần số

- Tính điểm IQ trung bình của bảng sau:

IQ Score	Frequency f	Class Midpoint x	$f \cdot x$
50–69	2	59.5	119.0
70–89	33	79.5	2623.5
90–109	35	99.5	3482.5
110–129	7	119.5	836.5
130–149	1	139.5	139.5
Totals:	$\Sigma f = 78$		$\Sigma(f \cdot x) = 7201.0$

$$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f} = \frac{7201.0}{78} = 92.3$$

Trung bình trọng số

- Khi mỗi giá trị dữ liệu có một trọng số khác nhau, w. Chúng ta có thể tính trung bình trọng số bằng công thức sau:

$$\overline{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$$

Ví dụ

- Kết quả học tập trong một học kỳ của một sinh viên như sau: môn 1 (điểm A, 3 chỉ); môn 2 (điểm A, 4 chỉ); môn 3 (điểm B, 3 chỉ); môn 4 (điểm C, 3 chỉ); môn 5 (điểm F, 1 chỉ).
- Hệ thống quy đổi điểm từ hệ chữ sang hệ số như sau:

  A=4, B=3, C=2, D=1, F=0
- Hãy tính điểm trung bình của sinh viên trên.

- *Hướng dẫn*: xem số tín chỉ như là trọng số, sử dụng điểm đã quy đổi từ hệ chữ để tính điểm trung bình

Ví dụ

$$\bar{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$$

$$= \frac{(3\times 4)+(4\times 4)+(3\times 3)+(3\times 2)+(1\times 0)}{3+4+3+3+1}$$

$$= \frac{43}{14} = 3.07$$

Trung vị (median)

- Giá trị trung vị *(median)*: là giá trị nằm chính giữa khi tập dữ liệu được sắp xếp (tăng dần hoặc giảm dần)

- Ký hiệu: $\tilde{x}$ *(gọi là x ngã)*

- Trung vị không bị ảnh hưởng bởi các giá trị ngoại lệ.

Trung vị (median)

- Cách tính trung vị:
  1. Sắp xếp dữ liệu (tăng dần hoặc giảm dần)
  2. Nếu số lượng giá trị dữ liệu là lẻ, thì trung vị chính là giá trị nằm chính giữa
  3. Nếu số lượng giá trị dữ liệu là chẵn, thì trung vị được tính là trung bình của hai giá trị nằm chính giữa

Ví dụ

Ví dụ: Tìm trung vị

5.40 1.10 0.42 0.73 0.48 1.10 0.66

Sắp xếp:

0.42 0.48 0.66 0.73 1.10 1.10 5.40

Median is 0.73

Ví dụ

Ví dụ: Tìm trung vị:

5.40 1.10 0.42 0.73 0.48 1.10

Sắp xếp

0.42 0.48 0.73 1.10 1.10 5.40

$$\frac{0.73 + 1.10}{2}$$

Median is 0.915

Yếu vị (mode)

- Yếu vị *(mode)*: là giá trị xuất hiện nhiều lần nhất.
- Một tập dữ liệu có thể có một hoặc nhiều yếu vị, hoặc có thể không có
- Yếu vị chỉ được dùng làm độ đo xu hướng hướng tâm đối với dữ liệu không có thứ tự

Ví dụ

Ví dụ:

a. 5.40 1.10 0.42 0.73 0.48 1.10 ⇦ Mode is 1.10

b. 27 27 27 55 55 55 88 88 99 ⇦ Bimodal - 27 & 55

c. 1 2 3 6 7 8 9 10 ⇦ No Mode

NỘI DUNG

- Độ đo xu hướng tập trung (central tendency)
- **Độ đo sự phân tán (variability)**
- Độ đo vị trí tương đối & Đồ thị hộp (boxplots)

Miền giá trị (range)

- Miền giá trị (range): là độ lệch giữa giá trị lớn nhất và giá trị bé nhất của dữ liệu.
- Miền giá trị: là độ đo sự biến thiên đơn giản nhất
- Miền giá trị dễ bị ảnh hưởng bởi các giá trị ngoại lệ (outliers)

Độ lệch chuẩn (standard deviation) của mẫu

- Độ lệch chuẩn (standard deviation) của một tập dữ liệu mẫu, thường được ký hiệu là s (hoặc sd) là đại lượng chỉ độ lệch của các giá trị dữ liệu so với giá trị trung bình.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

Độ lệch chuẩn (standard deviation) của mẫu

- Trong thực nghiệm, có thể tính nhanh giá trị độ lệch chuẩn bằng công thức sau:

$$s = \sqrt{\frac{n\left(\Sigma x^2\right) - (\Sigma x)^2}{n(n-1)}}$$

Một vài lưu ý

- Giá trị của độ lệch chuẩn luôn không âm
- Giá trị của độ lệch chuẩn vẫn bị ảnh hưởng bởi ngoại lệ
- Đơn vị của độ lệch chuẩn cũng là đơn vị của giá trị dữ liệu

Ví dụ

- Tìm độ lệch chuẩn của bộ dữ liệu sau:

$$22, 22, 26, 24$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(22-23.5)^2 + (22-23.5)^2 + (26-23.5)^2 + (24-23.5)^2}{4-1}}$$

$$= \sqrt{\frac{11}{3}} = 1.9149$$

Độ lệch chuẩn(standard deviation) của quần thể

- Tương tự như độ lệch chuẩn của mẫu, ta có công thức tính độ lệch chuẩn của quần thể.

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Phương sai

- Phương sai của một tập dữ liệu là bình phương của giá trị độ lệch chuẩn
- Phương sai mẫu: s^2
- Phương sai của quần thể: σ^2

Ký hiệu

s : độ lệch chuẩn của mẫu

s^2 : phương sai của mẫu

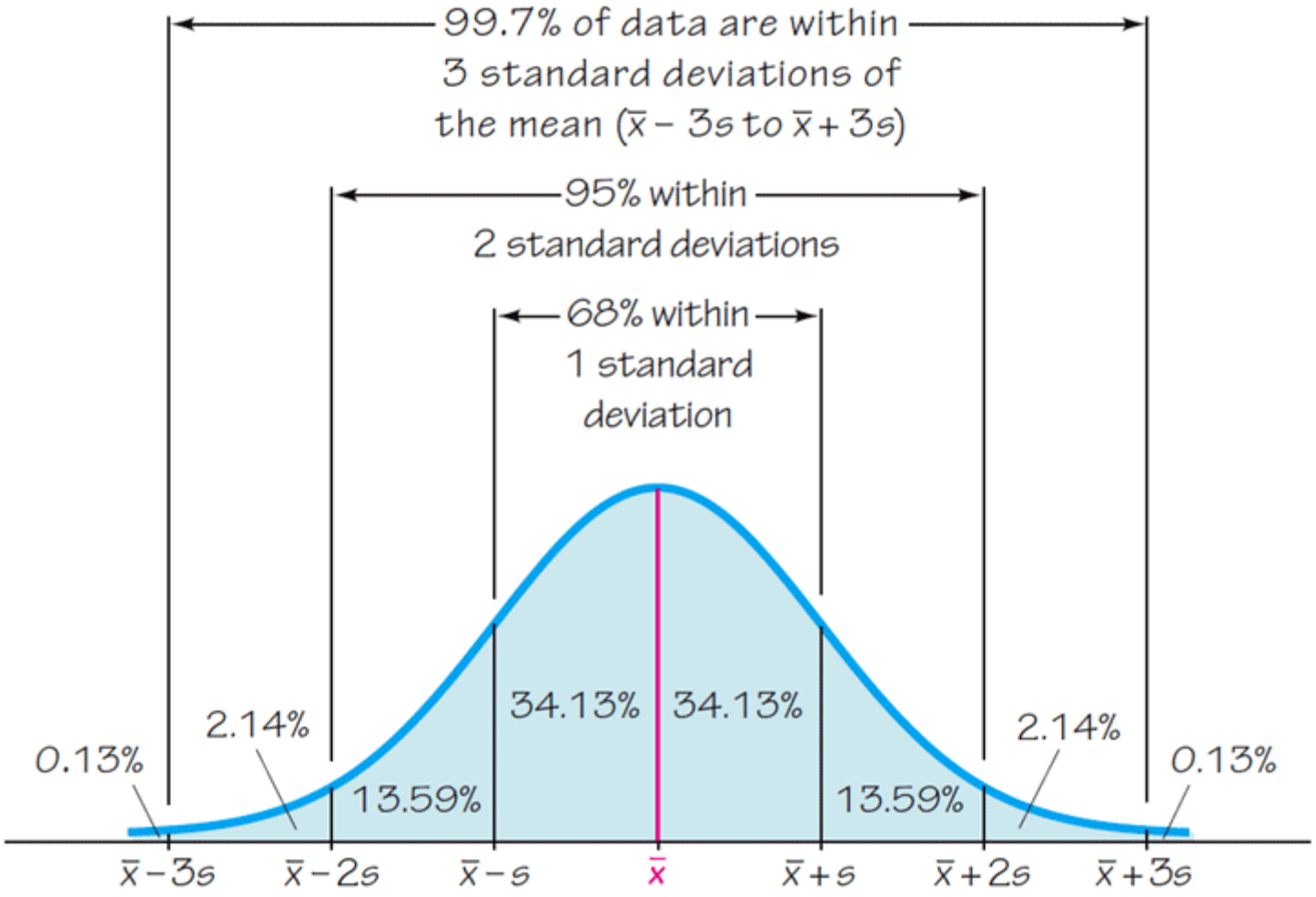
σ : độ lệch chuẩn của quần thể

σ^2: phương sai của quần thể

Quy tắc thực nghiệm (The Empirical Rule)

- Nếu tập dữ liệu có dạng hình chuông, hoặc xấp xỉ dạng hình chuông, thì các tính chất sau đây :
  - Khoảng 68% dữ liệu nằm trong đoạn $[\mu - \sigma, \mu + \sigma]$
  - Khoảng 95% dữ liệu nằm trong đoạn $[\mu - 2\sigma, \mu + 2\sigma]$
  - Khoảng 99.7% dữ liệu nằm trong đoạn $[\mu - 4\sigma, \mu + 3\sigma]$

Quy tắc thực nghiệm (The Empirical Rule)



Định lý Chebyshev

- Với một tổng thể bất kỳ có trung bình μ và độ lệch chuẩn σ , phần trăm các giá trị quan sát nằm trong khoảng $[\mu - k\sigma; \mu + k\sigma]$ bằng ít nhất $(1 - 1/k^2)$
- Ví dụ:

ít nhất	*nằm trong*
$(1 - 1/1^2) = 0\%$	$(\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	$(\mu \pm 3\sigma)$

Hệ số biến thiên (Coefficient of Variation)

- Hệ số biến thiên (Coefficient of Variation): được sử dụng để so sánh sự biến thiên của hai hay nhiều tập dữ liệu và mô tả mối liên hệ giữa độ lệch chuẩn và trung bình.
- Đơn vị tính: %
- Công thức tính:

Sample

$$cv = \frac{s}{\overline{x}} \cdot 100\%$$

Population

$$cv = \frac{\sigma}{\mu} \cdot 100\%$$

Ví dụ

- *Dữ liệu A có trung bình* $\bar{x}_A = 50$, *độ lệch chuẩn* $s_A = 5$

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

- *Dữ liệu B có trung bình* $\bar{x}_B = 100$, *độ lệch chuẩn* $s_B = 5$

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

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

NỘI DUNG

- Độ đo xu hướng tập trung (central tendency)
- Độ đo sự phân tán (variability)
- **Độ đo vị trí tương đối & Đồ thị hộp (boxplots)**

- Trong phần này, chúng ta sẽ khảo sát một số độ đo về vị trí tương đối của các giá trị dữ liệu với nhau trong một tập dữ liệu.
- Chúng ta sẽ xem xét các khái niệm: trị thống kê z, phân vị, tứ phân vị, và đồ thị hộp-râu

Trị thống kê z (z score)

- Trị thống kê z (z score): là sự khác biệt của giá trị dữ liệu với giá trị trung bình của tập dữ liệu được đo bằng số lần của độ lệch chuẩn.

Sample

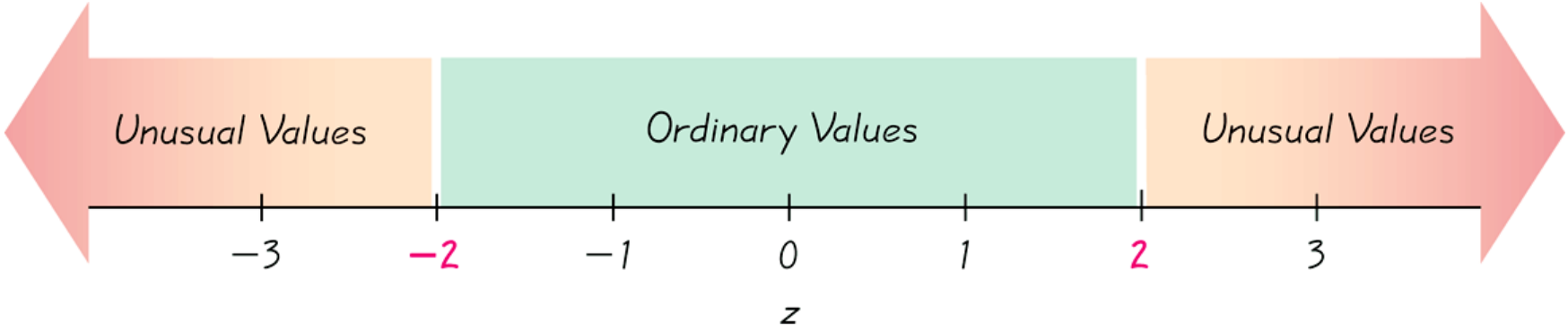
$$z = \frac{x - \bar{x}}{s}$$

Population

$$z = \frac{x - \mu}{\sigma}$$

Giá trị z thường được làm tròn với sai số là 0.01

Trị thống kê z (z score)



Các giá trị thông thường (ordinary):

$$2 \leq z \text{ score} \leq 2$$

Các giá trị bất thường (unusual):

$$z \text{ score} < -2 \text{ hoặc } z \text{ score} > 2$$

Ví dụ

- Một người trưởng thành có nhịp tim trung bình là 67.3 nhịp một phút, và độ lệch chuẩn là 10.3. Ông An có nhịp tim là 48 nhịp một phút. Hỏi nhịp tim ông An như vậy có bình thường hay không?

Ví dụ

- Một người trưởng thành có nhịp tim trung bình là 67.3 nhịp một phút, và độ lệch chuẩn là 10.3. Ông An có nhịp tim là 48 nhịp một phút. Hỏi nhịp tim ông An như vậy có bình thường hay không?

$$z = \frac{x - \bar{x}}{s} = \frac{48 - 67.3}{10.3} = -1.87$$

- Trả lời: Vì $-2 \leq z \leq 2$, nên có thể kết luận ông An có nhịp tim bình thường.

Phân vị (percentiles)

- Phân vị (percentiles): là độ đo về vị trí của dữ liệu. Có tất cả 99 phân vị, được ký hiệu: $P_1, P_2, \ldots, P_{99}$, chia tập dữ liệu thành 100 nhóm, mỗi nhóm chứa khoảng 1% dữ liệu.
- Công thức:

$$Phân\ vị\ của\ giá\ trị\ x = \frac{số\ giá\ trị\ nhỏ\ hơn\ x}{tổng\ số\ giá\ trị} * 100$$

Ví dụ

- Cho bảng số liệu dưới đây, hãy tìm phân vị của giá trị 23.

19	19	20	20	20	20	22	22	22	22
23	23	23	23	23	23	23	24	24	24
24	24	25	25	25	25	25	25	25	26
26	26	26	26	26	27	27	28	28	30

Ví dụ

- Cho bảng số liệu dưới đây, hãy tìm phân vị của giá trị 23.

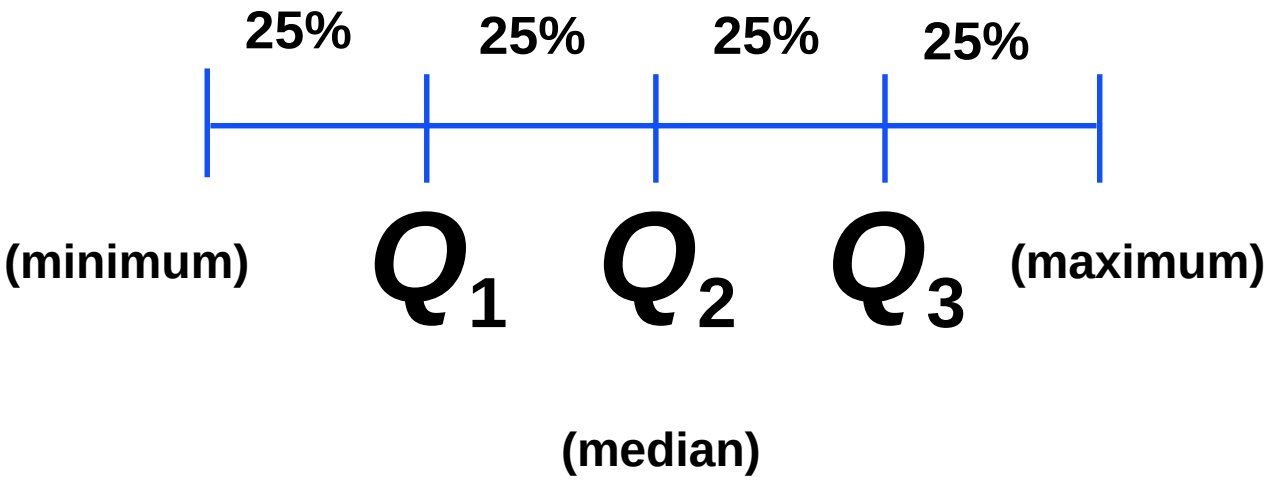
19	19	20	20	20	20	22	22	22	22
23	23	23	23	23	23	23	24	24	24
24	24	25	25	25	25	25	25	25	26
26	26	26	26	26	27	27	28	28	30

- *Giải:* có 10 giá trị nhỏ hơn 23, nên phân vị của giá trị 23 là: (10/40) * 100 = 25
- Hay nói cách khác giá trị 23 nằm ở phân vị thứ 25.

Tứ phân vị (quartiles)

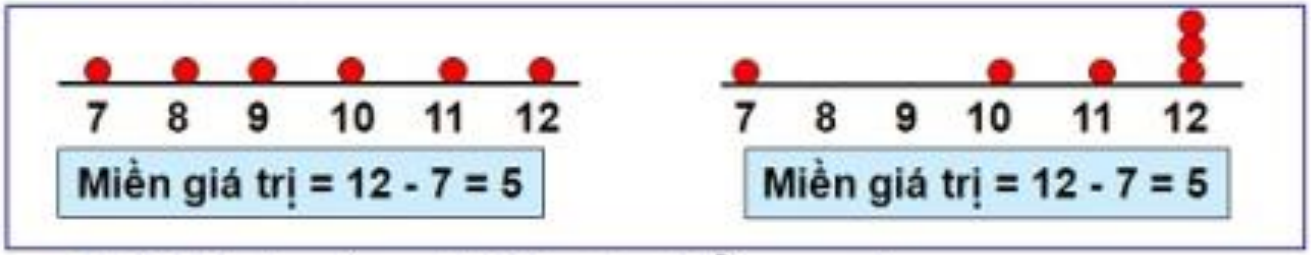
- Tứ phân vị (quartiles): ký hiệu là Q_1, Q_2, Q_3 chia tập dữ liệu đã được sắp xếp thành 4 nhóm, mỗi nhóm chứa khoảng 25% dữ liệu
  - Q_1 (phần tư thứ nhất): chia tập dữ liệu thành 2 nhóm 25% và 75%
  - Q_2 (phần tư thứ hai): chia tập dữ liệu thành 2 nhóm 50% và 50%. Q_2 chính là trung vị (median)
  - Q_3 (phần tư thứ ba): chia tập dữ liệu thành 2 nhóm 75% và 25%

Tứ phân vị (quartiles)

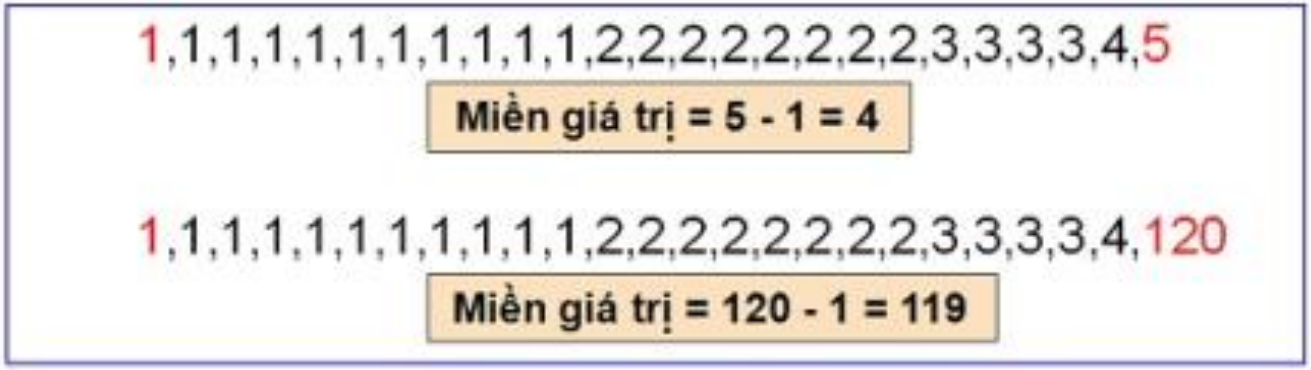


Nhược điểm của tứ phân vị

- Bỏ qua phân bố của dữ liệu



- Bị ảnh hưởng bởi các điểm oulier

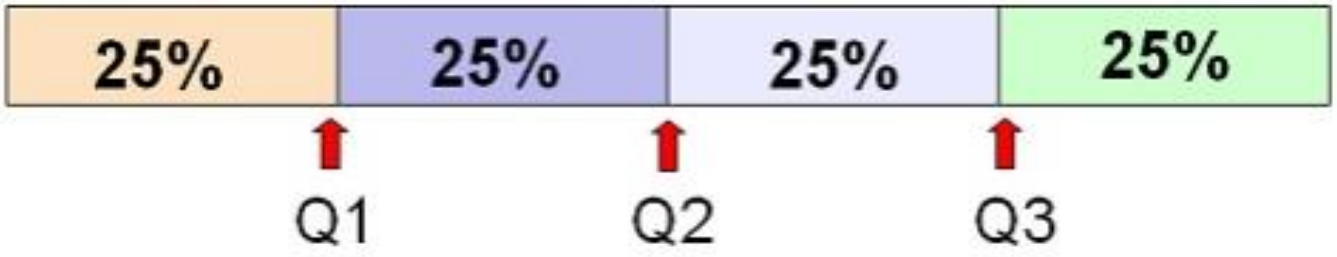


Miền phân vị (Interquartile Range – IQR)

- Miền phân vị (IQR): thường được sử dụng để loại bỏ các giá trị ngoại lệ (outliers)
- Công thức:

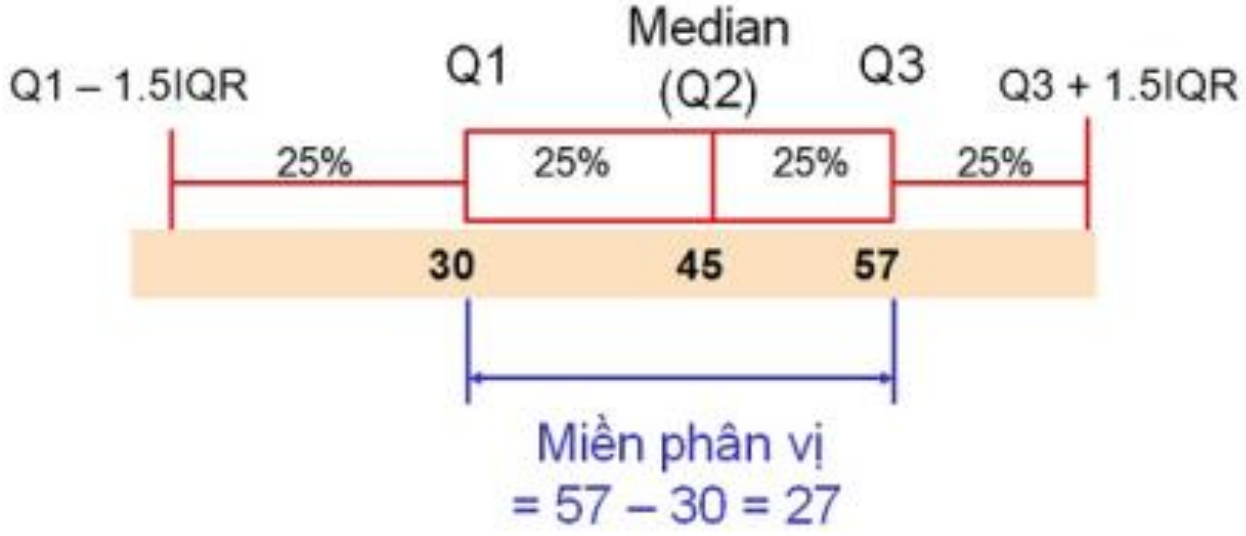
$$IQR = Q_3 - Q_1$$

- Để loại bỏ các ngoại lệ (outliers), ta bỏ các dữ liệu có giá trị nhỏ hơn Q_1 và lớn hơn Q_3



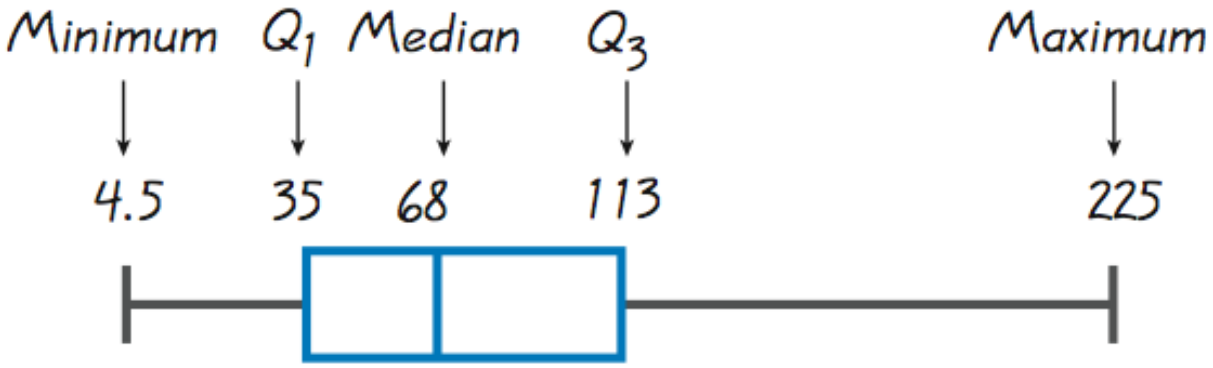
Đồ thị hộp (box-plot)

- Đồ thị hộp hoặc đồ thị hộp râu (box-plot or box-and-whisker-plot) là đồ thị của dữ liệu bao gồm một đường thẳng (râu) đi qua giá trị nhỏ nhất đến giá trị lớn nhất và một hộp được vẽ dựa vào ba giá trị Q_1, Q_2, Q_3

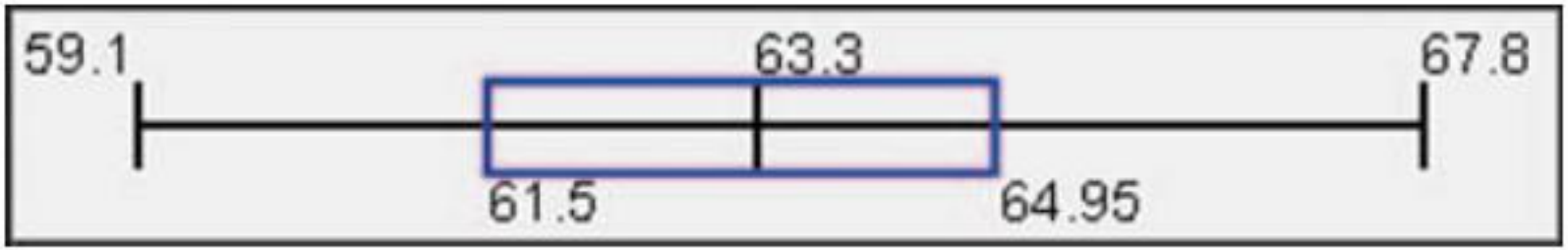


Đồ thị hộp (box-plot) – Cách vẽ

1. Tìm 5 giá trị: min, max, Q_1, Q_2, Q_3
2. Dựng một hình chữ nhật từ Q_1và Q_3 .Trong hình chữ nhật vẽ đường thẳng đi qua giá trị Q_2
3. Vẽ một đường thẳng từ hình chữ nhật (hộp) đến giá trị min và giá trị max.



Đồ thị hộp (box-plot) của phân phối chuẩn



Normal Distribution:
Heights from a Simple Random Sample of Women

Đồ thị hộp (box-plot) của phân phối bị lệch

Skewed Distribution:
Salaries (in thousands of dollars) of NCAA Football Coaches

Ngoại lệ (outlier)

- Ngoại lệ (outlier) là dữ liệu nằm rất xa so với phần lớn dữ liệu còn lại trong tập dữ liệu.
- Ảnh hưởng của ngoại lệ:
  - Ngoại lệ có thể ảnh hưởng lớn giá trị trung bình và độ lệch chuẩn của tập dữ liệu.
  - Ngoại lệ có thể làm sai lệch cái nhìn về phân phối của tập dữ liệu

Ngoại lệ (outlier)

- Một giá trị dữ liệu được gọi là ngoại lệ trong đồ thị hộp hiệu chỉnh nếu thỏa mãn:
  - Lớn hơn Q_3 một khoảng 1.5 x IQR

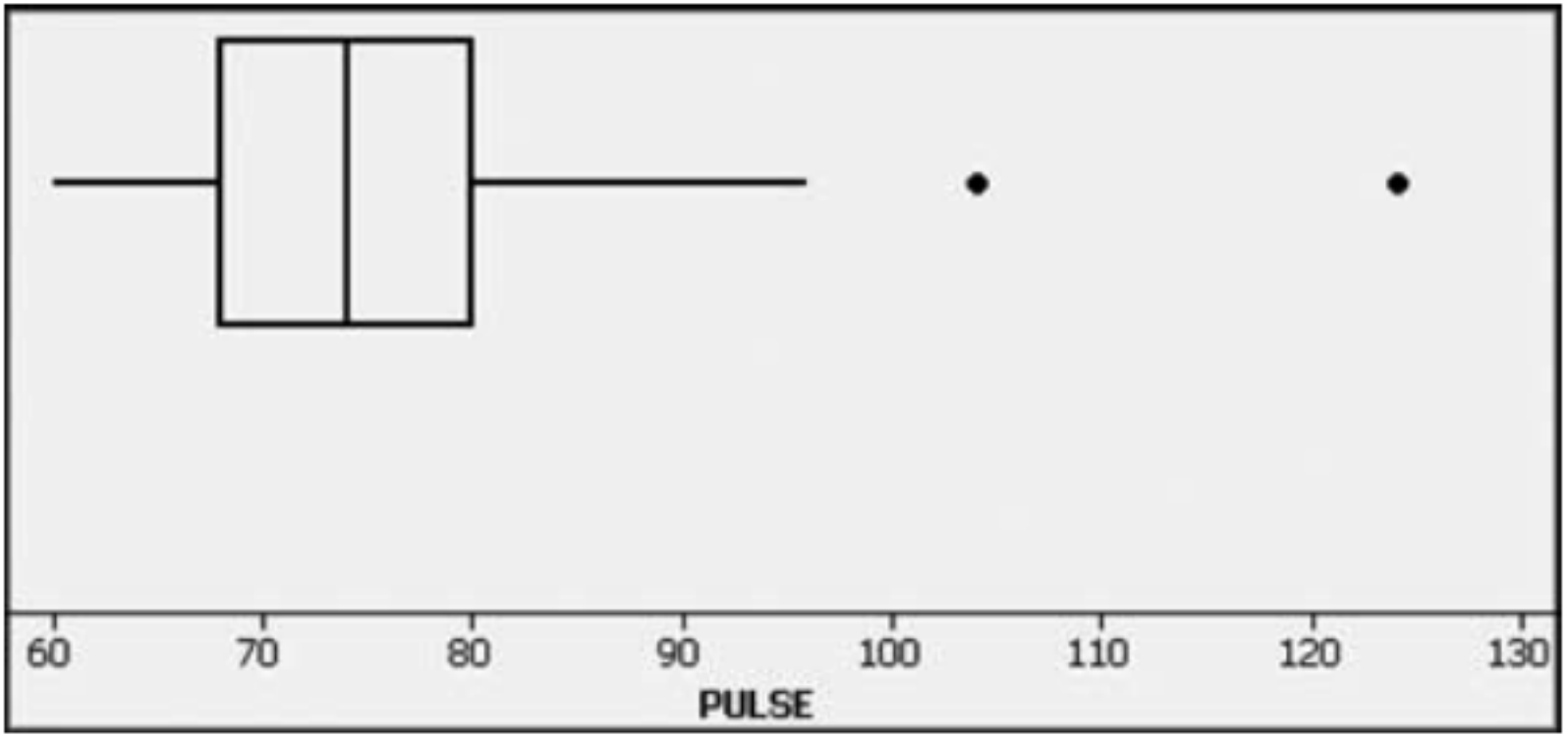
  Hoặc

  - Nhỏ hơn Q_1 một khoảng 1.5 x IQR

Đồ thị hộp hiệu chỉnh – Cách xây dựng

- Một đồ thị hộp hiệu chỉnh được xây dựng giống như đồ thị hộp ban đầu nhưng thêm các yếu tố sau:
  - Sử dụng dấu hoa thị (*) để mô tả các ngoại lệ
  - Một đường thẳng nối từ giá trị nhỏ nhất không phải ngoại lệ đến giá trị lớn nhất không phải ngoại lệ

Đồ thị hộp hiệu chỉnh – Cách xây dựng



THANK YOU