PHƯƠNG PHÁP NGHIÊN CỨU TRONG TÂM LÝ HỌC

BÀI 9

MÔ TẢ KẾT QUẢ

Tien.phamthithuy2386@hoasen.edu.vn

T11/2019

Tổng quan

A. Nhắc lại: Loại dữ kiện – *types of data*
B. Tóm tắt và biểu diễn dữ kiện bằng biểu đồ - *summarizing data in graphs*
C. Phân bố tần số – *frequency distribution*
D. Đo lường xu hướng trung tâm – *measures of central tendency*
E. Đo lường độ phân tán – *measures of dispersion (or variability)*

Mô tả vs. suy luận

- Thống kê mô tả
  - Mô tả mẫu nghiên cứu
  - Tóm tắt các xu hướng trong mẫu:
    - Xu hướng trung bình – central tendency
    - Phân tán – dispersion

- Thống kê suy luận
  - Cho phép suy luận từ các đặc điểm và xu hướng của mẫu sang cho quần thể.

Suy luận thống kê: Từ mẫu đến quần thể

- Là tập hợp có N phần tử, là mối quan tâm của nhà nghiên cứu
- Có nhiều biến ta không quan sát
-Các thông số (parameter) như:
  - μ – (population) mean
  - σ –standard error
  - β – regression coefficients

Quần thể - Population

-Là tập hợp n phần tử được chọn ngẫu nhiên từ quần thể
-Các biến giải thích và biến kết quả quan sát được
-Các chỉ số thống kê (statistic) như:
  - M - sample mean
  - SD - standard deviation

Mẫu – Sample

Inference – Suy luận

A. Nhắc lại bài cũ: Các loại dữ kiện

Discrete/
Continuous

Variable

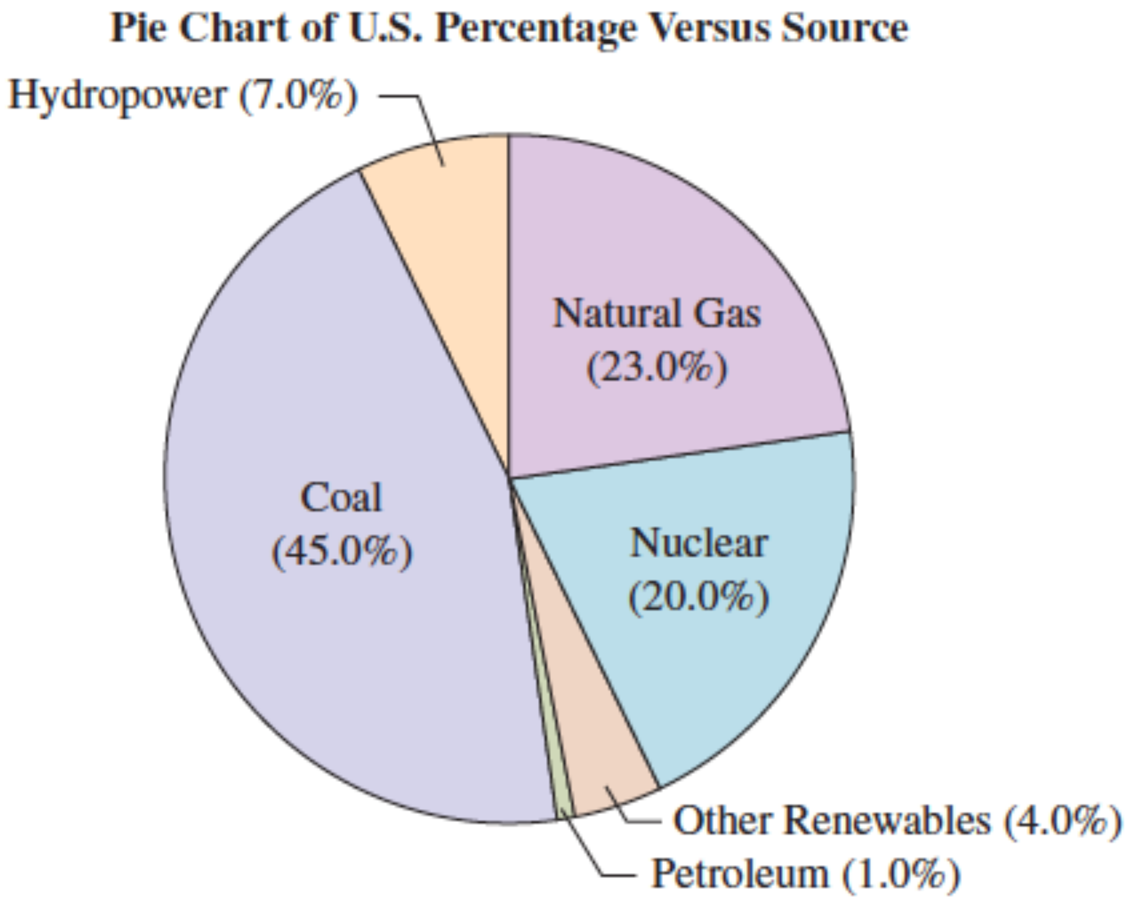
Categorical Quantitative

Level	Nominal	Ordinal	Interval	Ratio
Defining Feature	Distinct Categories	Ordered Categories	Meaningful Distances	Absolute Zero
Operations	$= \neq$	$< >$	$+ -$	$\times \div$

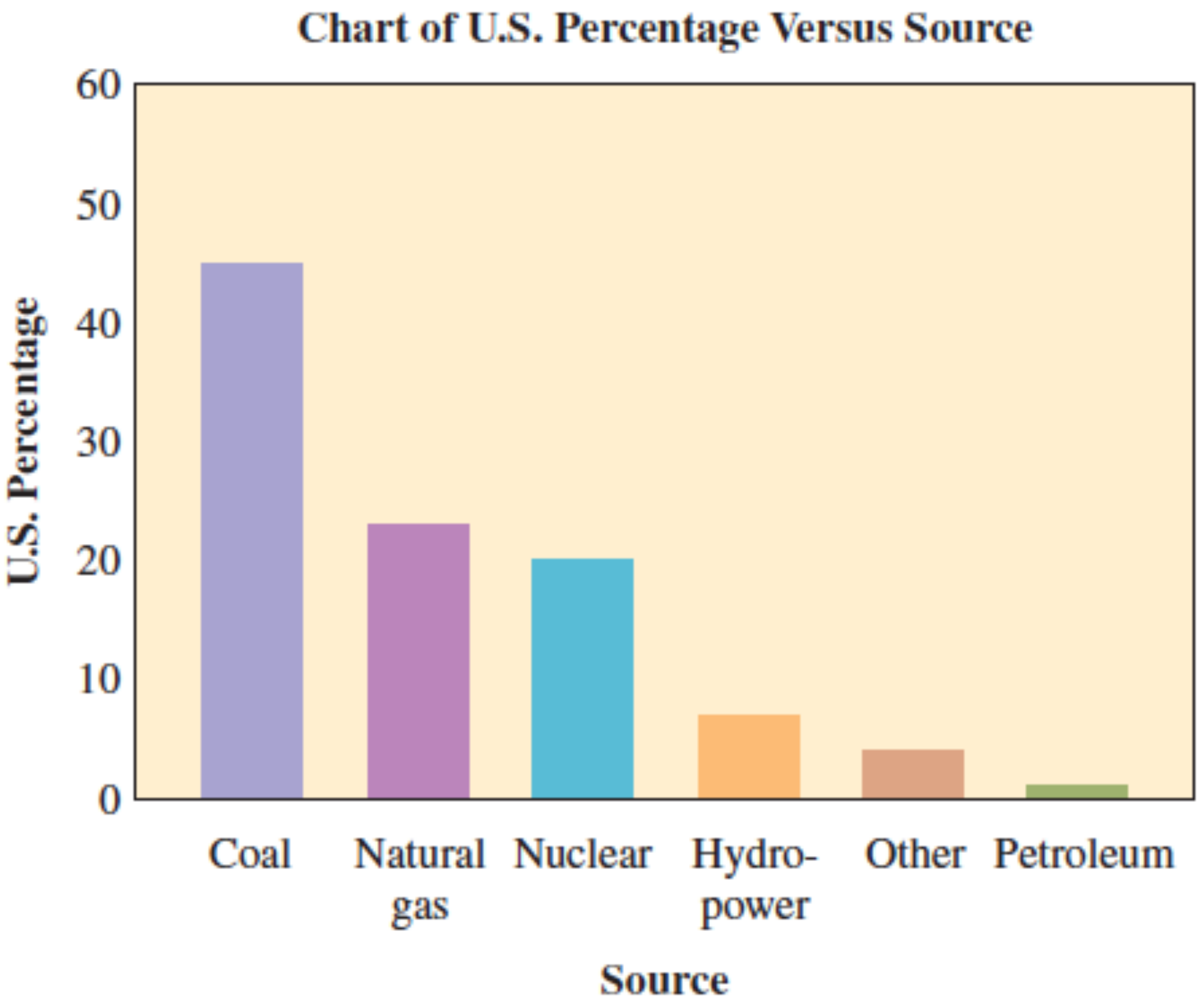
Rời rạc vs. Liên tục

- Biến rời rạc:
  - Chỉ có thể nhận một vài giá trị nhất định
  - Thường là số nguyên
    - Ví dụ: Thang Likert 5 điểm
- Biến liên tục:
  - Biến có thể nhận bất kỳ giá trị nào
  - Có thể đo tới bất kỳ mức độ chính xác nào bạn cần
    - Ví dụ: chiều cao, cân nặng, thời gian

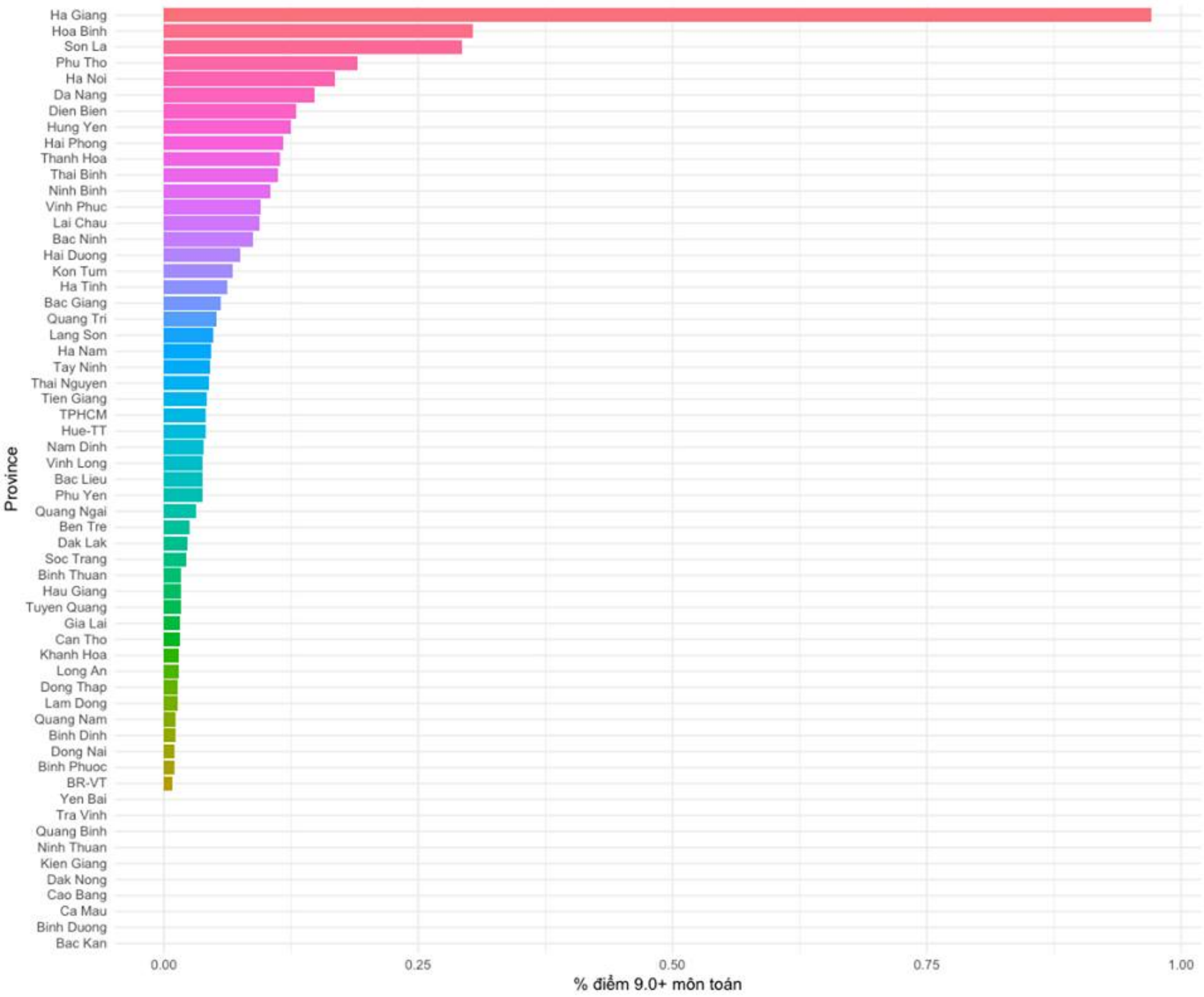
B. Biểu đồ: Biểu đồ bánh – Pie chart



Biểu đồ thanh – Bar graph

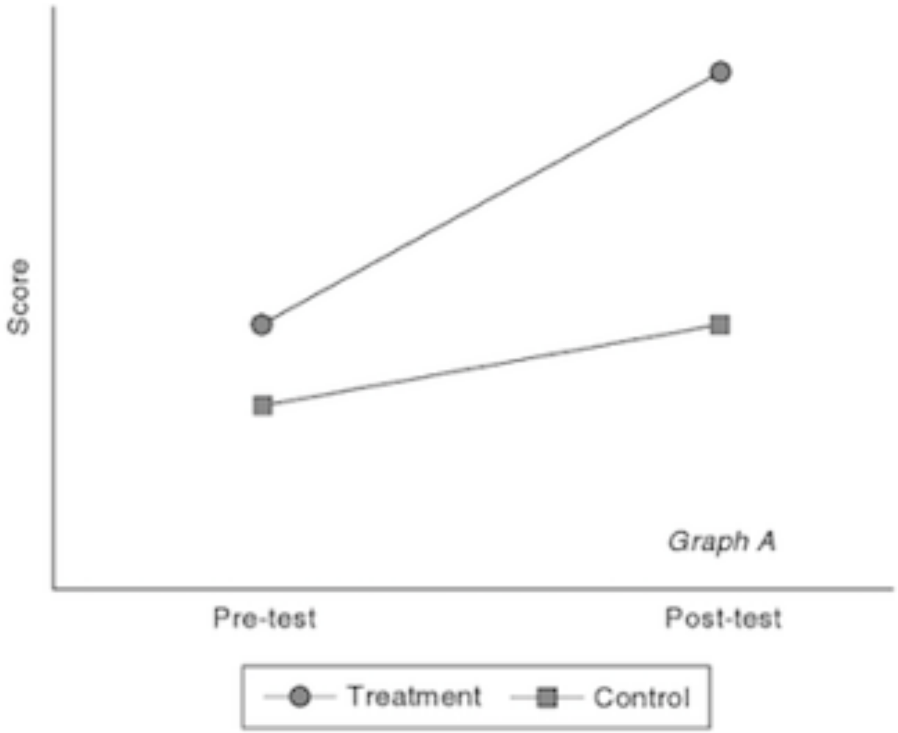


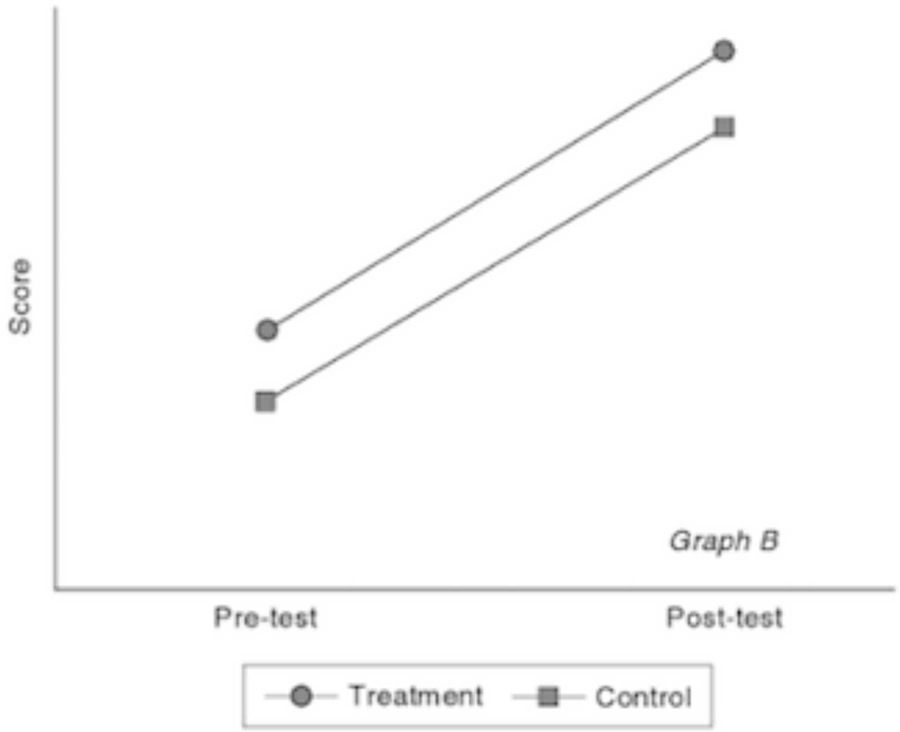
Biểu đồ thanh – bar graph



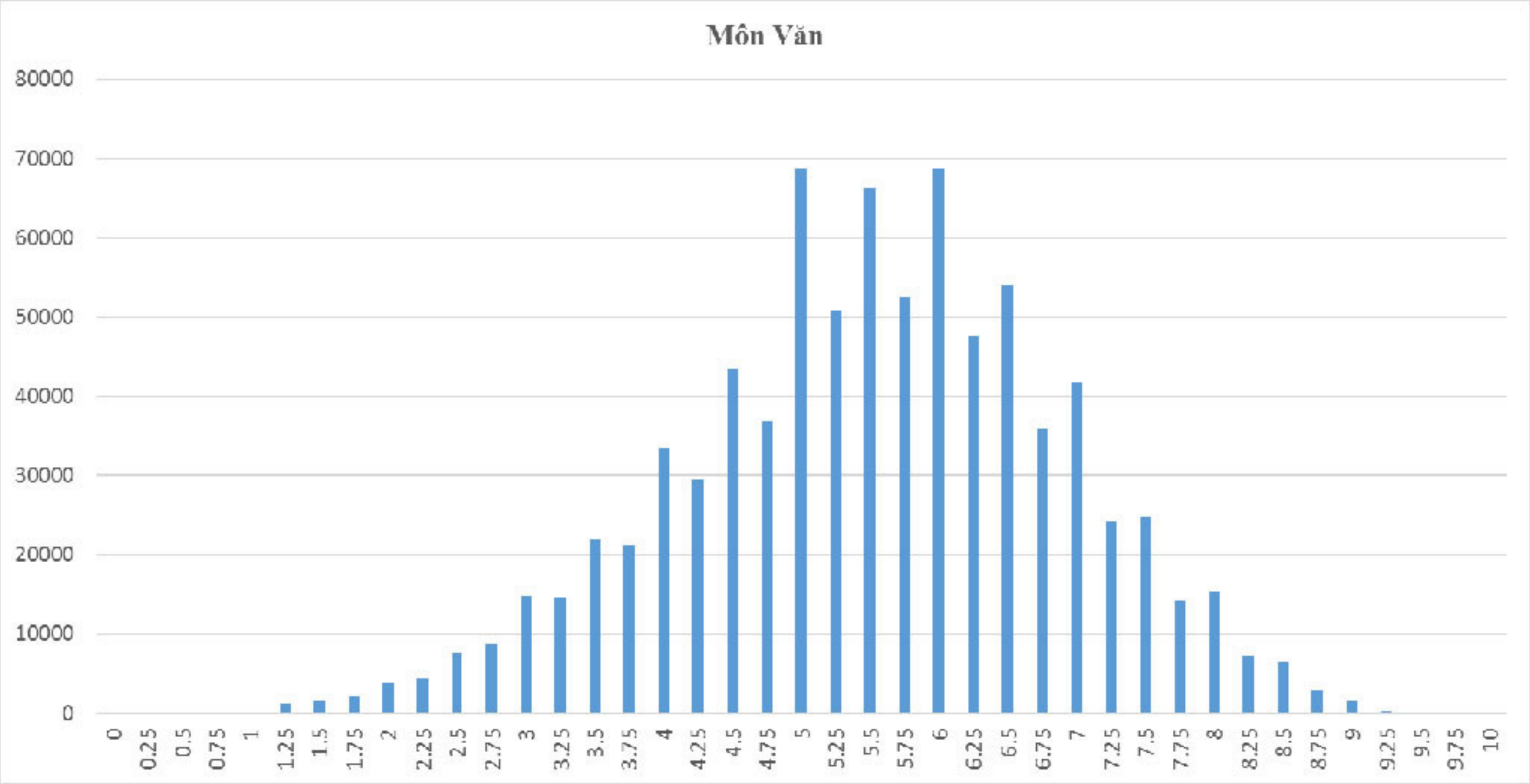
Tỉ trọng điểm môn toán từ 9 trở lên theo tỉnh thành. (Nguồn: FB Tuấn Nguyễn)

Biểu đồ đường – Line graph

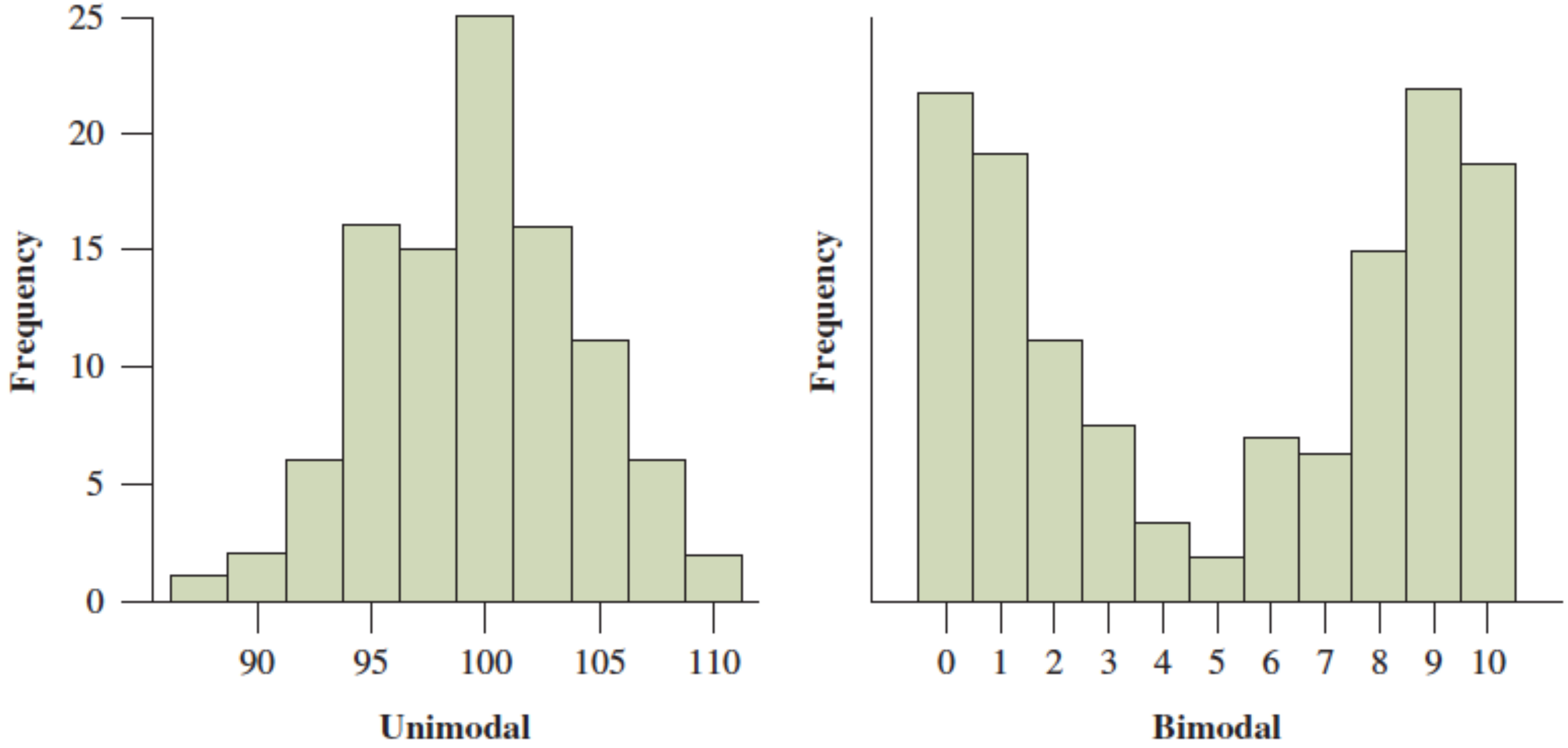


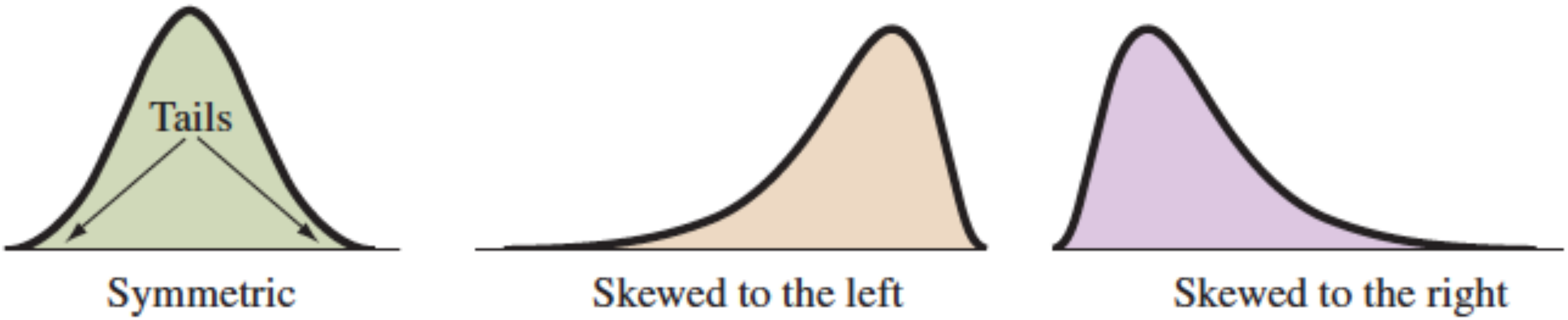


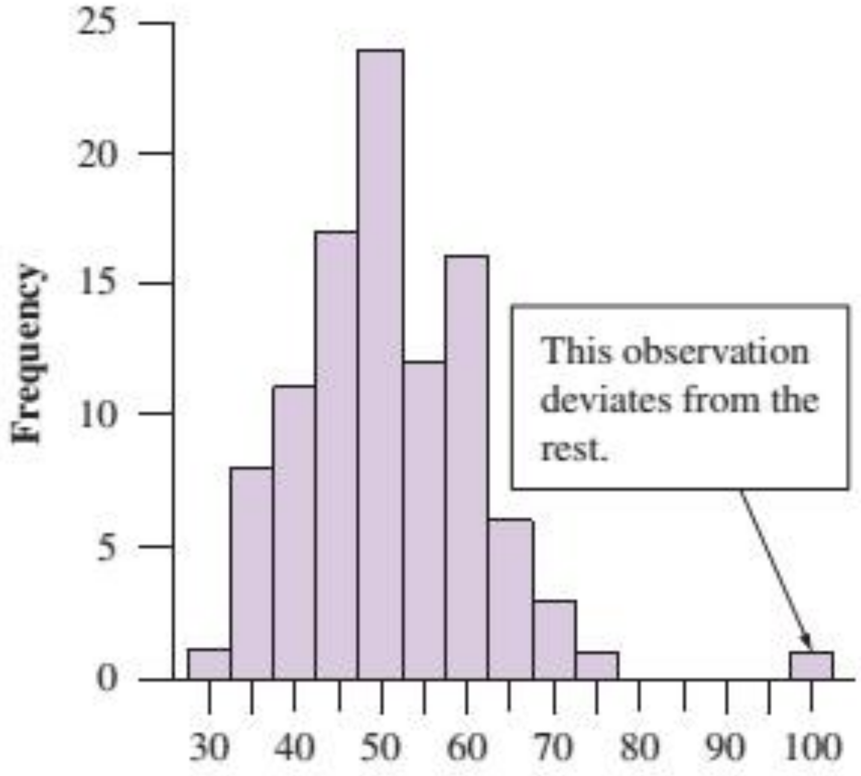
Biểu đồ tần số - Histogram

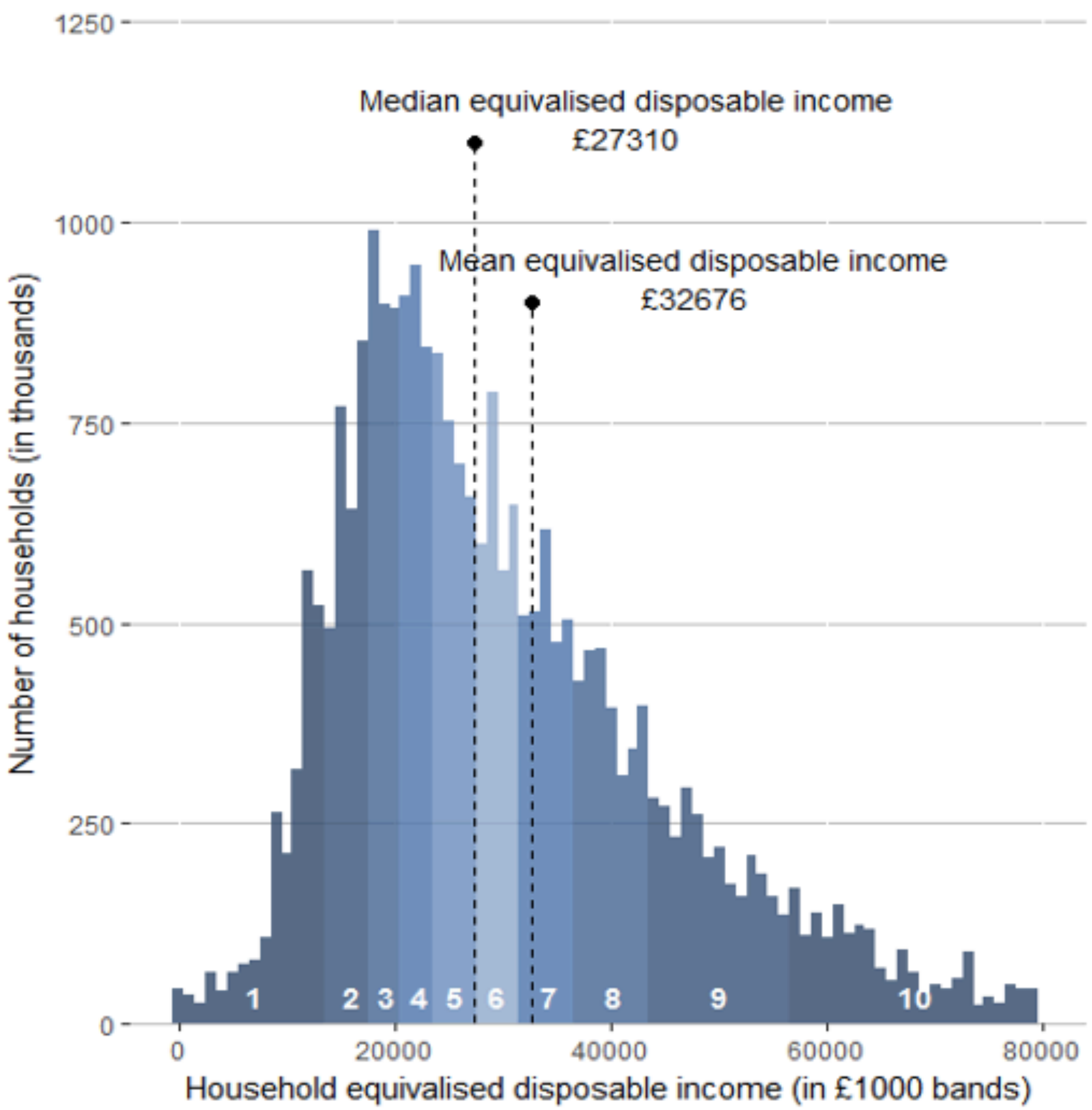


C. Hình dạng của phân bố tần số - Shape of frequency distribution

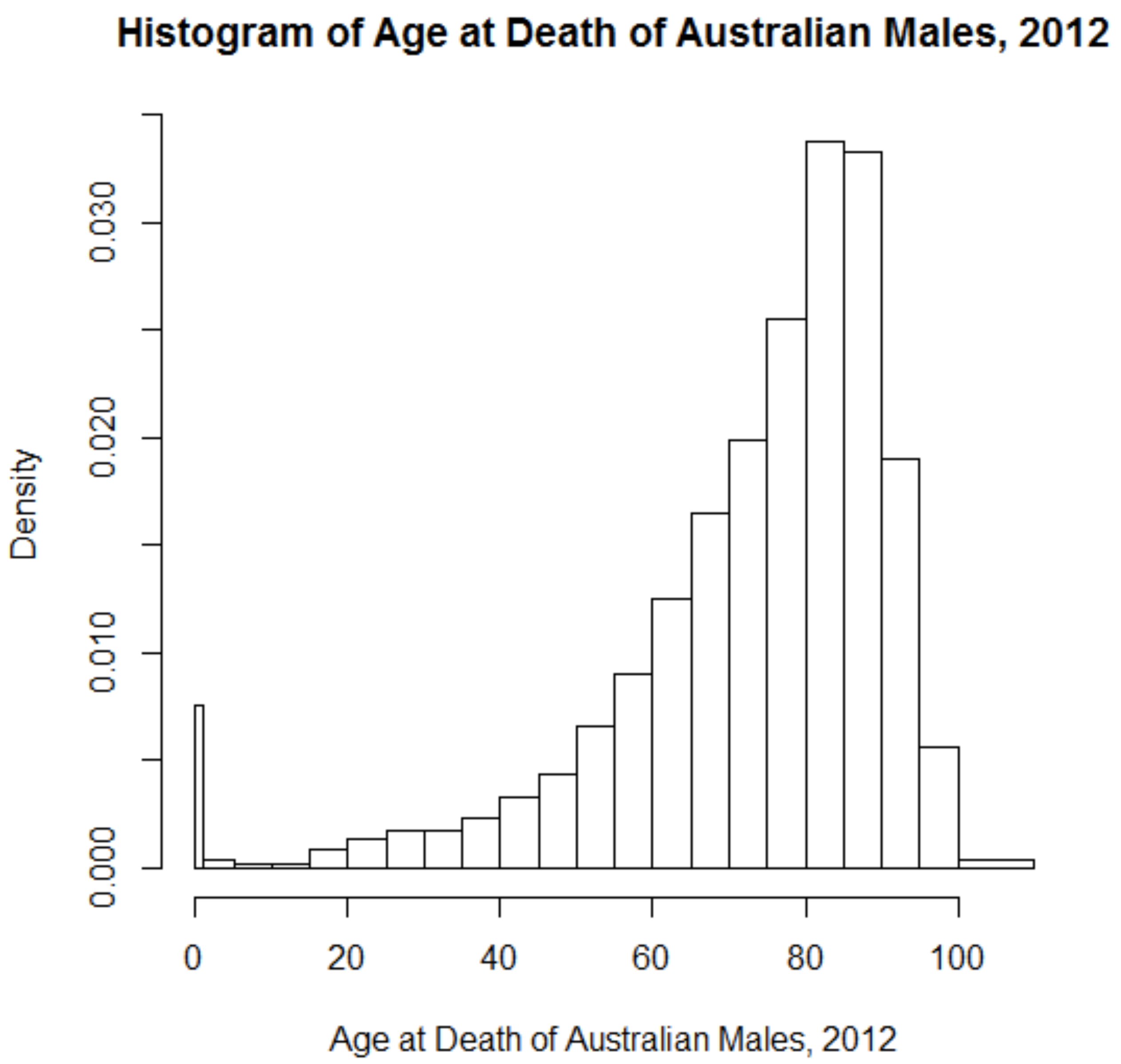



Tails
Symmetric
Skewed to the left
Skewed to the right


Frequency
0
5
10
15
20
25
30
40
50
60
70
80
90
100
This observation deviates from the rest.



Office for National Statistics
Figure: Distribution of UK household disposable income, financial year ending 2017


Histogram of Age at Death of Australian Males, 2012
Density
0.000
0.010
0.020
0.030
0
20
40
60
80
100
Age at Death of Australian Males, 2012

D. Đo lường xu hướng trung bình

- Một cách để trình bày, tóm tắt dữ kiện

43 46 32 10 33 28 60 58 66 8470

22 90 16 42 45 44 62 65 45 52 68

22 45 61 32 18 75 33 28 12 34

33 47 56 82 37 23 26 48 91 60

Đo lường xu hướng trung bình

Mode = giá trị có tần số cao nhất

Trung vị (median) = giá trị ở giữa

Số trung bình (mean) = trung bình cộng

Mode

Giá trị xuất hiện thường xuyên nhất:

v.d. 4 6 7 7 9 7 3 9

Mode = 7

(+) Dễ tính toán, dễ hiểu

v.d. 2 6 8 1 4 7 3 5 92 92

Mode = 92

(-) Nguy cơ trở nên thiếu tính đại diện

Trung vị - Median

Giá trị giữa của dãy giá trị khi xếp theo thứ tự

1 3 6 8 9 11 12

Trung vị = 8

(+) Ít bị ảnh hưởng bởi giá trị ngoại vi – *outliers* (điểm cực trị - *untypical extreme scores*)

1 3 6 8 9 11 92

Trung vị = 8

(-) Không đại diện cho tất cả các giá trị trong toàn bộ dữ liệu

Khi có 2 trung vị

Trường hợp có hai trung vị (tức là khi tổng số quan sát là số chẵn)

v.d. 3 4 6 8 8 8

Hai điểm giữa là 6 và 8

Ta lấy trung bình cộng của chúng

Trung vị = (6 + 8) / 2 = 7

Giá trị trung bình – Mean

Lấy tổng của tất cả giá trị của các quan sát chia cho tổng số quan sát

v.d. 3 2 2 5 3

Mean = (3 + 2 + 2 + 5 + 3) / 5 = 15/5 = 3

(+) Là giá trị duy nhất đại diện cho tất cả các giá trị quan sát được

v.d. 3 2 2 5 23

Mean = 7

(-) Dễ bị ảnh hưởng bởi giá trị ngoại vi

Công thức tính giá trị trung bình

$$\overline{X} = \frac{\sum X}{N}$$

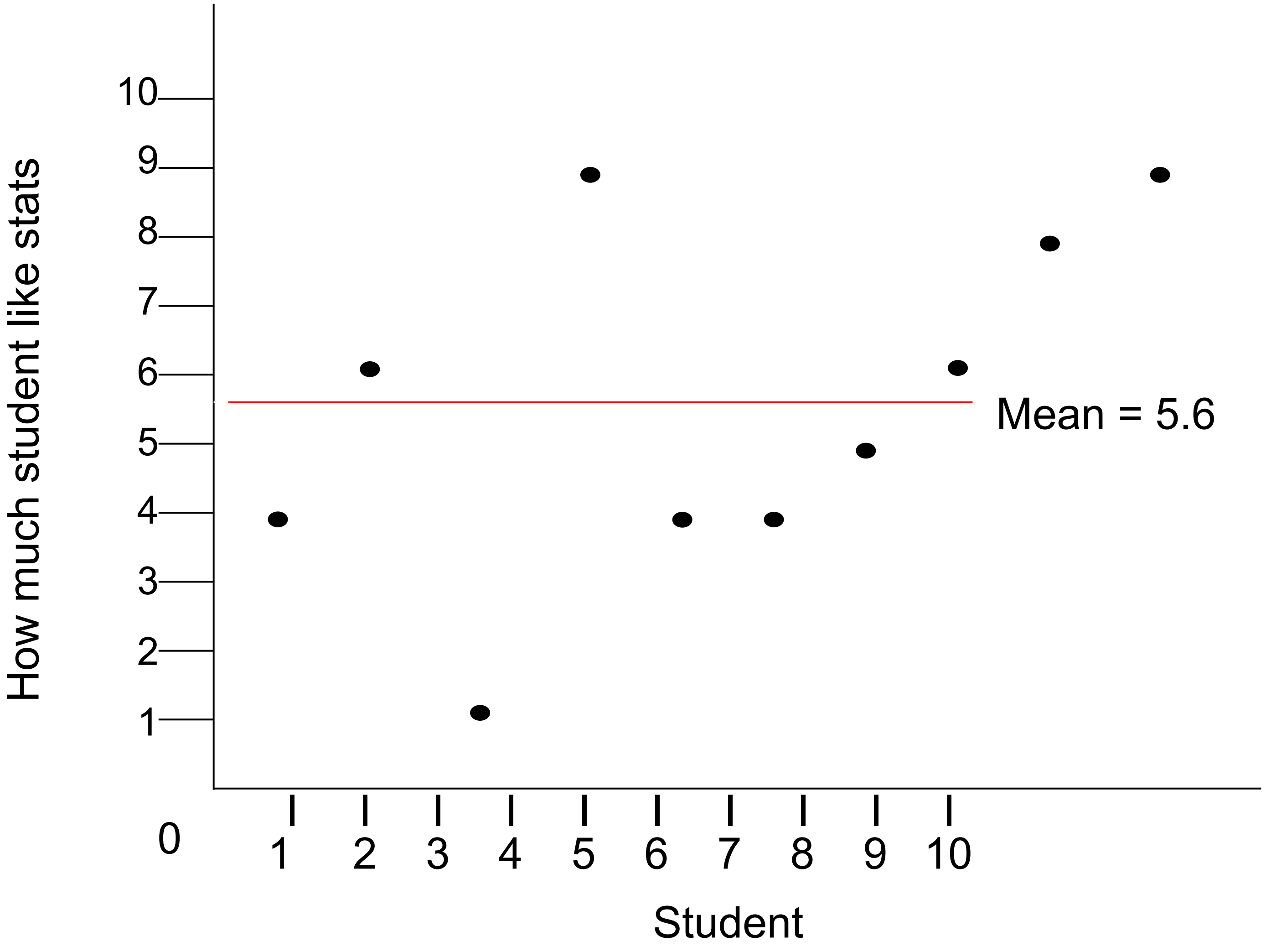
$\overline{X}$ = Mean

$\sum$ = The sum of

X = A score

N = Number of scores

Có thể xem GTTB (mean) là một mô hình dự báo đơn giản



Student	How much like stats
1	4
2	6
3	1
4	9
5	4
6	4
7	5
8	6
9	8
10	9

Mean = 5.6

E. Đo lường xu hướng phân tán

- Khoảng (*Range*)
- Khoảng tứ phân vị (*Interquartile Range*)
- Độ lệch chuẩn (*Standard Deviation*)

Khoảng - Range

Là khoảng khác biệt giữa quan sát có giá trị lớn nhất và quan sát có giá trị nhỏ nhất

22   25   30   42   88   102

Range = Max – Min

102 – 22 = 80

Chỉ phản ánh hai điểm cực trị của toàn bộ dữ liệu

Interquatile Range (IQR)

Các điểm cực trị ảnh hưởng đến khoảng

22 22 23 25 27 30 30 32 32 33 33 106

Range = 106 – 22 = 84. Nhưng hầu hết quan sát rơi vào giá trị từ 22 đến 33

IQR

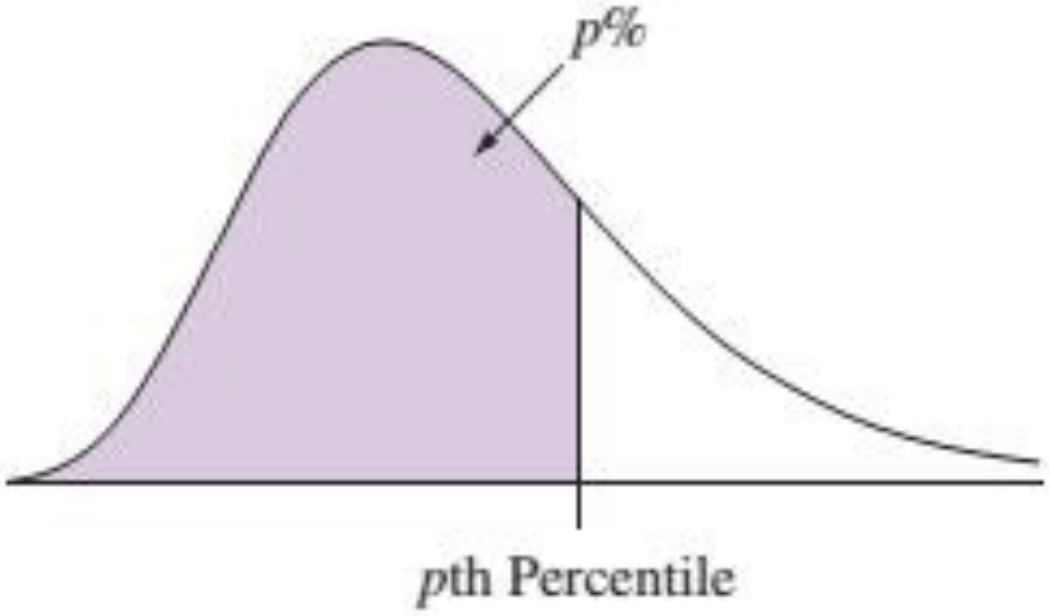
22 22 23 25 27 30 30 32 32 33 33 106

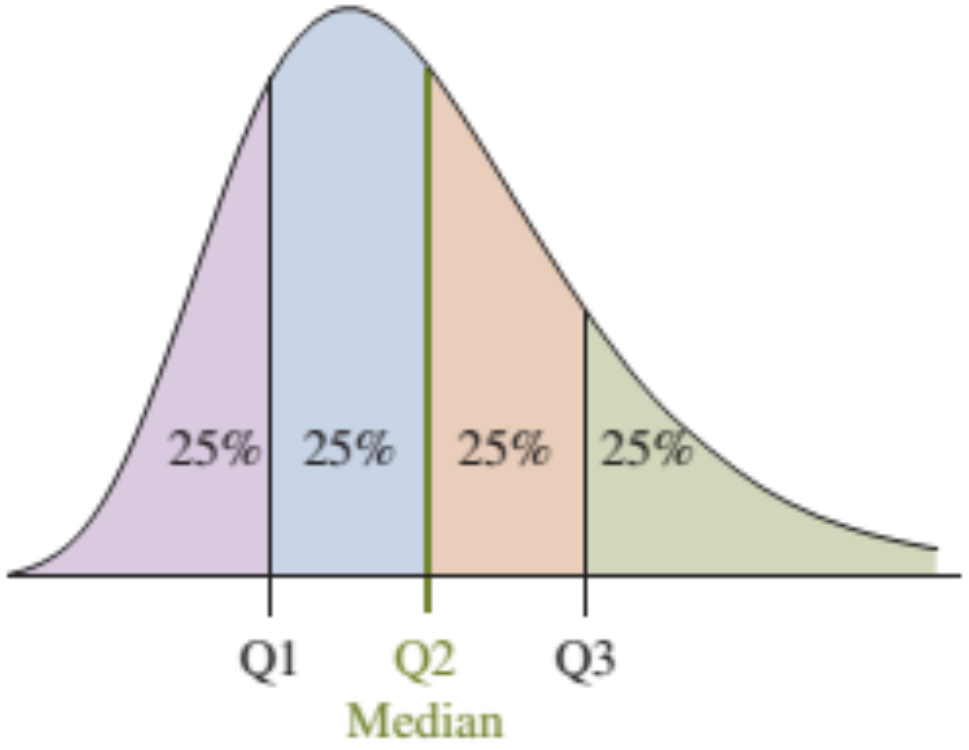
25th percentile

Median

75th percentile

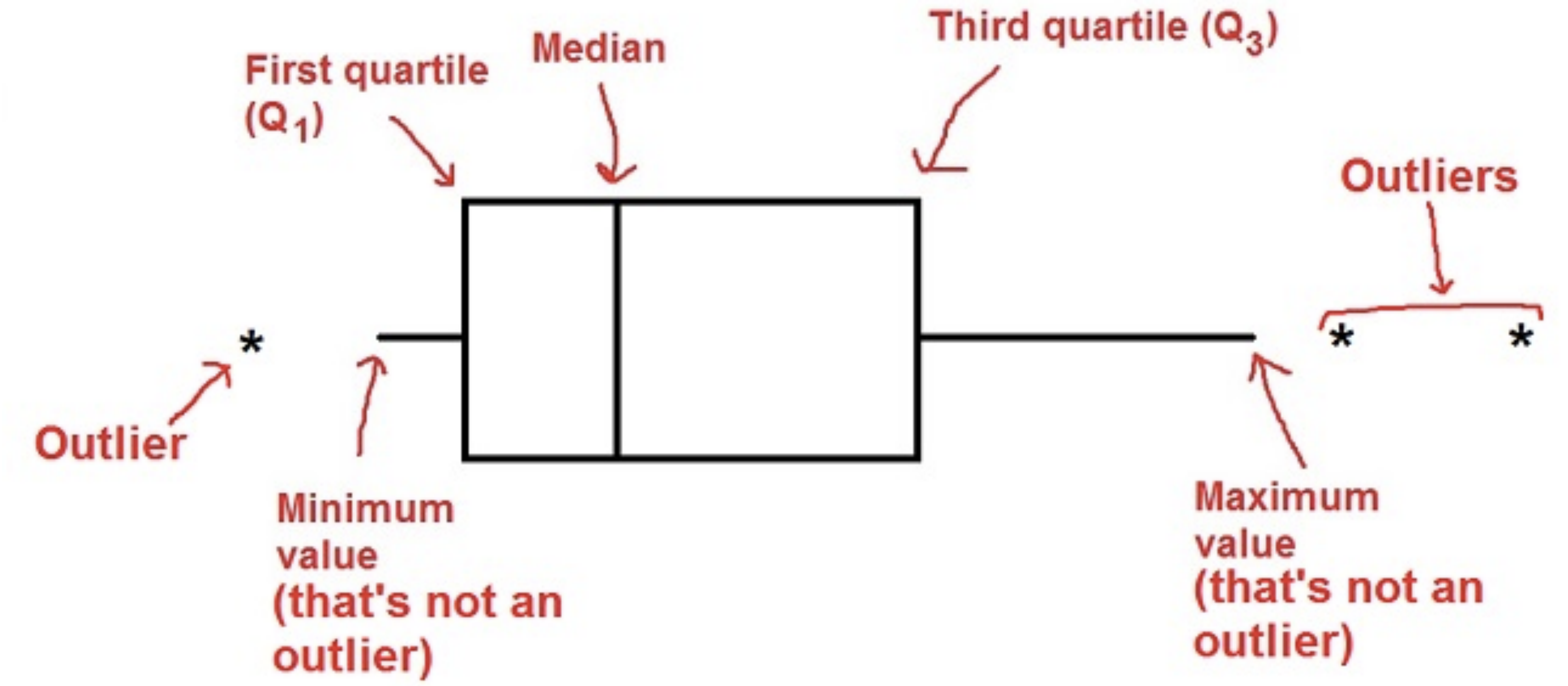
Interquatile Range (IQR)





Khoảng tứ phân vị (IQR): khoảng được ‘tỉa bớt’
IQR = Q3 – Q1

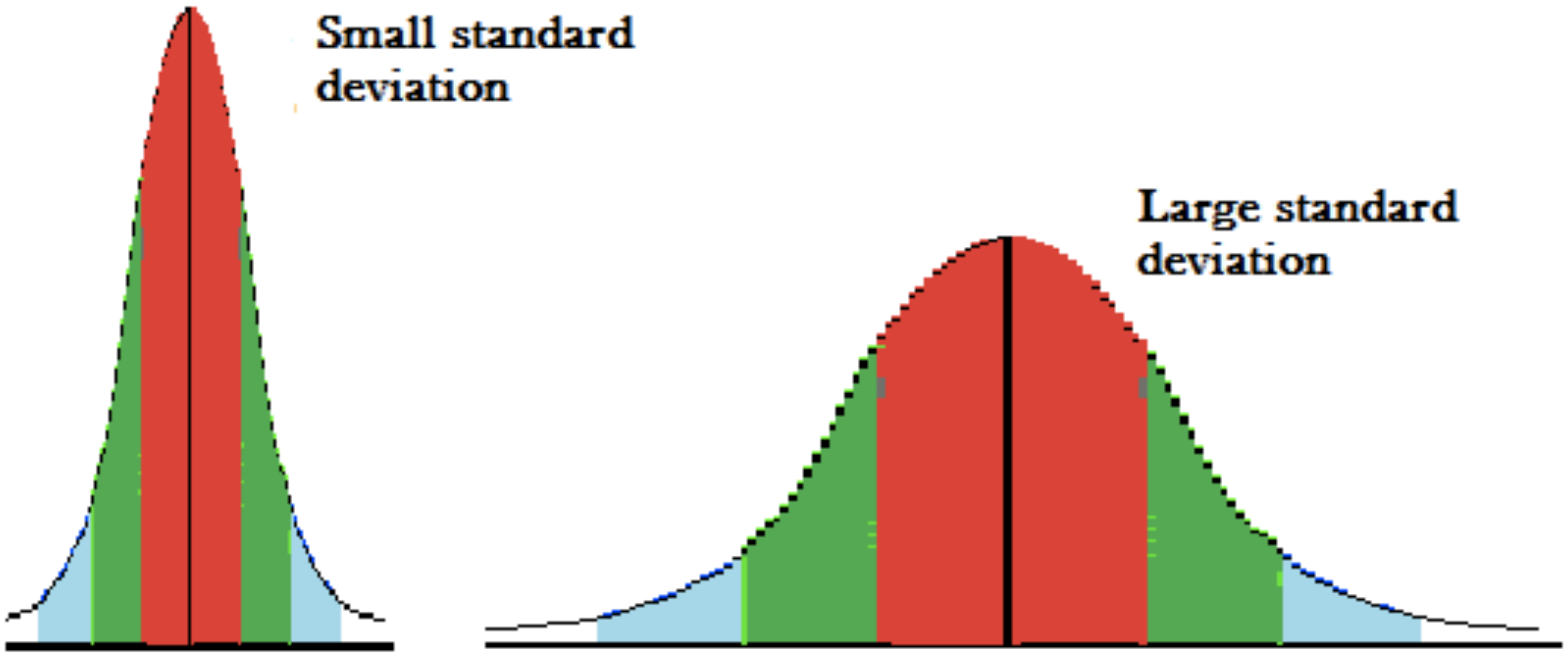
Biểu đồ hộp: Box plot



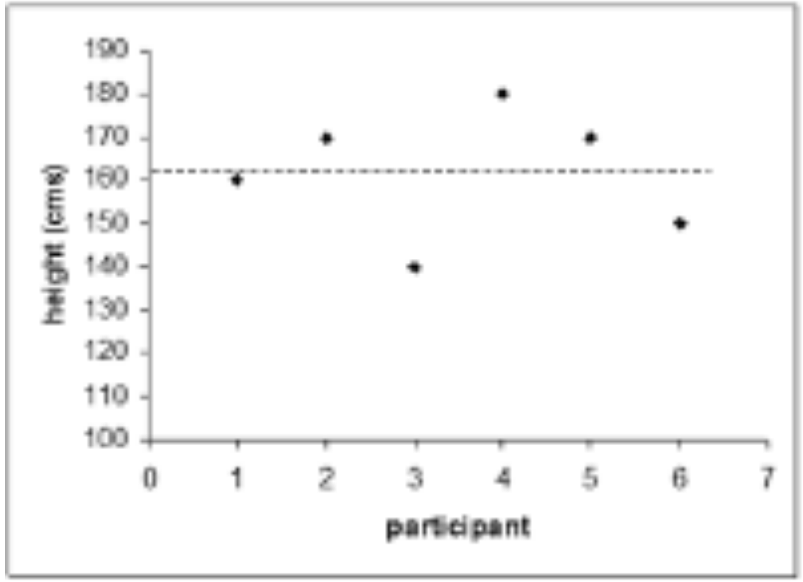
Độ lệch chuẩn – Standard deviation (SD)

- Độ lệch chuẩn thể hiện khoảng lệch của từng quan sát so với giá trị trung bình.
- Nó đại diện cho tất cả các giá trị trong toàn bộ dữ kiện
- Là một đại lượng rất hữu ích để mô tả dữ liệu

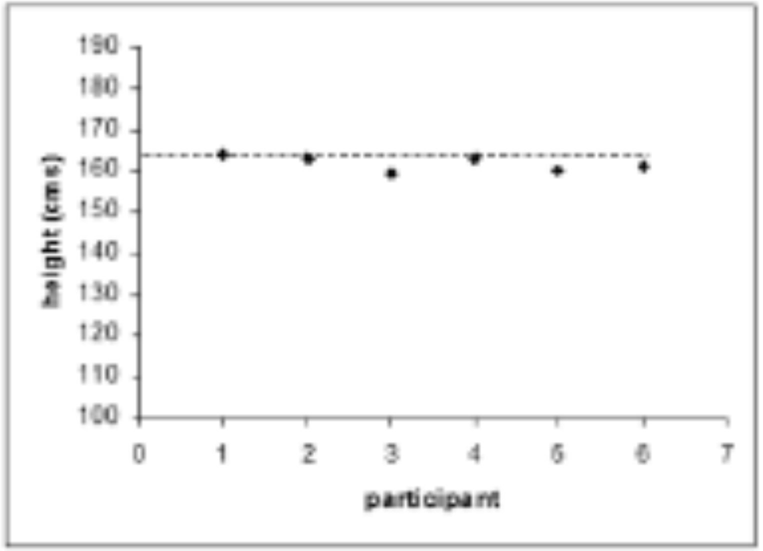
Độ lệch chuẩn – Standard deviation (SD)



2a.



2b.



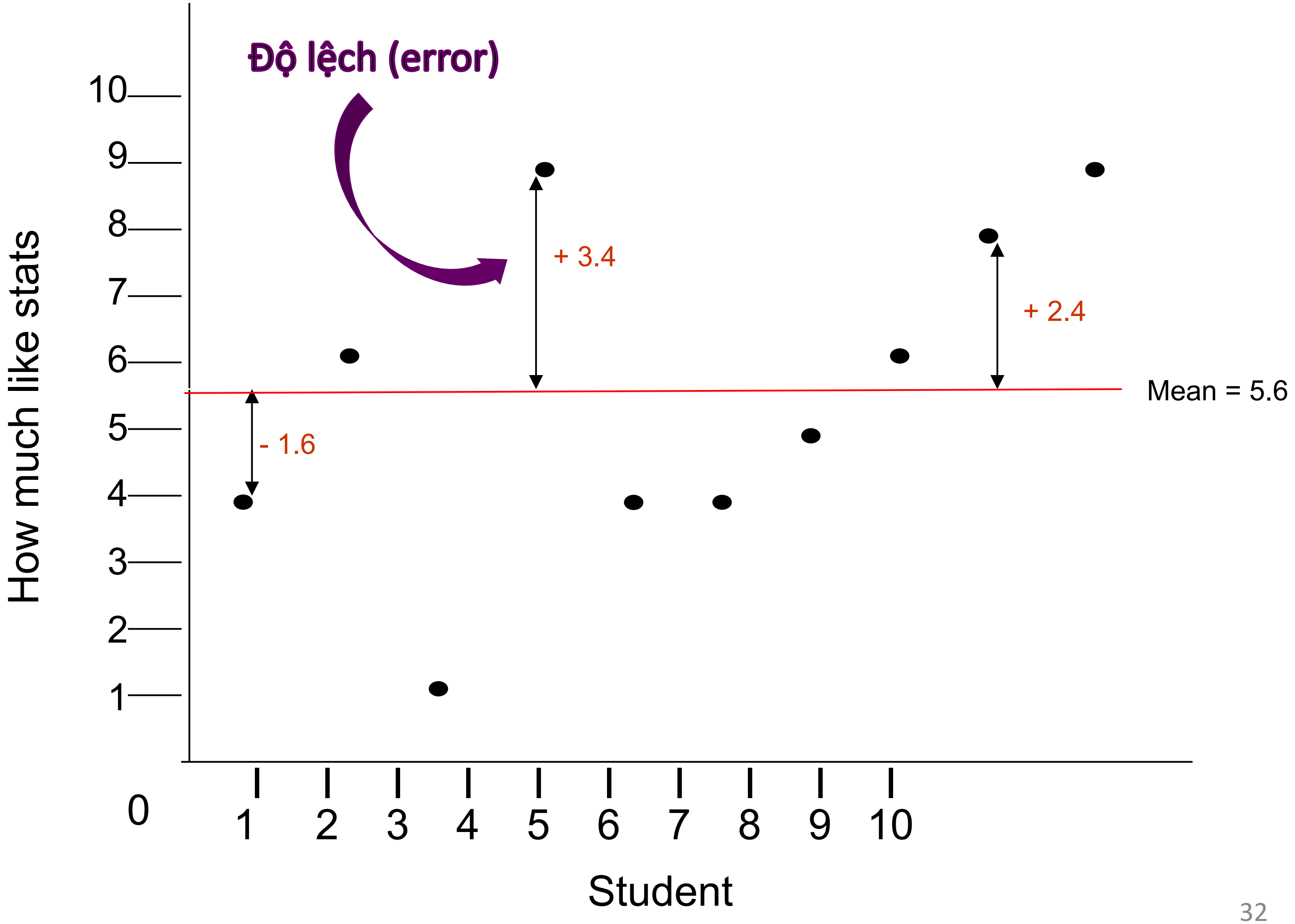
Biểu đồ chiều cao của người tham gia nghiên cứu. Biểu đồ 2a và 2b có cùng giá trị trung bình là 1m60.
Nhưng để ý rằng, tính biến thiên của từng quan sát riêng lẻ của hai biểu đồ khác nhau.

2a) Biến thiên lớn giữa các quan sát riêng lẻ (large variation)
2b) Biến thiên ít giữa các quan sát riêng lẻ (small variation)

... trở lại với mô hình GTTB đơn giản

Student	How much like stats	Differences (error)
1	4	-1.6
2	6	0.4
3	1	- 4.6
4	9	3.4
5	4	-1.6
6	4	-1.6
7	5	-0.6
8	6	0.4
9	8	2.4
10	9	3.4

Mean = 5.6



Cách tính độ lệch chuẩn: Độ lệch (error)

Student	How much like stats	Differences (error)
1	4	-1.6
2	6	0.4
3	1	- 4.6
4	9	3.4
5	4	-1.6
6	4	-1.6
7	5	-0.6
8	6	0.4
9	8	2.4
10	9	3.4
Total	**56**	**0**

Mean = 5.6

Độ lệch của một quan sát so với giá trị trung bình là:

$$x - \bar{x}$$

Mỗi quan sát đều có một độ lệch - *deviation*, nhận giá trị dương hoặc âm tùy theo vị trí của nó nhỏ hơn hay lớn hơn giá trị trung bình. Tổng các độ lệch này luôn bằng 0.

1. Sum of squares

Student	How much like stats	Differences from mean (error)	Squared differences
1	4	-1.6	2.56
2	6	0.4	0.16
3	1	- 4.6	21.16
4	9	3.4	11.56
5	4	-1.6	2.56
6	4	-1.6	2.56
7	5	-0.6	0.36
8	6	0.4	0.16
9	8	2.4	5.76
10	9	3.4	11.56
Total	**56**	**0**	**58.4**

Mean = 5.6

Vì vậy, người ta phải tính tổng bình phương của chúng (sum of squares), ký hiệu là:

$$\mathbf{SS} = \sum (X - \overline{X})^2$$

$= (-1.6)^2 + (0.4)^2 + \ldots$v.v…

$= (2.56) + 0.16) + \ldots$v.v….

= <u>58.4</u>

Đây chính là tổng giá trị mà các quan sát riêng lẻ chệch ra khỏi GTTB được bình phương lên.

2. Phương sai (variance)

Sum of squared errors (SS) là đại lượng đo lường tính biến thiên tốt hơn so với độ lệch (error) ban đầu.

Tuy nhiên, nó phụ thuộc nhiều vào số lượng quan sát mà ta có. Càng có nhiều quan sát, SS càng lớn.

Vì vậy, ta chia SS cho tổng số quan sát. Đây chính là số trung bình của tổng bình phương các độ lệch

$$Variance = \frac{SS}{n} = \frac{\sum(x-\bar{x})^2}{n}$$

$$= \frac{58.4}{10} = 5.84$$

Với:
n là cỡ mẫu
n – 1 là bậc tự do (degree of freedom)

Thực tế: $$s^2 = \frac{sum\_of\_squared\_deviation}{sample\_size - 1} = \frac{\sum(x-\bar{x})^2}{n-1}$$

3. Độ lệch chuẩn

$$SD = \sqrt{\frac{\Sigma(X-\bar{X}^2}{N-1}}$$

Với:

Σ: tổng

X: giá trị quan sát được

$\overline{X}$: giá trị trung bình của X

n: cỡ mẫu

n – 1: bậc tự do

Độ lệch chuẩn thể hiện khoảng lệch tiêu biểu hay là khoảng lệch trung bình của một quan sát so với giá trị trung bình. Độ lệch chuẩn càng lớn thì dữ liệu càng biến thiên mạnh.

Đo lường xu hướng phân tán

- Tổng bình phương độ lệch – sum of squared errors (SS)
- Phương sai – variance
- Độ lệch chuẩn – standard deviation (SD)

=> Cùng thể hiện (1) tính biến thiên của dữ liệu, (2) khả năng dự báo của mô hình giá trị trung bình; (3) độ lệch (gây ra bởi sai lệch giữa quan sát đơn lẻ so với GTTB).

Tóm tắt bài học

- Việc phân loại và nhận biết loại dữ kiện (cấp độ đo lường) rất cần thiết vì nó quyết định nhiều vấn đề trong suốt quá trình nghiên cứu, từ chọn biểu đồ, chọn đại lượng biểu diễn xu hướng tập trung, cho đến loại kiểm định thống kê.
- Mean, mode, median đều có điểm mạnh và yếu, tùy trường hợp mà dùng.
- Độ lệch chuẩn và giá trị trung bình (mean) là hai thông số mô tả dữ liệu quan trọng và được dùng phổ biến nhất.