Tóm tắt dữ liệu

Bài 2

Linh Truong

Jan, 2024

Nội dung

- ĐẶC ĐIỂM DỮ LIỆU
- MÔ TẢ BẰNG CÁC ĐỘ ĐO
- ĐỊNH VỊ DỮ LIỆU
- MÔ TẢ BẰNG BIỂU ĐỒ
- TÓM TẮT

Mục tiêu

- Biết một số đặc điểm của dữ liệu
- Lựa chọn được độ đo và đồ thị phù hợp để mô tả dữ liệu

ĐẶC ĐIỂM DỮ LIỆU

Khi khảo sát dữ liệu cần tìm hiểu một số đặc điểm như sau:

- **Xu hướng tập trung (Central tendancy)**: dữ liệu tập trung ở đâu? giá trị nào có thể đại diện được cho dữ liệu? Dữ liệu có xu hướng tập trung hay không?
- **Sự phân tán (Variability)**: dữ liệu phân tán bao nhiêu? liệu có giá trị bất thường nào không?
- **Hình dạng phân phối (shape)**: phân phối dữ liệu có đối xứng hay bị lệch? Phẳng hay nhọn?

MÔ TẢ BẰNG CÁC ĐỘ ĐO

Độ đo xu hướng tập trung

Statistic	Detail
mean ($\bar{x}$)	$\frac{1}{n}\sum_{i=1}^{n} x_i$
median	Giá trị chính giữa khi dữ liệu được sắp
mode	Giá trị xuất hiện nhiều nhất
midrange	$\frac{x_{min}+x_{max}}{2}$
geometric mean	$\sqrt[n]{x_1 x_2 \dots x_n}$

Độ đo xu hướng tập trung

Trung bình

Giả sử dữ liệu có n giá trị $x_1, x_2, \ldots, x_n$. Giá trị trung bình sẽ được tính:

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

Trung bình theo trọng số

Trường hợp dữ liệu được tổ chức theo dạng bảng phân phối tần số như sau:

Giá trị	x_1	x_2	x_3	.....	x_k
Tần số	w_1	w_2	w_3	.....	w_k

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

Ví dụ

Một mẫu thu nhận được có các giá trị sau:

$$4, 4, 5, 5, 5, 5, 6, 6, 7, 7$$

Tính giá trị trung bình?

5.45

Một mẫu thu nhận được có các bảng phân phối tần số như sau:

Giá trị	**4**	**5**	**6**	**7**	**300**
Tần số	2	4	2	2	1

Tính giá trị trung bình mẫu.

30

Ví dụ

Chú ý

Trường hợp bảng phân phối dữ liệu chứa dữ liệu dạng khoảng, có thể dùng dữ liệu trung tâm để đại diện cho khoảng đó

Bảng khảo sát về thời gian chơi game ở lứa tuổi thiếu niên:

Thời gian chơi (giờ)	**0 - 3.5**	**3.5 - 7.5**	**7.5 - 11.5**	**11.5 - 16.5**	**15.5 - 19.5**
Số lượng thiếu niên	3	7	12	7	9

Tính thời gian chơi game trung bình của những người tham gia khảo sát.

Độ đo xu hướng tập trung

Trung vị (median)

Là giá trị trung tâm của dãy các giá trị dữ liệu đã được sắp.

Ví dụ

Tìm trung vị của dãy: 1, 2, 3, 4, 5, 6, 7, 8, 9

5 (1, 2, 3, 4, **5**, 6, 7, 8, 9)

Tìm trung vị của dãy: 1, 2, 2, 2, 3, 4, 5, 7, 7, 8

3.5 (1, 2, 2, 2, **3, 4**, 5, 7, 7, 8)

Độ đo xu hướng tập trung

Yếu vị (mode)

Là giá trị dữ liệu có tần số xuất hiện nhiều nhất. Nếu mọi giá trị dữ liệu đều có tần số như nhau, ta nói dữ liệu không có yếu vị

Ví dụ

Khảo sát tuổi của 6 sinh viên, người ta thu được số liệu:

19, 22, 20, 21, 22, 23

Tìm giá trị yếu vị?

22

Ví dụ

Bảng dữ liệu dưới đây khảo sát bằng cấp của phụ huynh các đối tượng tham gia khảo sát:

Data set

Participant	A	B	C	D	E	F
Parents' education level	Bachelor's degree	Master's degree	High school diploma	Bachelor's degree	Doctoral degree	Master's degree

Tìm giá trị yếu vị?

Có hai yếu vị: **Bachelor's degree** và **Master's degree**

Độ đo xu hướng tập trung

Trung bình hình học

Giả sử mẫu có n giá trị $x_1, x_2, \ldots, x_n$. Giá trị trung bình hình học sẽ được tính:

$$\sqrt[n]{x_1 x_2 \ldots x_n} = (x_1 x_2 \ldots x_n)^{\frac{1}{n}}$$

- Trung bình hình học thường chính xác hơn giá trị trung bình số học khi thể hiện phần trăm những thay đổi theo thời gian.

Ví dụ

- Giả sử bạn đầu tư số tiền \$100 với lãi suất là r_1 năm thứ nhất, lãi suất là r_2 năm thứ 2, ..., lãi suất là r_n năm thứ n
  - Tổng số tiền sau năm 1: $A_1 = 100(1 + r_1)$
  - Tổng số tiền sau năm 2: $A_2 = A_1(1 + r_2) = 100(1 + r_1)(1 + r_2)$
  - Tổng số tiền sau năm 3: $A_3 = A_2(1 + r_3) = 100(1 + r_1)(1 + r_2)(1 + r_3)$
  - $\ldots$
  - Tổng số tiền sau năm n: $A_n = 100(1 + r_1)(1 + r_2)\ldots(1 + r_n)$

- Tỷ suất lợi nhuận gộp trung bình sau n năm:

$$r_g = \sqrt[n]{(1 + r_1)(1 + r_2)\ldots(1 + r_n)} - 1$$

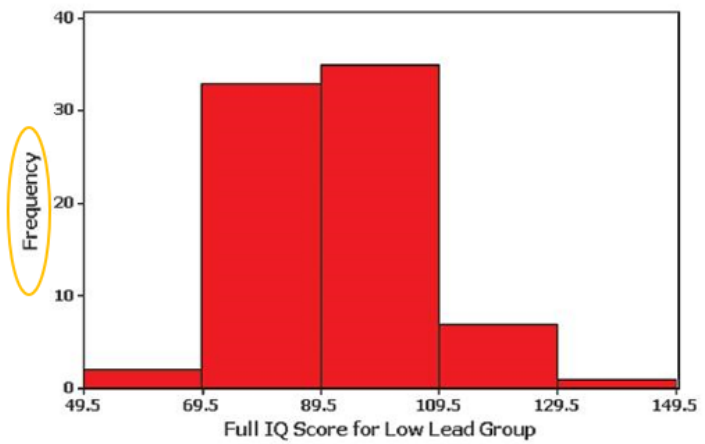
Hình dạng phân phối

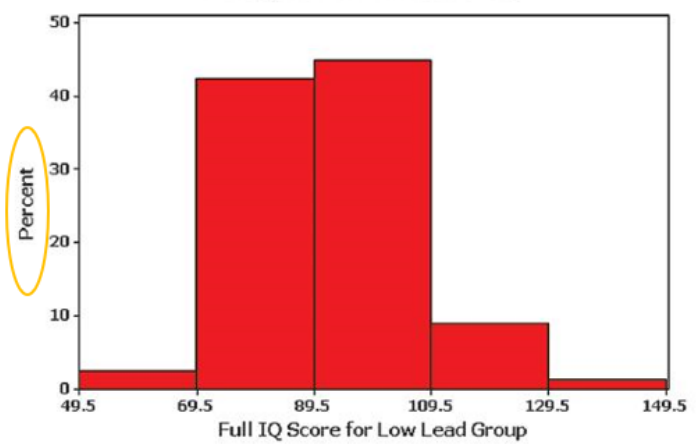
Histogram

Đồ thị Histogram dùng để biểu diễn phân phối tần số của dữ liệu, với một trục thể hiện giá trị, trục còn lại thể hiện tần số hoặc tần số tương đối. Có thể dùng đồ thị histogram để xem hình dạng của phân phối dữ liệu.

Bảng dữ liệu

IQ Score	Frequency	Relative Frequency
50-69	2	2.6%
70-89	33	42.3%
90-109	35	44.9%
110-129	7	9.0%
130-149	1	1.3%





Ví dụ

Dữ liệu số sách đọc trong tháng

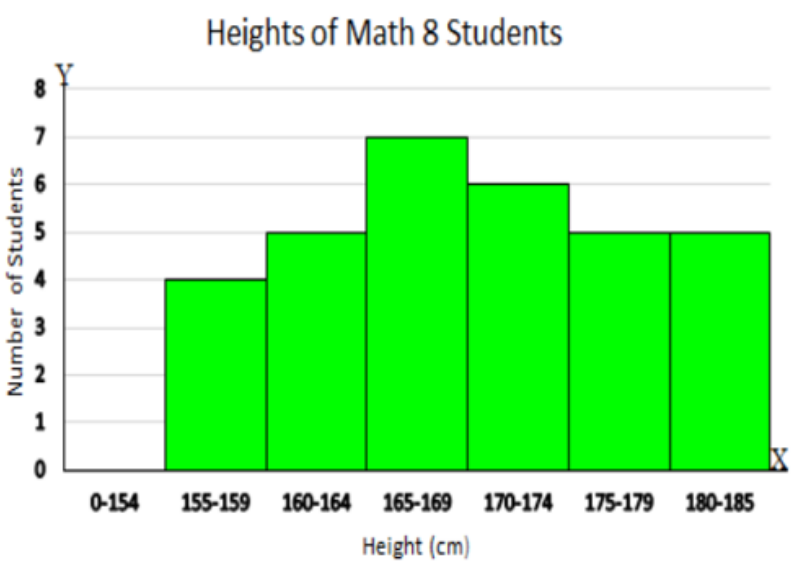
Number of books read in a month	Frequency
2	1
3	3
4	5
5	2
6	1

Ví dụ

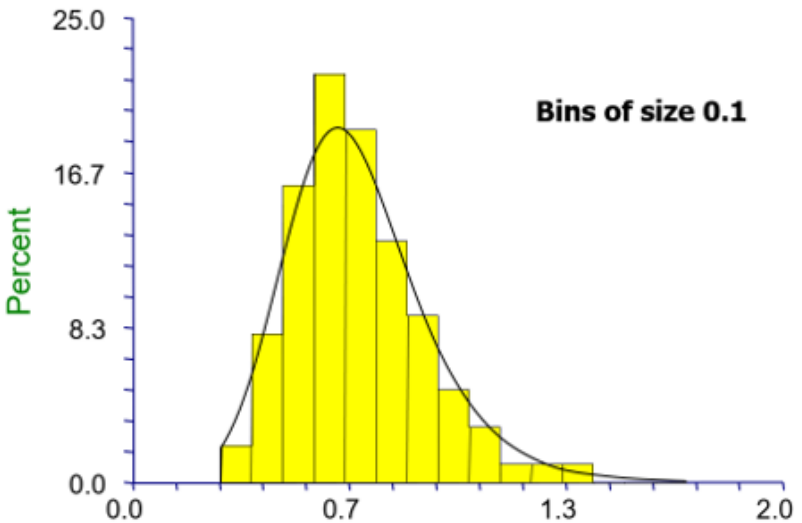
Dữ liệu chiều cao của một số sinh viên lớp Toán

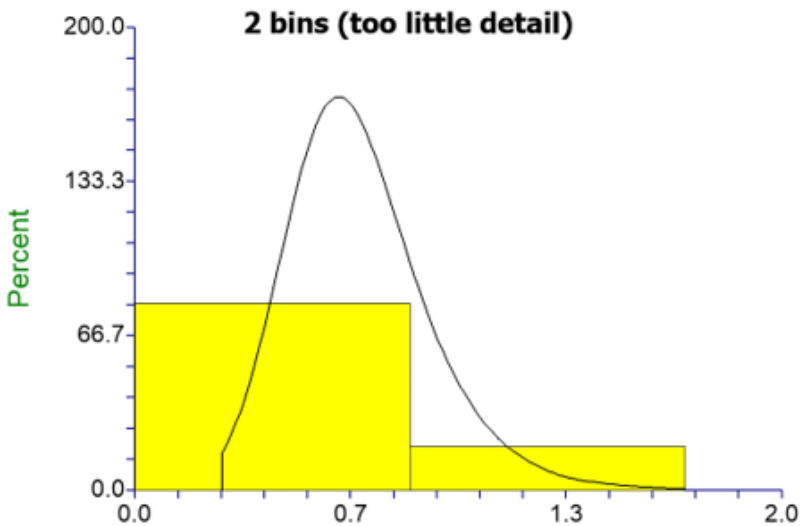
158	177	175	183	159	172	160	168
169	167	161	157	162	164	168	176
156	179	174	165	163	170	173	180
174	181	176	174	180	167	182	169

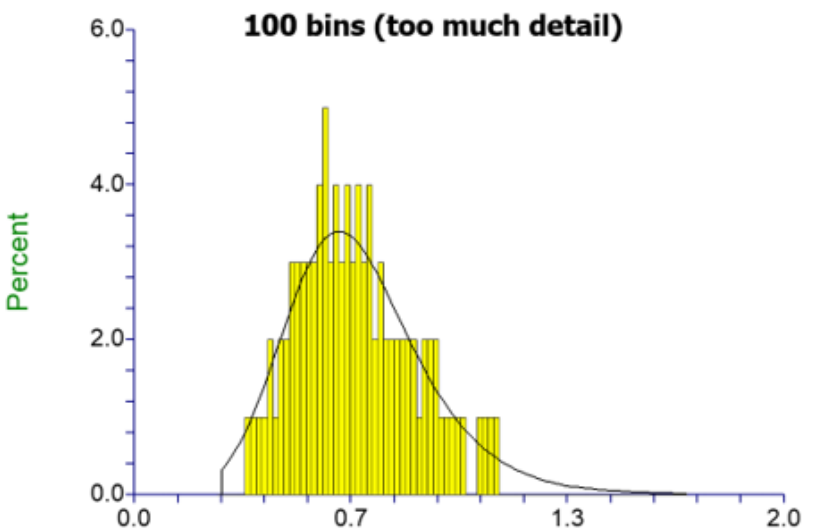
Intervals	Tally	Frequency
0 - 154		0
155 - 159	𝍸	4
160 - 164	𝍸	5
165 - 169	𝍸 ||	7
170 - 174	𝍸 |	6
175 - 179	𝍸	5
180 - 189	𝍸	5



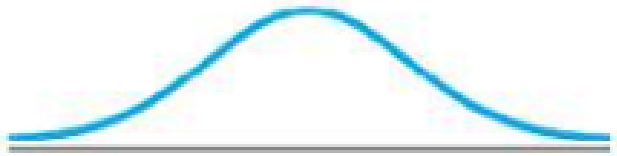
Chọn bao nhiêu cột?







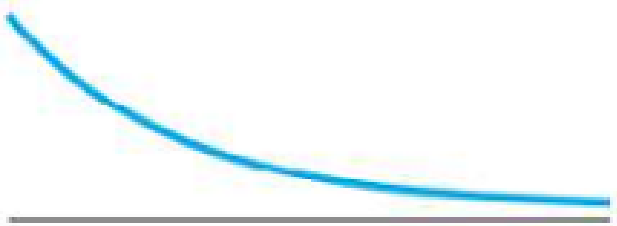
Một số hình dạng phân phối



(a) Bell shaped

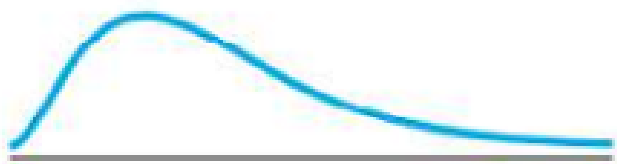
(b) Triangular

(c) Uniform (or rectangular)

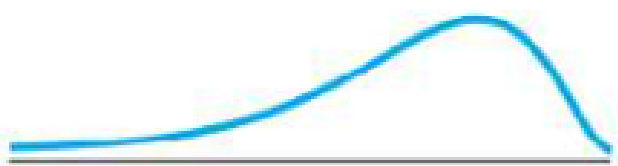


(d) Reverse J shaped

(e) J shaped

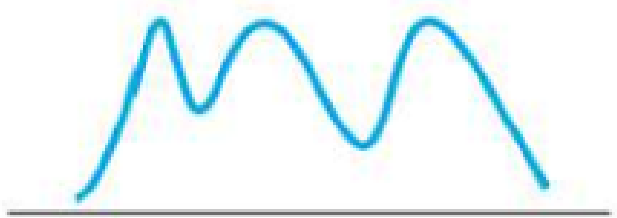


(f) Right skewed



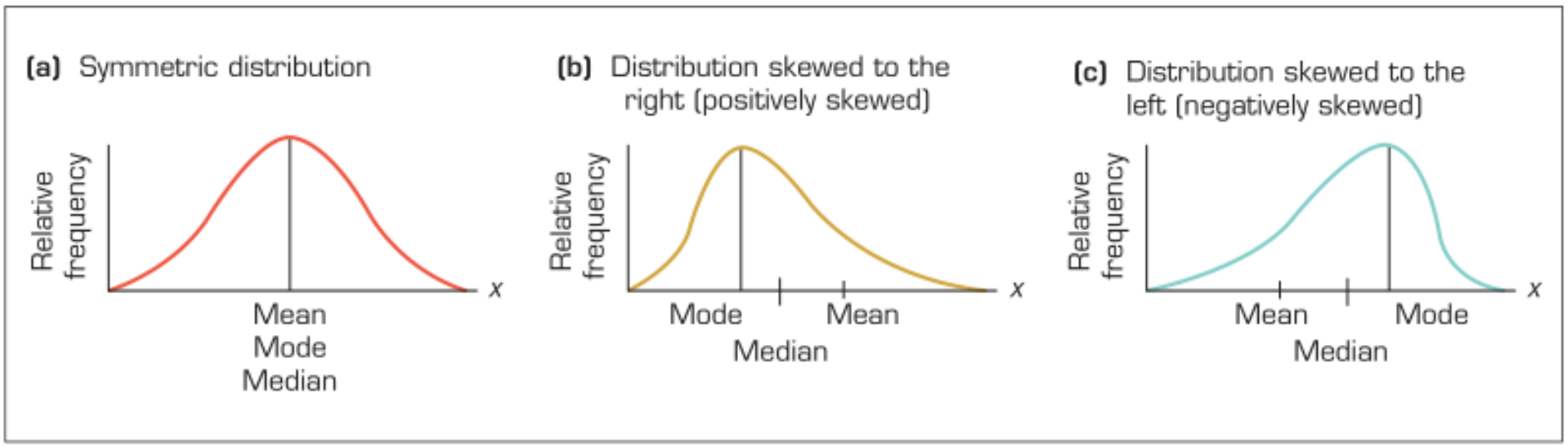
(g) Left skewed

(h) Bimodal



(i) Multimodal

So sánh


(a) Symmetric distribution
Relative frequency
x
Mean
Mode
Median
(b) Distribution skewed to the right (positively skewed)
Relative frequency
x
Mode
Mean
Median
(c) Distribution skewed to the left (negatively skewed)
Relative frequency
x
Mean
Mode
Median

Độ đo sự phân tán (Central Tendency)

Statistic	Formula
range	$x_{max} - x_{min}$
sample variance (s^2)	$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
sample standard deviation (s)	$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

ĐỊNH VỊ DỮ LIỆU

Câu hỏi

> Giả sử trong kỳ thi Đại Học bạn đạt được số điểm khá cao là 25 điểm (điểm tối đa là 30). Tuy nhiên, theo chỉ tiêu nhà trường chỉ lấy 60% tổng số hồ sơ đăng ký dự thi. Làm thế nào bạn biết được mình có được tuyển hay không?

Nếu có thể xác định được vị trí của giá trị 25 trong toàn bộ phân phối điểm thì có thể biết kết quả.

Hoặc nếu biết được ngưỡng điểm có thể loại bớt 40% thí sinh có điểm thấp thì có thể biết kết quả xét tuyển

Khi đó giá trị của ngưỡng điểm được gọi là phân vị thứ 40 của dữ liệu.

Percentile

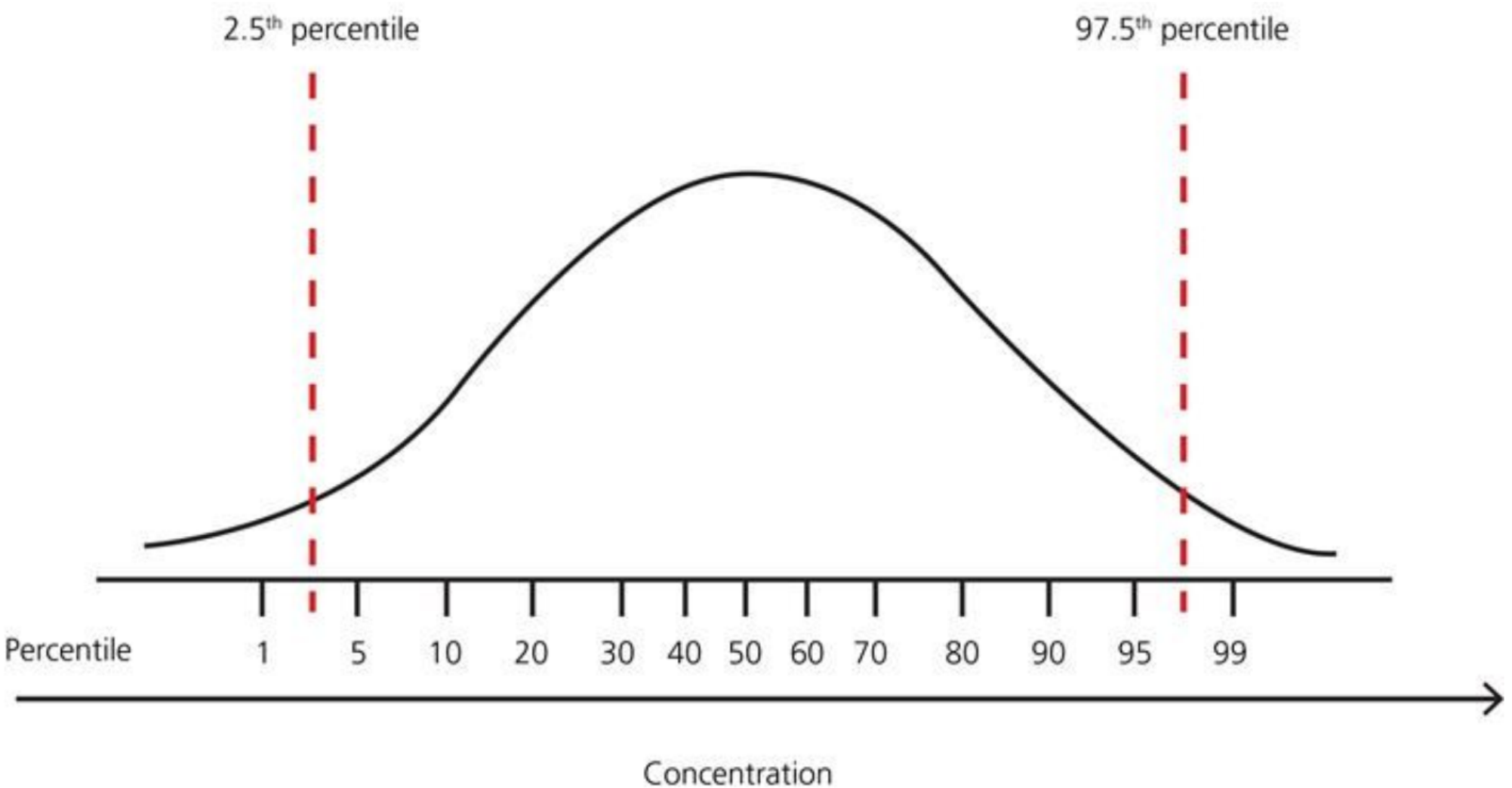
Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:

That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Percentile


2.5th percentile
97.5th percentile
Percentile
1
5
10
20
30
40
50
60
70
80
90
95
99
Concentration

Tứ phân vị (Quartiles)

Tứ phân vị

Nếu ta sắp dữ liệu tăng dần sau đó chọn ra 3 điểm Q_1, Q_2, Q_3 để chia toàn bộ dữ liệu ra làm 4 phần đều nhau, mỗi phần chiếm 25% của dữ liệu. Các điểm này được gọi là các tứ phân vị. Trong đó:

- Q_1 là phân vị thứ 25
- Q_2 là phân vị thứ 50 (trung vị)
- Q_3 là phân vị thứ 74

Miền phân vị (IQR - Interquartile range)

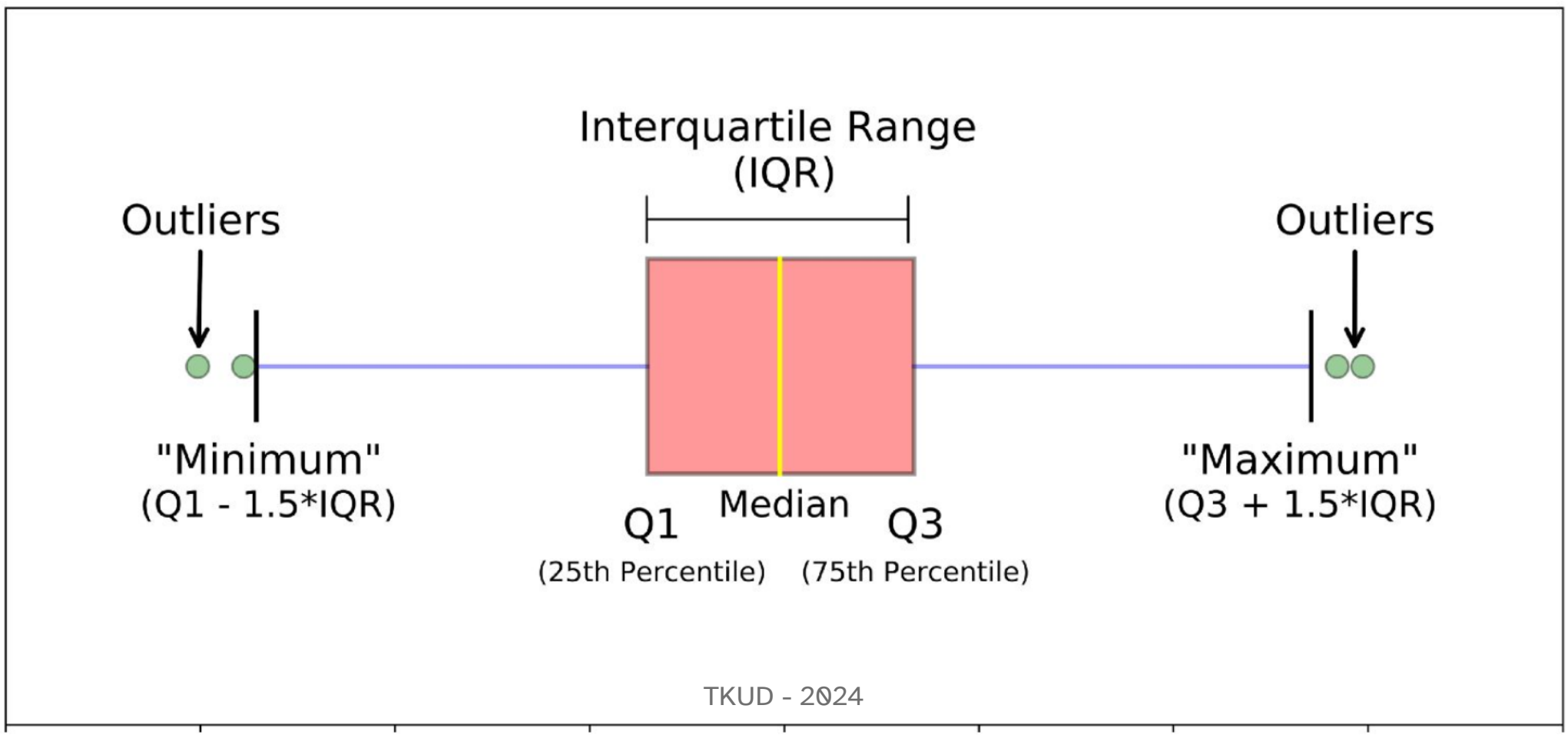
Là khoảng cách giữa hai giá trị phân vị Q_3 và Q_1.

$$IQR = Q_3 - Q_1$$

Miền phân vị thường được dùng để đo độ biến thiên của dữ liệu

Boxplot

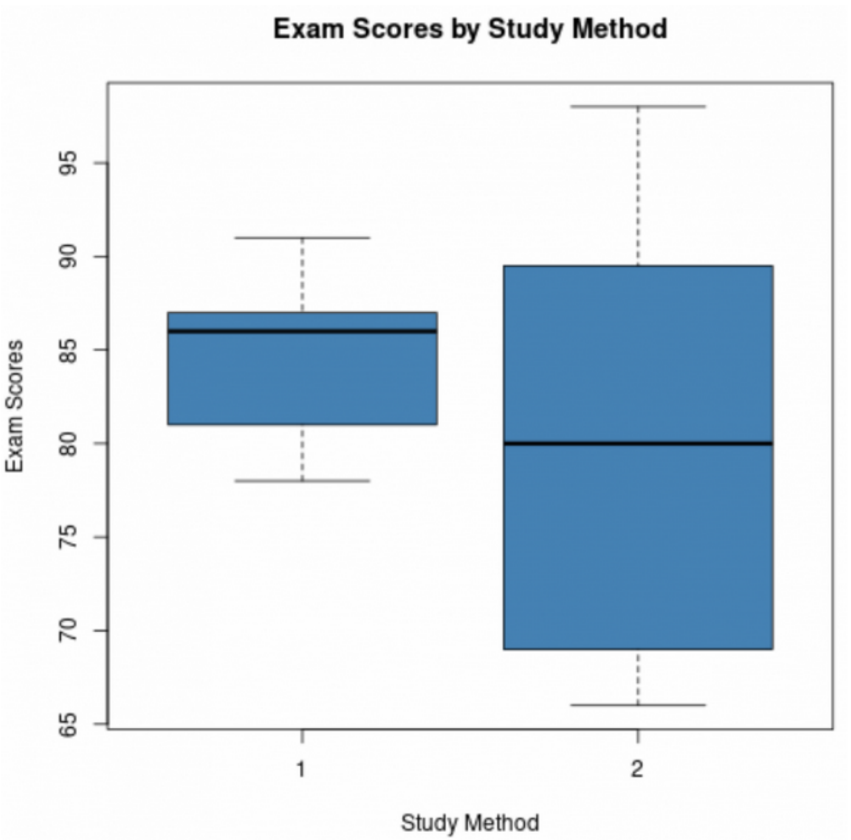
- Đồ thị hộp (boxplot) được vẽ dựa trên các điểm tứ phân vị
- Có thể sử dụng để mô tả độ tập trung, độ phân tán, ngoại lệ của dữ liệu



Ví dụ

Dữ liệu điểm thi dưới đây thể hiện điểm thi của sinh viên sử dụng hai phương pháp khác nhau để chuẩn bị cho kỳ thi:

- **Phương pháp 1**: 78, 78, 79, 80, 80, 82, 82, 83, 83, 86, 86, 86, 86, 87, 87, 87, 88, 88, 88, 91
- **Phương pháp 2**: 66, 66, 66, 67, 68, 70, 72, 75, 75, 78, 82, 83, 86, 88, 89, 90, 93, 94, 95, 98



MÔ TẢ BẰNG BIỂU ĐỒ

Trực quan hóa

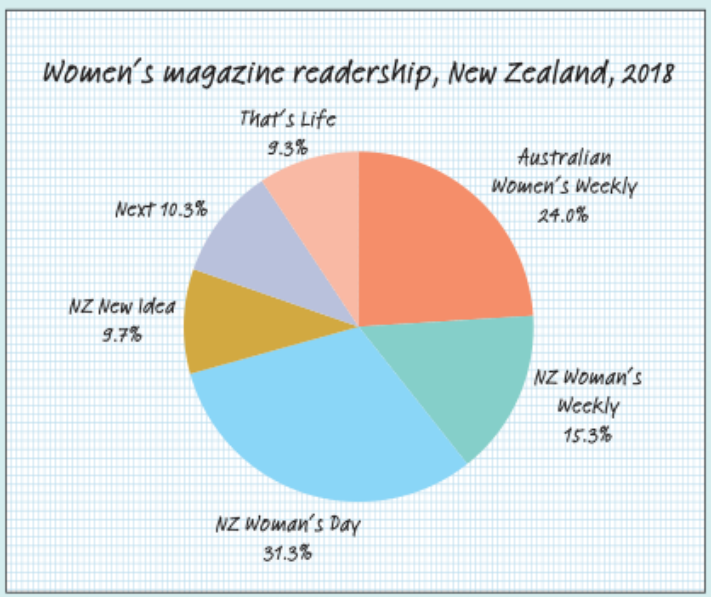
- Một đồ thị vẽ tốt, giúp dễ đọc và mang lại nhiều thông tin có giá trị
- Vậy lựa chọn đồ thị nào để biểu diễn:
  - Cho biến định tính?
  - Cho biến định lượng?
  - Mối quan hệ giữa các biến

Bar chart (đồ thị cột)

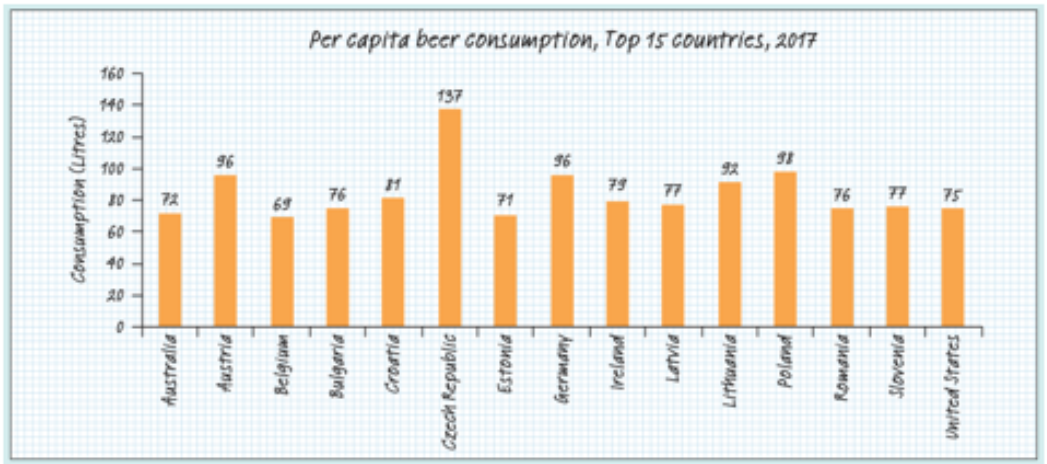
Pie chart (đồ thị tròn)

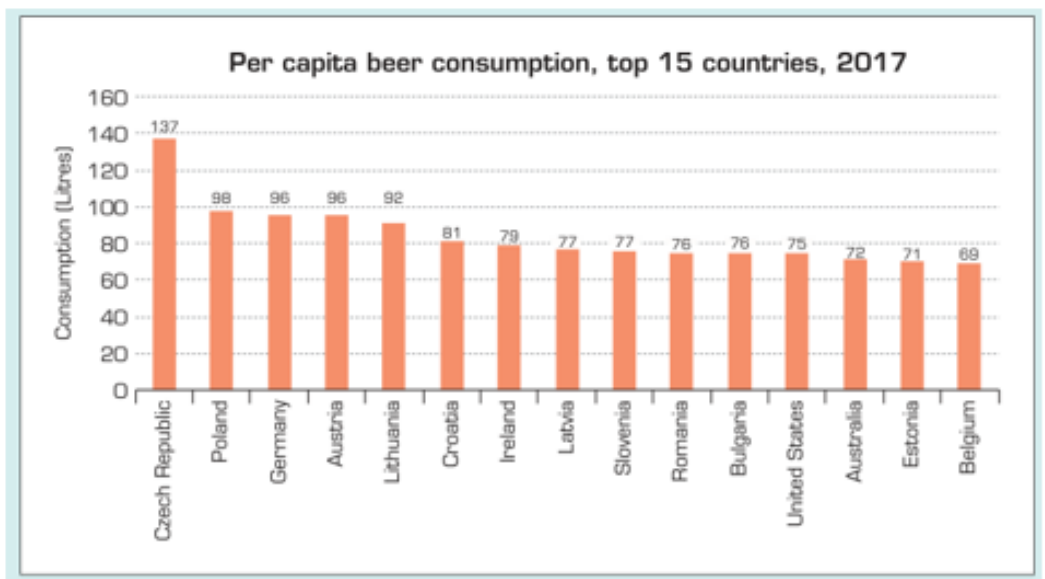
Magazine	Proportion of readers (in percentages)	Angle of the slice
Australian Women's Weekly (1)	24.0	24.0 × 3.6 = 86.4°
NZ Woman's Weekly (2)	15.3	15.3 × 3.6 = 55.2°
NZ Woman's Day (3)	31.3	31.3 × 3.6 = 112.8°
New Idea (4)	9.7	9.7 × 3.6 = 34.8°
Next (5)	10.3	10.3 × 3.6 = 37.2°
That's Life (6)	9.3	9.3 × 3.6 = 33.6°
Total	**100.0**	**360°**

FIGURE 3.2 Pie chart for Example 3.1

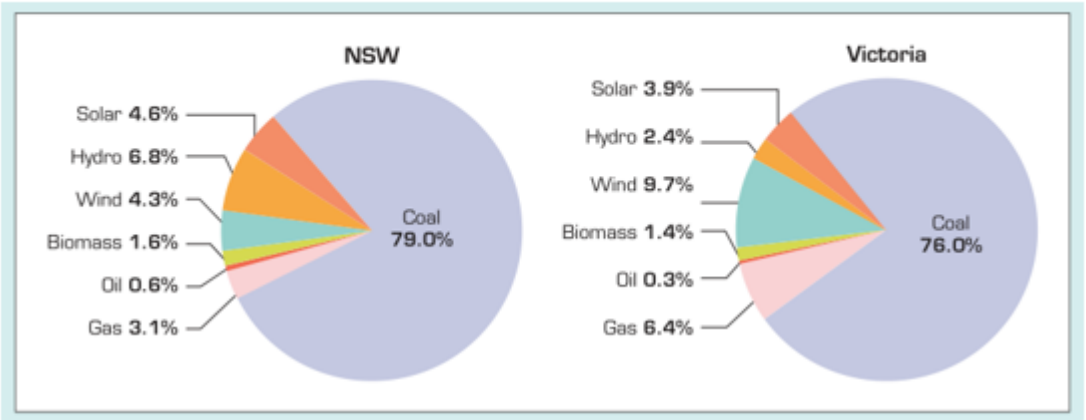


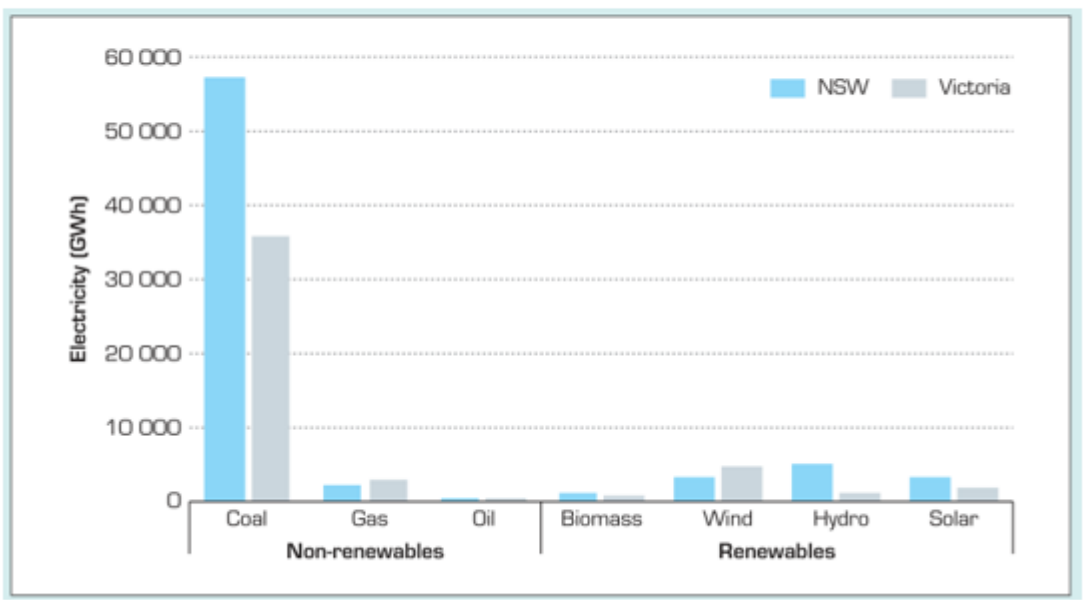
So sánh


Per capita beer consumption, Top 15 countries, 2017
Consumption (Litres)
160
140
120
100
80
60
40
20
0
72
96
69
76
81
137
71
96
79
77
92
98
76
77
75
Australia
Austria
Belgium
Bulgaria
Croatia
Czech Republic
Estonia
Germany
Ireland
Latvia
Lithuania
Poland
Romania
Slovenia
United States

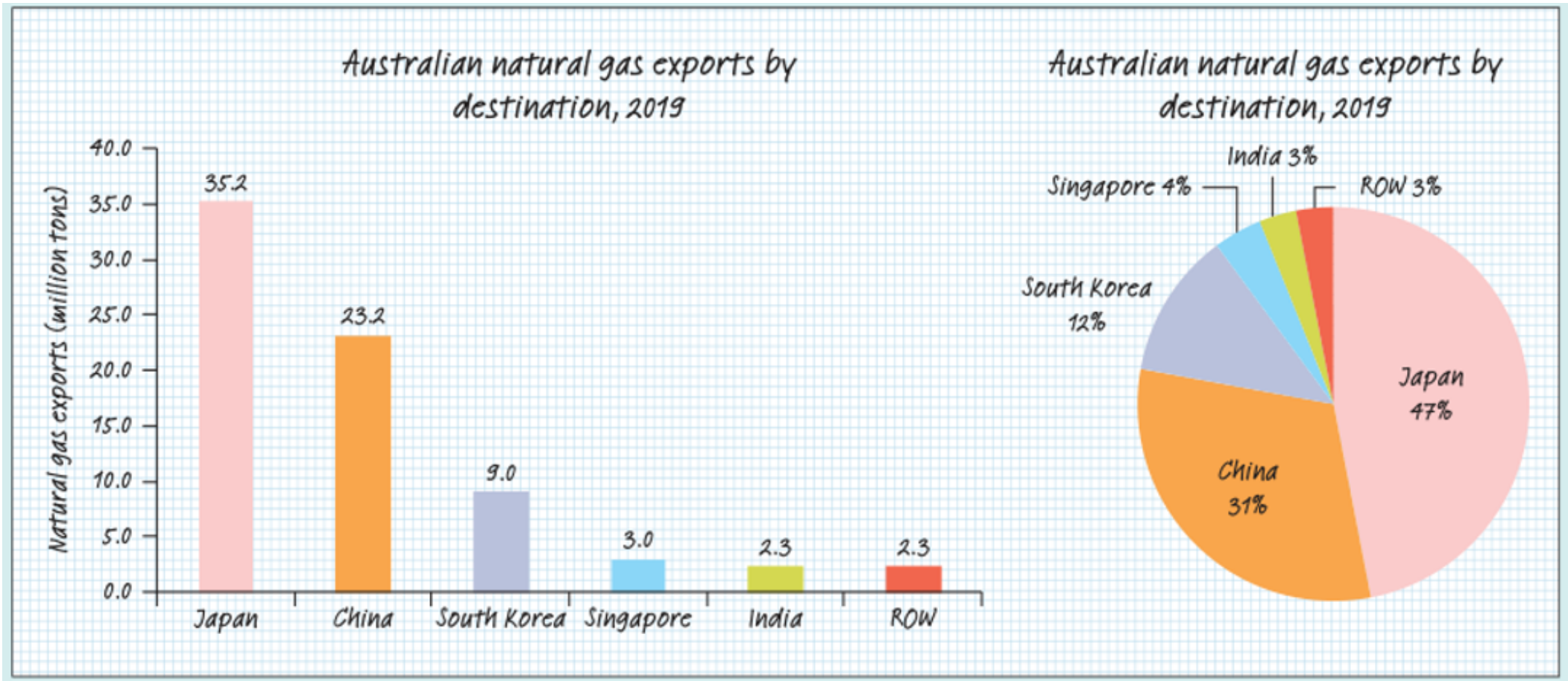

Per capita beer consumption, top 15 countries, 2017
Consumption (Litres)
160
140
120
100
80
60
40
20
0
137
98
96
96
92
81
79
77
77
76
76
75
72
71
69
Czech Republic
Poland
Germany
Austria
Lithuania
Croatia
Ireland
Latvia
Slovenia
Romania
Bulgaria
United States
Australia
Estonia
Belgium

So sánh


NSW
Solar 4.6%
Hydro 6.8%
Wind 4.3%
Biomass 1.6%
Oil 0.6%
Gas 3.1%
Coal 79.0%
Victoria
Solar 3.9%
Hydro 2.4%
Wind 9.7%
Biomass 1.4%
Oil 0.3%
Gas 6.4%
Coal 76.0%


60 000
50 000
40 000
30 000
20 000
10 000
0
Electricity (GWh)
NSW
Victoria
Coal
Gas
Oil
Biomass
Wind
Hydro
Solar
Non-renewables
Renewables

Chọn đồ thị nào?



Ví dụ

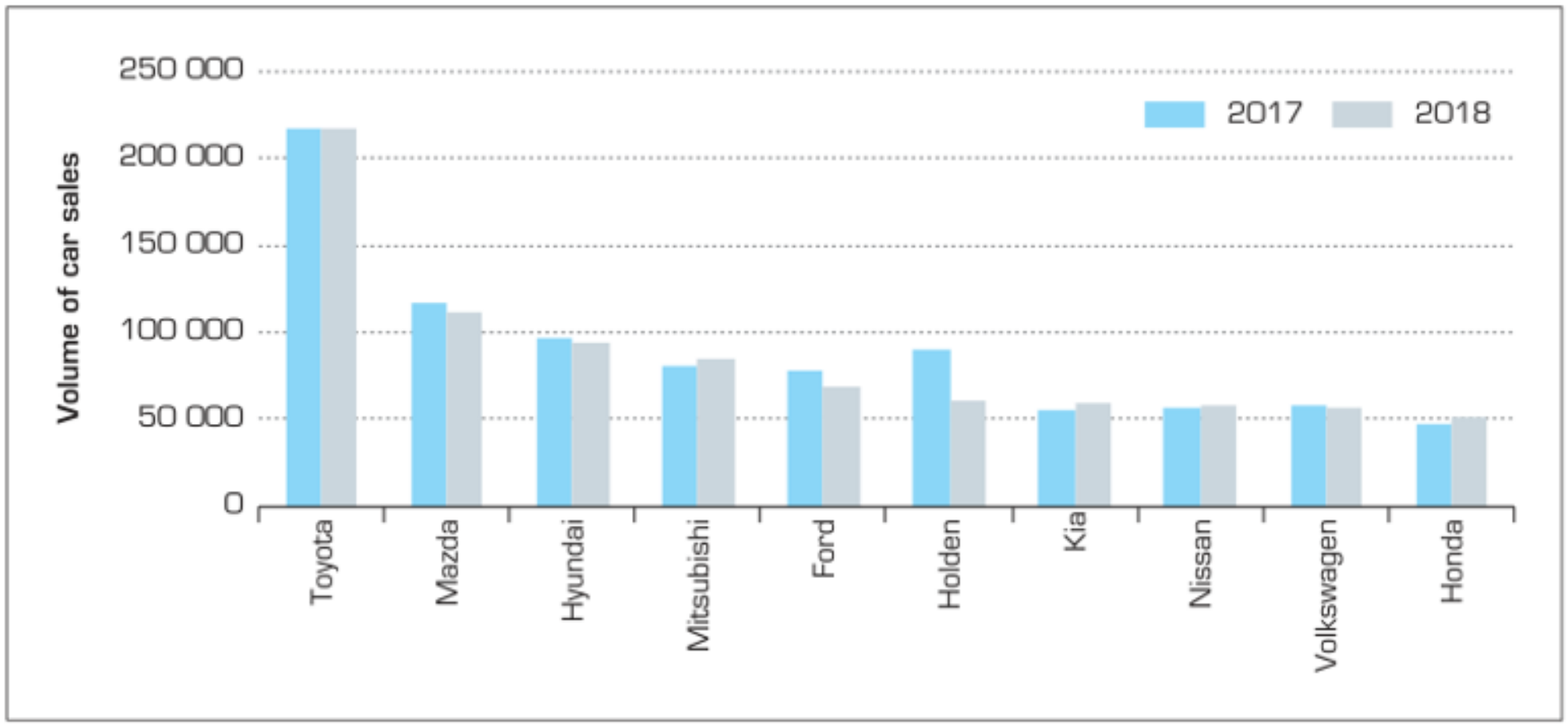
TABLE 3.5 Top 10 best-selling new passenger vehicle sales in Australia, by manufacturer 2017 and 2018

Rank	Model	Volume of sales		Change (%)	Market share	
		2017	2018		2017	2018
1	Toyota	216 566	217 061	0.2	24.2	25.2
2	Mazda	116 349	111 280	–4.4	13.0	12.9
3	Hyundai	97 013	94 187	–2.9	10.8	10.9
4	Mitsubishi	80 654	84 944	5.3	9.0	9.9
5	Ford	78 161	69 081	–11.6	8.7	8.0
6	Holden	90 306	60 751	–32.7	10.1	7.0
7	Kia	54 737	58 815	7.5	6.1	6.8
8	Nissan	56 594	57 699	2.0	6.3	6.7
9	Volkswagen	58 004	56 620	–2.4	6.5	6.6
10	Honda	46 783	51 525	10.1	5.2	6.0
	Total	**895 167**	**861 963**		**100.0**	**100.0**

Source: Federal Chamber of Automotive Industries, 2019.

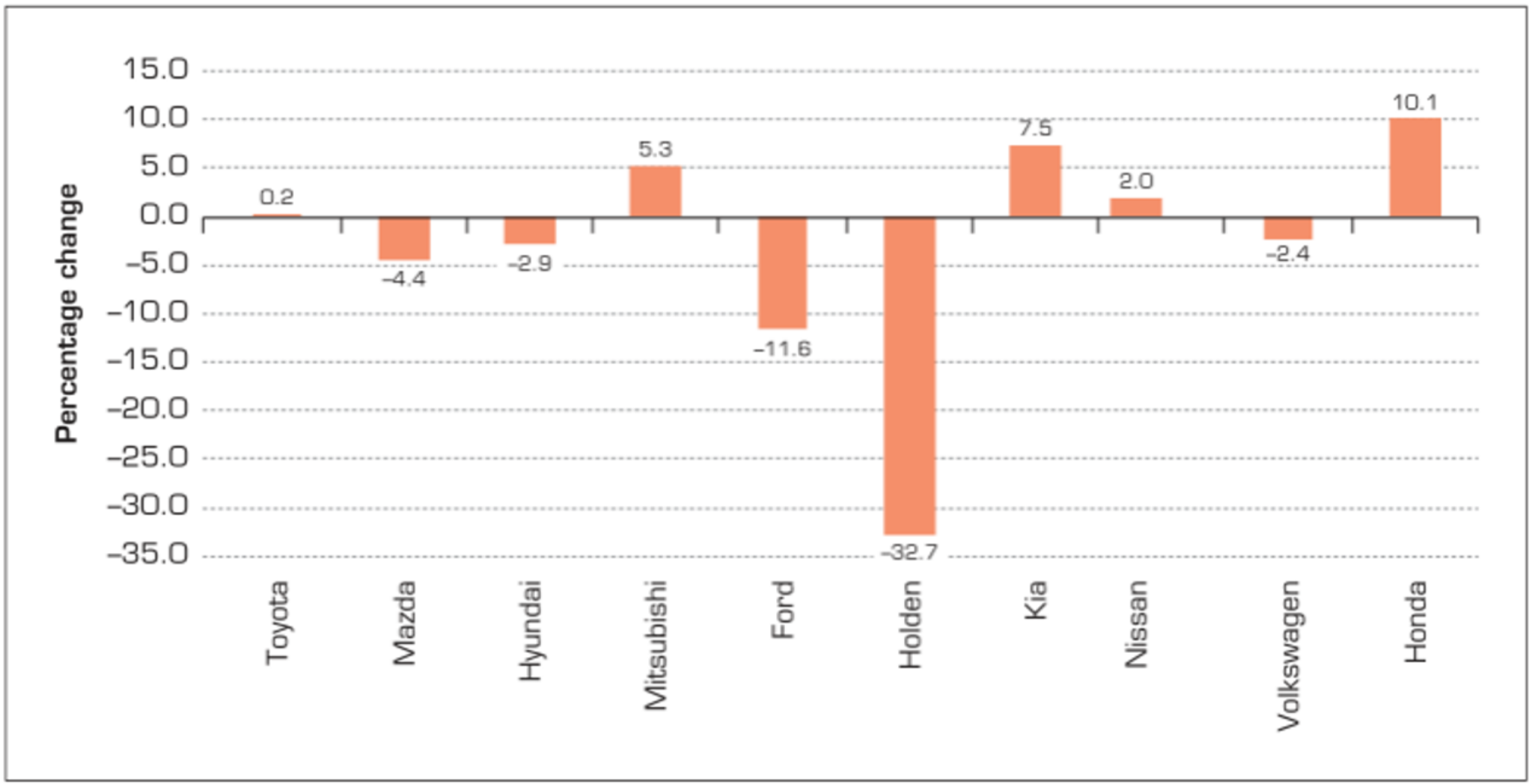
Ví dụ

FIGURE 3.6 Bar chart of new passenger vehicle sales, 10 best-selling manufacturers, 2017 and 2018



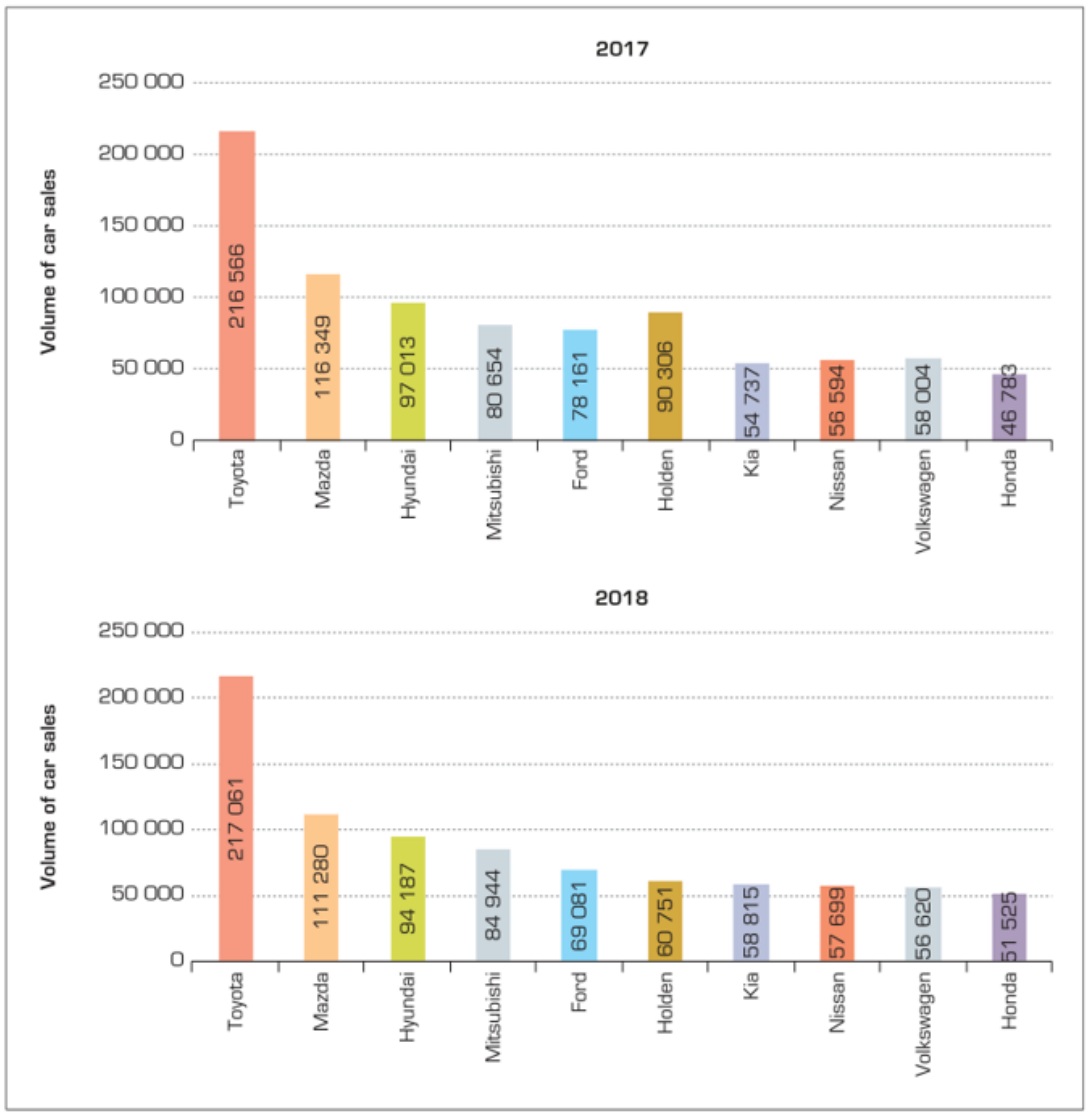
Ví dụ

FIGURE 3.7 Bar chart emphasising change in sales by manufacturer between 2017 and 2018



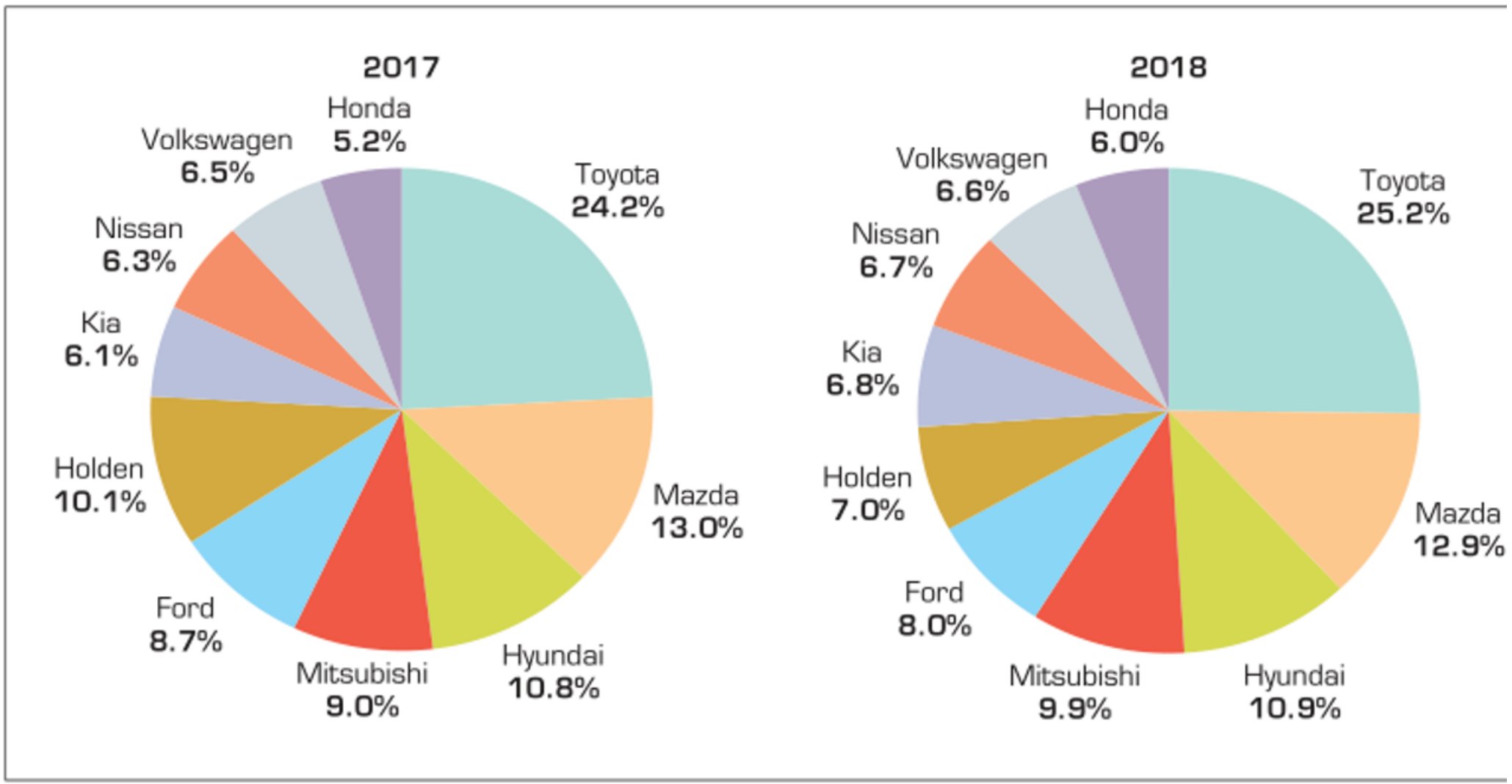
Ví dụ

FIGURE 3.8 Bar charts emphasising sales profile by year, 2017 and 2018



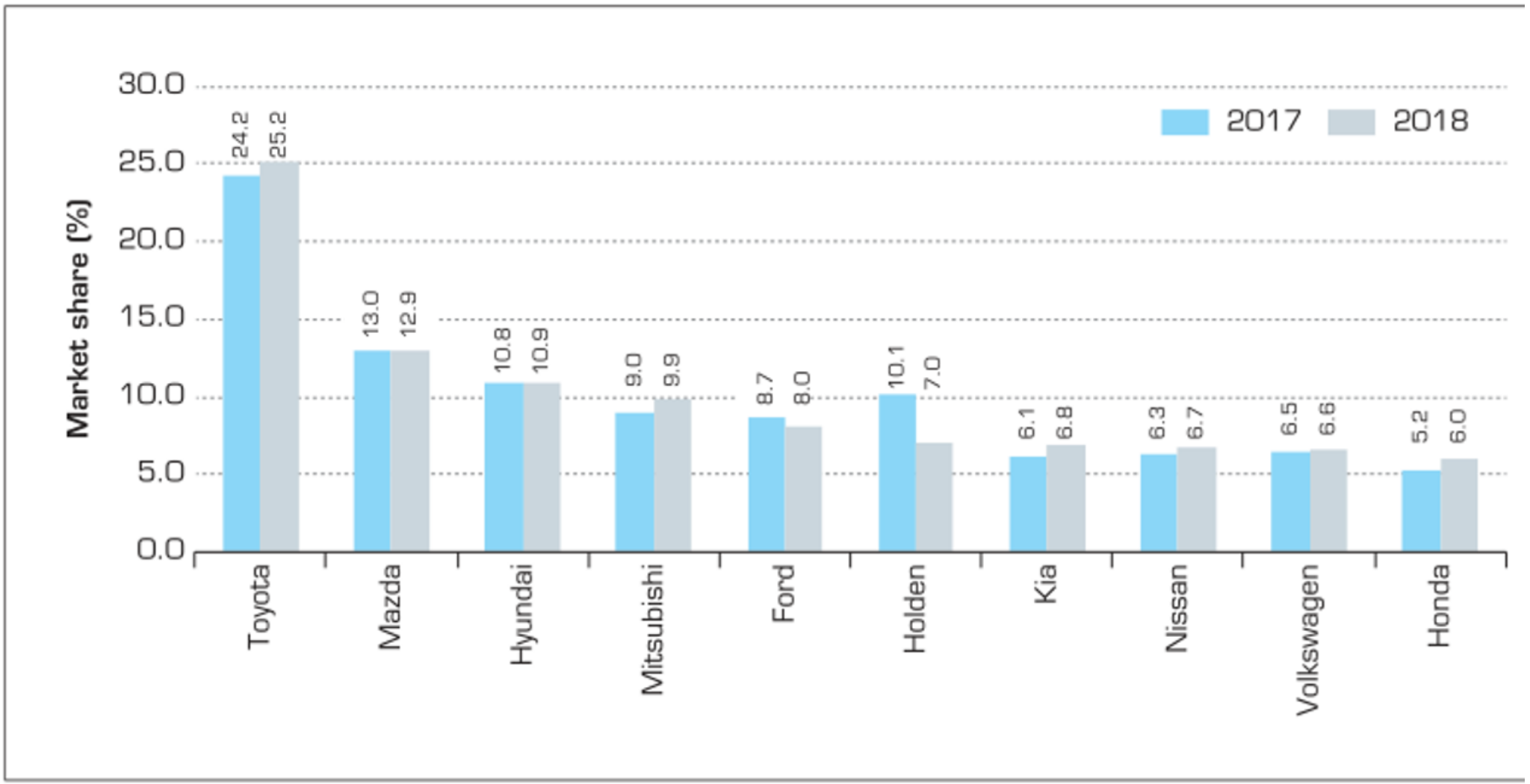
Ví dụ

FIGURE 3.9A Pie charts emphasising change in market share, 2017 vs 2018



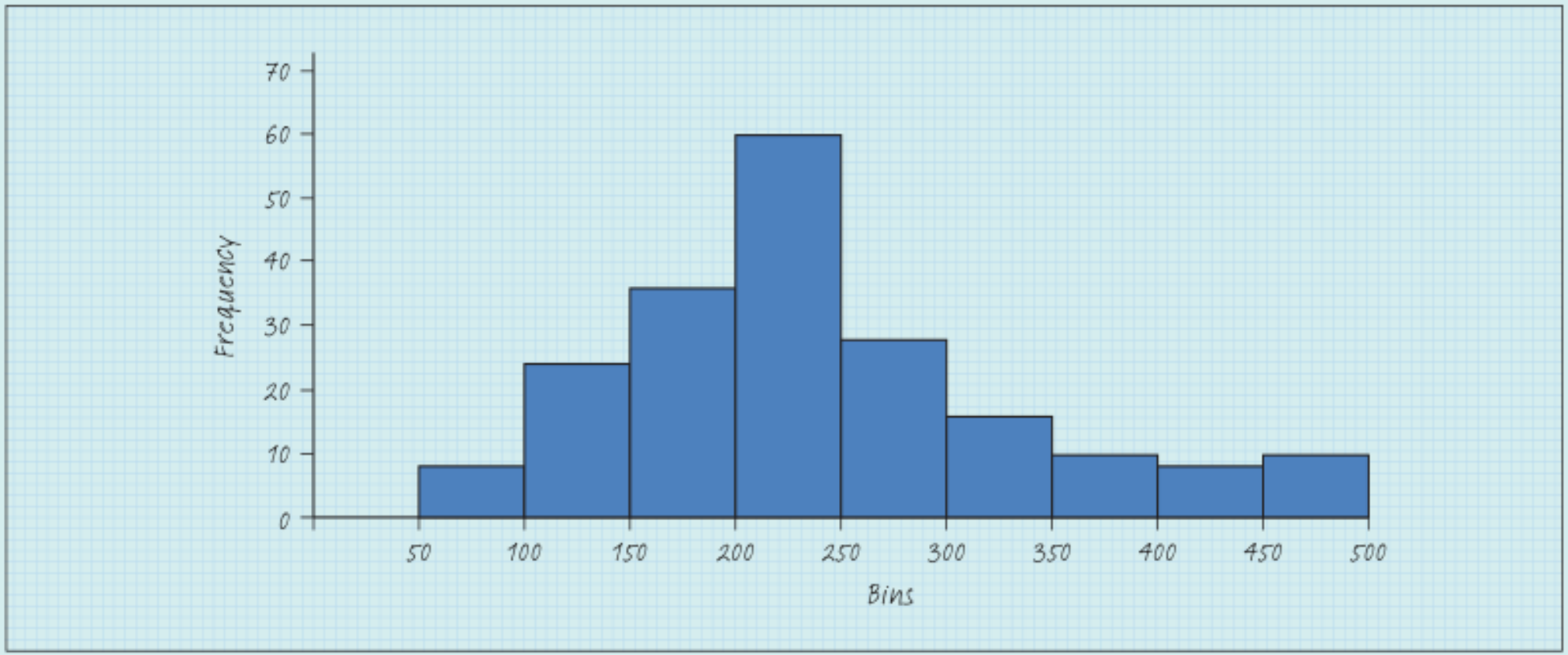
Ví dụ

FIGURE 3.9B Bar chart emphasising change in market shares, 2017 vs 2018

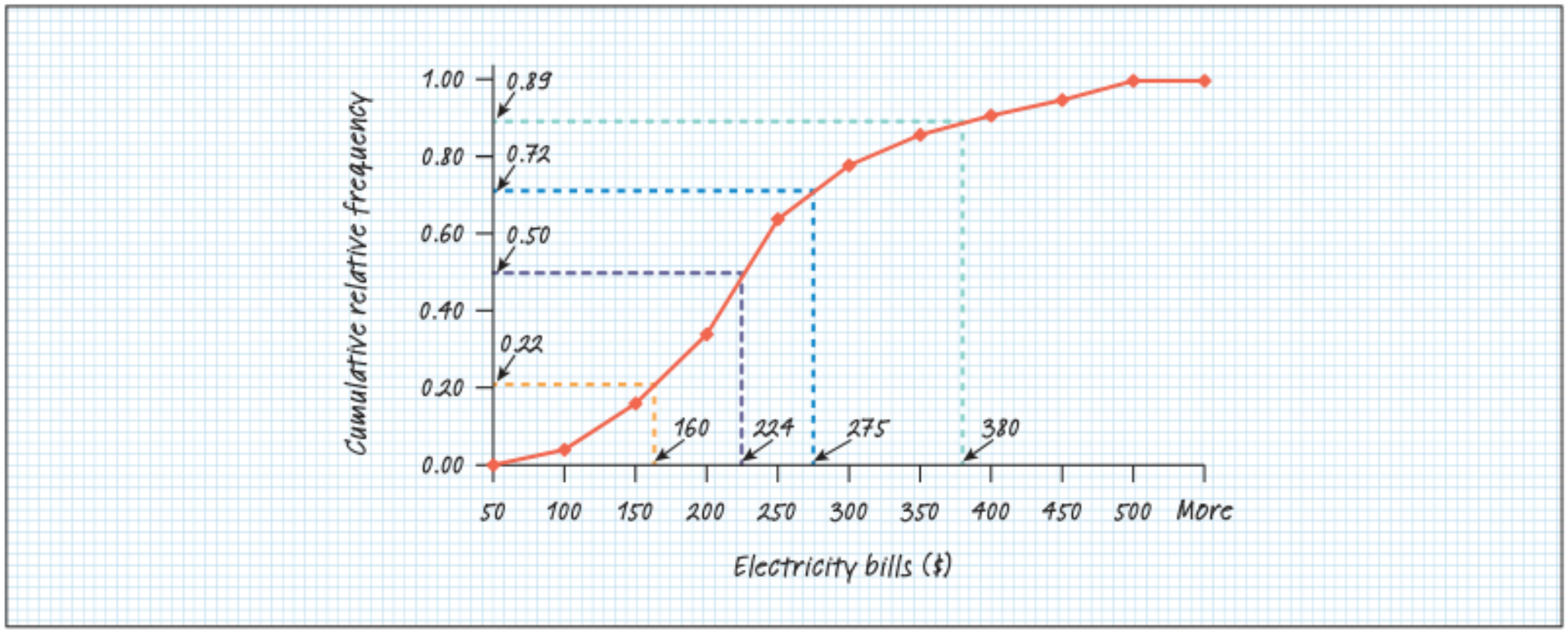


Histogram (đồ thị phân phối tần số)

FIGURE 4.1 Histogram of electricity bills of 200 new Brisbane customers

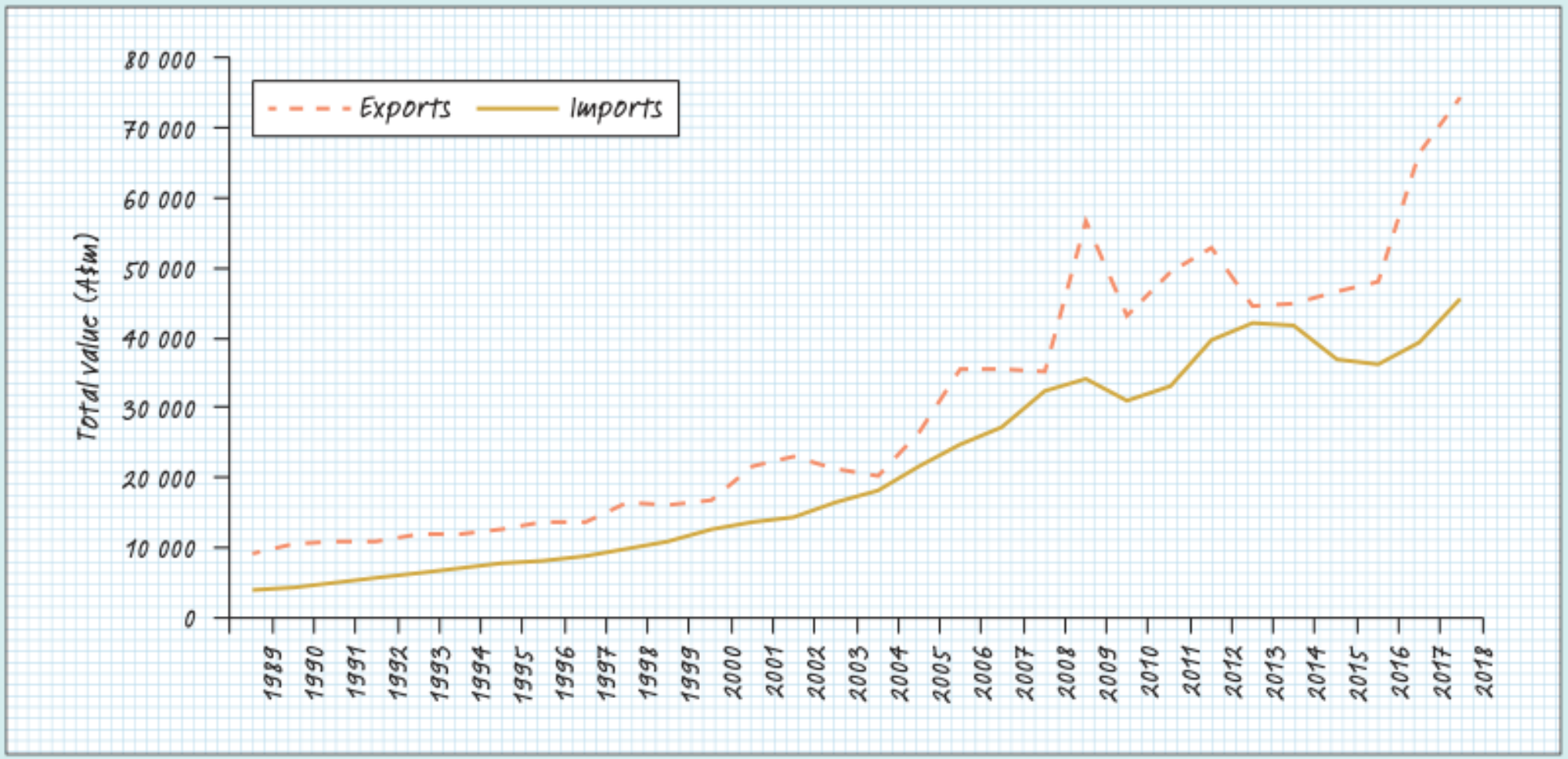


Cumulative Distribution Function (CDF)



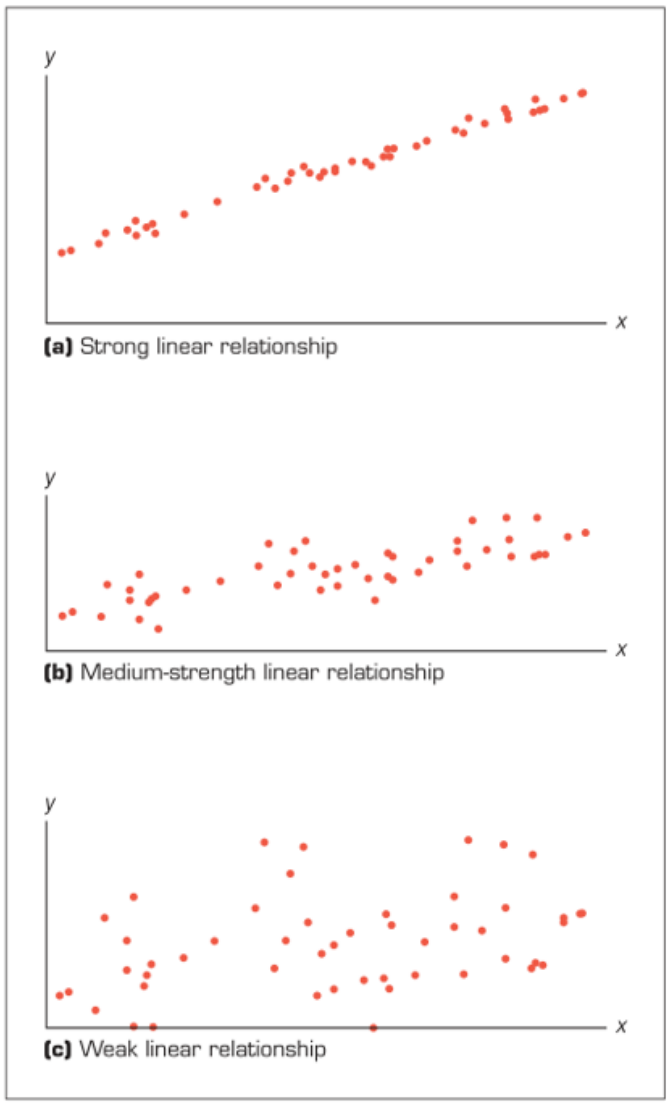
Timeseries chart

FIGURE 4.10 Line chart of Queensland's overseas exports and imports, 1989–2018



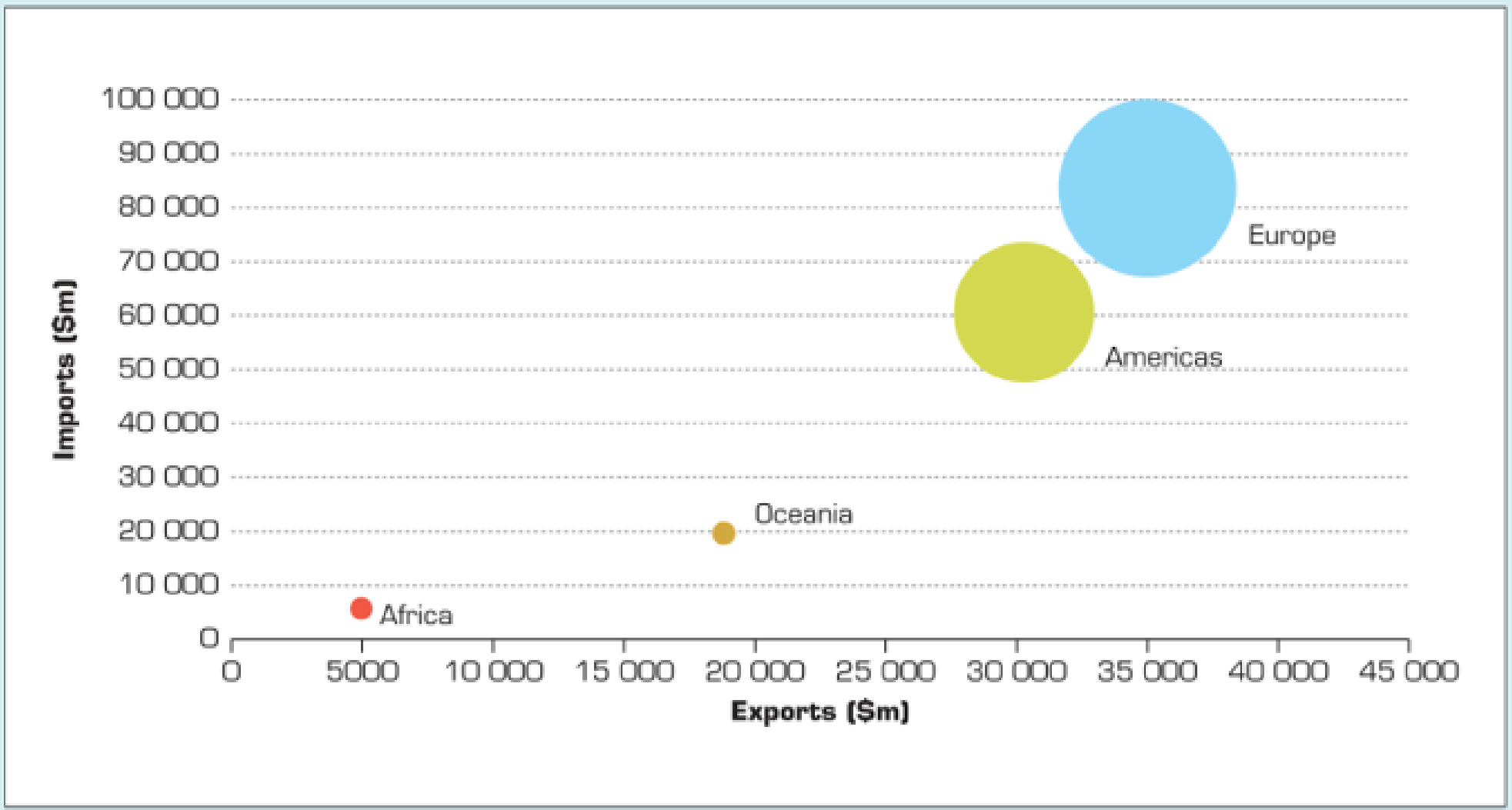
Scatterplot (đồ thị tán xạ)

FIGURE 4.13 Scatter diagrams depicting linearity



Ba biến

FIGURE 4.16 Australian merchandise trade (in $m), Four regions, 2018



Bài tập

	WITHOUT OUTLIERS	WITH OUTLIERS
DATA	4, 4, 5, 5, 5, 5, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 7, 7, 300
MEAN		
MEDIAN		
MODE		
SD		
RANGE		

Bài tập

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

Table 2.27

What is the best estimate for the mean number of hours spent playing video games?

TÓM TẮT

- Biết được các đặc trưng cần khảo sát khi khám phá dữ liệu
- Biết được các loại đồ thị cũng như các độ đo phù hợp để mô tả dữ liệu