

Chương 4 Thống kê mô tả

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

Tổng quan về thống kê

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

- 2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
- 2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
- 2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

- 3.1 Các đặc trưng đo lường khuynh hướng tập trung
- 3.2 Độ đo sự biến thiên
- 3.3 Khảo sát hình dáng phân phối của dữ liệu
- 3.4 Phân tích dữ liệu thăm dò

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
- **Thống kê suy diễn**
 - Ước lượng, kiểm định thống kê
 - Phân tích mối liên hệ
 - Dự đoán, ...

Dữ liệu và thống kê

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

- 2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
- 2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
- 2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

- 3.1 Các đặc trưng đo lường khuynh hướng tập trung
- 3.2 Độ đo sự biến thiên
- 3.3 Khảo sát hình dáng phân phối của dữ liệu
- 3.4 Phân tích dữ liệu thăm dò

- **Dữ liệu (data)** là thông tin có được từ những quan sát, những phép đếm, những đo đạc, hoặc các câu trả lời.
- **Thống kê (statistics)** là khoa học về thu thập, tổ chức, phân tích, và giải thích dữ liệu để đưa ra các quyết định.
- **Tổng thể (population)** là toàn bộ tập hợp tất cả các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hoặc định lượng nào đó.
- **Mẫu (sample)** là một tập con của một tổng thể.

Tham số và thống kê

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

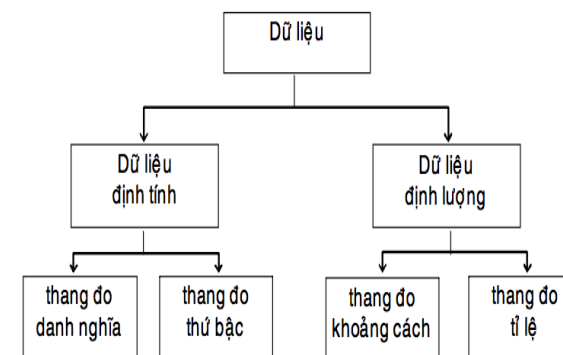
- 2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
- 2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
- 2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

- 3.1 Các đặc trưng đo lường khuynh hướng tập trung
- 3.2 Độ đo sự biến thiên
- 3.3 Khảo sát hình dáng phân phối của dữ liệu
- 3.4 Phân tích dữ liệu thăm dò

- **Tham số (parameter)** là một mô tả số về một đặc trưng của một *tổng thể*.
- **Thống kê (statistic)** là một mô tả số về một đặc trưng của một *mẫu*.

Tham số \longrightarrow Tổng thể
Thống kê \longrightarrow Mẫu



XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phân loại dữ liệu

• Dữ liệu thời điểm: dữ liệu được thu thập ở cùng hoặc xấp xỉ vào cùng một thời điểm.

• Dữ liệu chuỗi thời gian: dữ liệu thu thập được qua nhiều giai đoạn thời gian.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ: Dữ liệu thời điểm

TABLE 1.1 DATA SET FOR 25 MUTUAL FUNDS

Fund Name	Fund Type	Net Asset Value (\$)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Brown Cap Small	DE	35.73	15.85	1.20	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Overseas	IE	48.39	23.46	0.90	4-Star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-Star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-Star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.10	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Sh-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

Source: Morningstar Funds500 (2008).

Dữ liệu trong bảng 1.1 là dữ liệu thời điểm vì họ mô tả 5 biến của 25 công ty trong danh sách SP 500 tại cùng một thời điểm.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ: Dữ liệu chuỗi thời gian

FIGURE 1.1 U.S. AVERAGE PRICE PER GALLON FOR CONVENTIONAL REGULAR GASOLINE

Source: Energy Information Administration, U.S. Department of Energy, May 2009.

Dữ liệu trong Hình 1.1 cho thấy giá trung bình của mỗi gallon xăng không chì tại Mỹ từ tháng 3/2006 đến tháng 5/2009.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Các nhánh của thống kê

Nghiên cứu thống kê có hai nhánh chính: **thống kê mô tả** và **thống kê suy luận**.

• **Thống kê mô tả:** Liên quan đến việc thu thập, tổ chức, xử lý dữ liệu để biến đổi dữ liệu thành thông tin; tổng hợp dữ liệu(tính trung bình mẫu, phương sai mẫu, trung vị, ...) và và trình bày dữ liệu (dùng bảng và đồ thị).

• **Thống kê suy luận:** Liên quan đến việc sử dụng một mẫu để rút ra kết luận về một tổng thể. Suy diễn thống kê là xử lý các thông tin có được từ đó đưa ra các cơ sở cho những dự đoán , dự báo và các ước lượng, kiểm định giả thuyết thống kê.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dạng phân phối của dữ liệu

3.4 Phân tích dữ liệu thâm đo

Phân phối tần số (Frequency distribution)

Bảng tóm tắt dữ liệu thể hiện số lượng (tần số) của giá trị dữ liệu trong mỗi nhóm riêng biệt. Ví dụ:

TABLE 2.1 DATA FROM A SAMPLE OF 50 SOFT DRINK PURCHASES

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Coke Classic
Sprite	Coke Classic	
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

TABLE 2.2 FREQUENCY DISTRIBUTION OF SOFT DRINK PURCHASES

Soft Drink	Frequency
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dạng phân phối của dữ liệu

3.4 Phân tích dữ liệu thâm đo

Bảng phân phối tần suất và tần suất phần trăm của nước ngọt

TABLE 2.3 RELATIVE FREQUENCY AND PERCENT FREQUENCY DISTRIBUTIONS OF SOFT DRINK PURCHASES

Soft Drink	Relative Frequency	Percent Frequency
Coke Classic	.38	38
Diet Coke	.16	16
Dr. Pepper	.10	10
Pepsi	.26	26
Sprite	.10	10
Total	1.00	100

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dạng phân phối của dữ liệu

3.4 Phân tích dữ liệu thâm đo

Phân phối tần suất

• Phân phối tần suất (Relative frequency distribution): Bảng tóm tắt dữ liệu thể hiện tỷ lệ các giá trị dữ liệu trong mỗi nhóm riêng biệt. Với n là số quan sát trong tập dữ liệu (cỡ mẫu):

tần suất

tần số

n

• Phân phối tần suất phần trăm (Percent frequency distribution): Bảng tóm tắt dữ liệu thể hiện tỷ lệ % của các giá trị dữ liệu trong mỗi nhóm riêng biệt.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dạng phân phối của dữ liệu

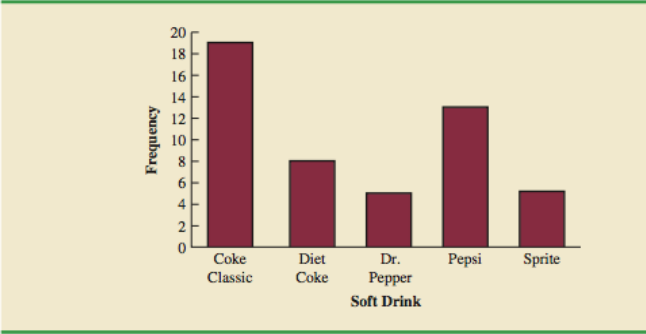
3.4 Phân tích dữ liệu thâm đo

Biểu đồ thanh (Bar graph)

Biểu đồ mô tả dữ liệu định tính đã được tóm tắt trong bảng phân phối tần số, tần suất hoặc tần suất phần trăm.

Ví dụ: Biểu đồ thanh các lần mua nước ngọt

FIGURE 2.1 BAR CHART OF SOFT DRINK PURCHASES



XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ hình tròn (Pie chart)

Biểu đồ tóm tắt dữ liệu dựa trên các phần trong một đường tròn tương ứng với các tần suất cho mỗi nhóm. Ví dụ: Biểu đồ hình tròn các loại nước ngọt được mua

FIGURE 2.2

PIE CHART OF SOFT DRINK PURCHASES

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ

Thời gian kiểm toán (ngày)

TABLE 2.4

YEAR-END AUDIT TIMES (IN DAYS)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

TABLE 2.5

FREQUENCY DISTRIBUTION FOR THE AUDIT TIME DATA

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Bảng phân phối tần số của dữ liệu trên

Lưu ý: **Trị số giữa nhóm** (class midpoint): Giá trị chính giữa giá trị nhỏ nhất và giá trị lớn nhất. Ví dụ với dữ liệu trên ta có trị số giữa của 5 nhóm trên là 12, 17, 22, 27 và 32.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phân phối tần số (Frequency distribution)

Tương tự như định nghĩa trong tóm tắt dữ liệu định tính, đó là bảng tóm tắt dữ liệu thể hiện số lượng (tần số) của các phần tử trong mỗi nhóm không chồng lấn. Tuy nhiên, ở đây cần xác định các nhóm không chồng lấn.

Các bước xác định các nhóm:

- Xác định số lượng các nhóm riêng biệt (thường từ 5 đến 20 nhóm);
- Xác định độ rộng của mỗi nhóm

$$\text{độ rộng của nhóm} = \frac{\text{Giá trị lớn nhất} - \text{Giá trị nhỏ nhất}}{\text{số nhóm}};$$

- Xác định các giới hạn của nhóm (được lựa chọn sao cho mỗi giá trị của quan sát thuộc về một và chỉ một nhóm).

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phân phối tần suất (Relative frequency distribution)

- Phân phối tần suất** (Relative frequency distribution) và **Phân phối tần suất phần trăm** (Percent frequency distribution): tương tự trong tóm tắt dữ liệu định tính.

Ví dụ: Bảng phân phối tần suất và tần suất phần trăm của dữ liệu trên

TABLE 2.6

RELATIVE FREQUENCY AND PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

Audit Time (days)	Relative Frequency	Percent Frequency
10–14	.20	20
15–19	.40	40
20–24	.25	25
25–29	.10	10
30–34	.05	5
Total	1.00	100

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

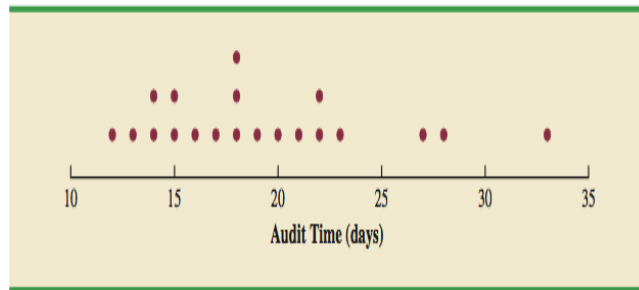
3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Đồ thị điểm (Dot plot)

Đồ thị tóm tắt dữ liệu bằng các điểm nằm trên các giá trị dữ liệu biểu diễn trên trục ngang.
Ví dụ: Đồ thị điểm cho thời gian kiểm toán

FIGURE 2.3 DOT PLOT FOR THE AUDIT TIME DATA



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

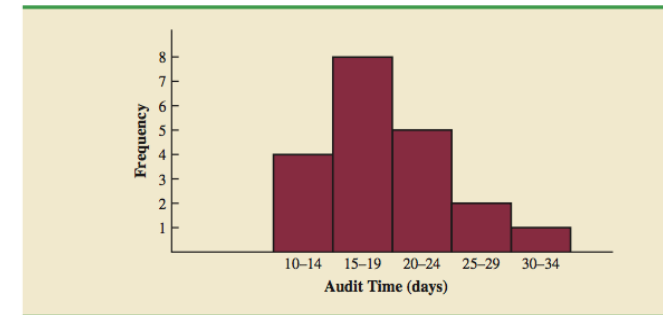
3.4 Phân tích dữ liệu thăm dò

Biểu đồ phân phối (Histogram)

Biểu đồ thể hiện phân phối tần số, phân phối tần suất hoặc phân phối tần suất phần trăm của dữ liệu định lượng xây dựng bằng cách đặt khoảng giá trị nhóm trên trục ngang và tần số, tần suất hoặc tần suất phần trăm trên trục thẳng đứng.

Ví dụ: Biểu đồ phân phối thời gian kiểm toán

FIGURE 2.4 HISTOGRAM FOR THE AUDIT TIME DATA



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

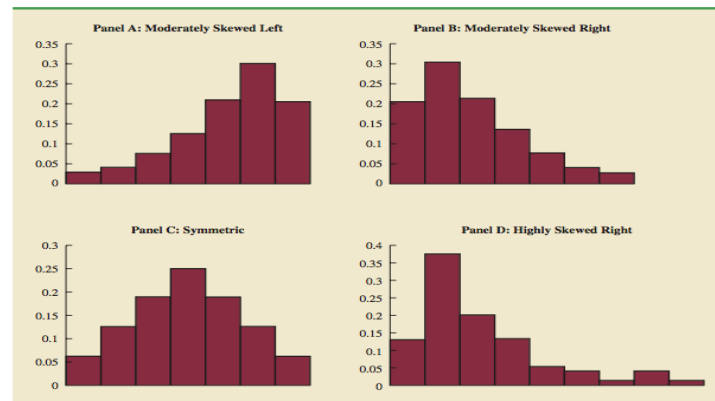
3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ phân phối (Histogram) dùng để cung cấp thông tin về hình dáng của một phân phối.

Ví dụ: Biểu đồ phân phối mô tả các hình dáng phân phối

FIGURE 2.5 HISTOGRAMS SHOWING DIFFERING LEVELS OF SKEWNESS



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Dáng điệu của phân phối

- Dáng điệu của phân phối là **đối xứng (symmetric)** nếu các giá trị quan trắc cân bằng xung quanh trung tâm.
- Dáng điệu của phân phối là **bất đối xứng (skewed)** nếu dữ liệu quan trắc không phân bố đối xứng xung quanh trung tâm.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

- **Phân phối tần số tích lũy** (Cumulative frequency distribution): cho thấy số lượng các giá trị dữ liệu ít hơn hoặc bằng giới hạn trên của mỗi nhóm.
- **Phân phối tần suất tích lũy** (Cumulative relative frequency distribution): cho thấy tỷ lệ của số các giá trị dữ liệu nhỏ hơn hoặc bằng giới hạn trên của mỗi nhóm.
- **Phân phối tần suất phần trăm tích lũy** (Cumulative percent frequency distribution): cho thấy tỷ lệ phần trăm giá trị dữ liệu nhỏ hơn hoặc bằng giới hạn trên của mỗi nhóm.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ

Phân phối tần số tích lũy, tần suất tích lũy và tần suất phần trăm tích lũy cho dữ liệu thời gian kiểm toán

TABLE 2.7

CUMULATIVE FREQUENCY, CUMULATIVE RELATIVE FREQUENCY, AND CUMULATIVE PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	.20	20
Less than or equal to 19	12	.60	60
Less than or equal to 24	17	.85	85
Less than or equal to 29	19	.95	95
Less than or equal to 34	20	1.00	100

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ khác:

Chọn ngẫu nhiên 20 ngày mùa đông có nhiệt độ cao và đo nhiệt độ (Đv: độ F) được số liệu như sau

24 35 17 21 24 37 26 46 58 30
32 13 12 38 41 43 44 27 53 27

Hãy lập bảng phân bố tần số cho số liệu này.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Giải

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần
12, 13, 17, 21, 24, 24, 26, 27, 27, 30
32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Xác định phạm vi giá trị của dữ liệu (range): $58 - 12 = 46$
- Chọn số khoảng cần chia: 5 (thông thường từ 5 đến 15)
- Xác định độ rộng của khoảng: 10 (làm tròn $46/5$)
- Xác định biên của các khoảng: từ 10 đến dưới 20, từ 20 đến dưới 30, . . . , từ 50 đến dưới 60
- Đếm số giá trị dữ liệu nằm trong mỗi khoảng

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Giải (tt)

Dữ liệu được sắp xếp theo thứ tự tăng dần:
12, 13, 17, 21, 24, 24, 26, 27, 27, 30
32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Khoảng	Tần số	Tần suất	Phần trăm
[10,20)	3	0.15	15
[20,30)	6	0.30	30
[30,40)	5	0.25	25
[40,50)	4	0.20	20
[50,60)	2	0.10	10
Tổng	20	1.00	100

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ nhánh lá (Stem-and-leaf)

Biểu đồ nhánh lá (cành lá) được sử dụng để hiển thị cả thứ tự và hình dạng của một bộ dữ liệu cùng một lúc.

Biểu đồ nhánh lá (Stem-Leaf)

- Biểu đồ stem-leaf cung cấp một cái nhìn trực quan về bộ dữ liệu x_1, x_2, \dots, x_n , với mỗi x_i gồm ít nhất hai chữ số.
- Biểu đồ stem-leaf có nhiều thuận lợi trong việc tìm các đặc trưng của dữ liệu như các phân vị, các tứ phân vị, trung vị, mode.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Giải (tt)

Khoảng	Tần số
[10, 20)	3
[20, 30)	6
[30, 40)	5
[40, 50)	4
[50, 60)	2

(Không có khoảng cách giữa các cột)

Histogram : Daily High Temperature

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Để xây dựng một biểu đồ stem-leaf, ta thực hiện theo các bước sau:

1 Sắp xếp dữ liệu theo thứ tự tăng dần

2 Chia các giá trị sắp xếp thành hai phần: phần gốc **stem**, gồm một (hoặc vài) chữ số đầu tiên, và phần lá **leaf**, gồm các chữ số còn lại.

3 Liệt kê các giá trị stem vào một cột dọc.

4 Ghi lại leaf cho mỗi quan sát vào bên cạnh stem của nó.

5 Viết các đơn vị cho các stem và leaf lên đồ thị.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ Stem-Leaf

Ví dụ

• Sắp xếp dữ liệu:
21, 24, 24, 26, 27, 27, 30, 32, 38, 41

• Hoàn thành biểu đồ stem - leaf:

Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ Stem-Leaf

Một ví dụ khác

Sử dụng đơn vị hàng trăm cho stem
(đơn vị lá = 10)

Data:

Stem	Leaves
6	1 3 6
7	2 2 5 8
8	3 4 6 6 9 9
9	1 3 3 6 8
10	3 5 6
11	4 7
12	2

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ Stem-Leaf

Một ví dụ khác

Ví dụ: Bảng 2.8 Số câu trả lời trong bài kiểm tra năng lực

TABLE 2.8

NUMBER OF QUESTIONS ANSWERED CORRECTLY ON AN APTITUDE TEST

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

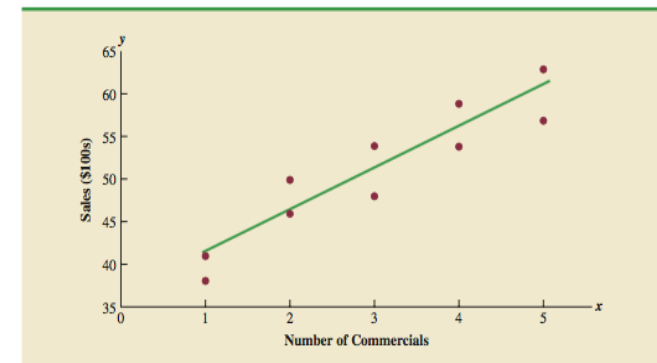
3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Với dữ liệu trong Bảng 2.8, ta có biểu đồ nhánh lá sau:

6	8	9
7	2	3 3 5 6 6
8	0	1 1 2 3 4 5 6
9	1	2 2 2 4 5 5 6 7 8 8
10	0	0 2 4 6 6 6 7 8
11	2	3 5 5 8 9 9
12	4	6 7 8
13	2	4
14		1

Những con số bên trái đường thẳng đứng (6, 7, ..., 12, 13) tạo thành nhánh, và mỗi chữ số ở bên phải đường thẳng đứng là một lá. Ví dụ: xem xét hàng đầu tiên, ta có nhánh là 6 và lá là 8 và 9. Hàng này chỉ ra rằng hai giá trị dữ liệu là 68 và 69.



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

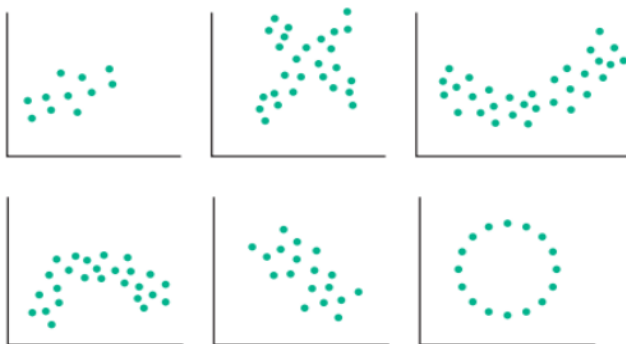
3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Đồ thị phân tán (Scatter diagram)

Đồ thị phân tán (scatter plot) được sử dụng để xác định mối liên hệ giữa hai biến X và Y .



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

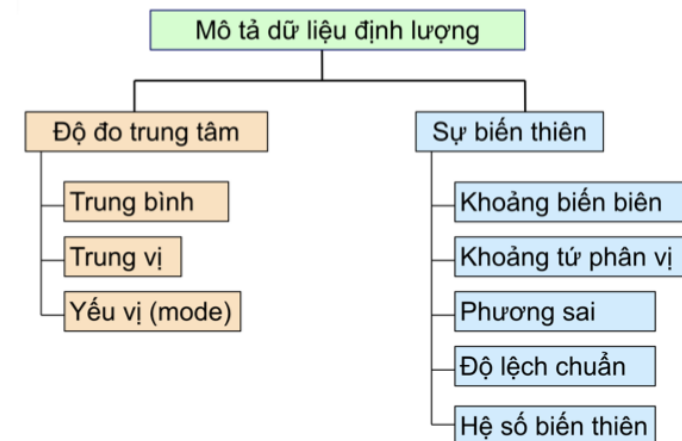
3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Mô tả dữ liệu định lượng



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

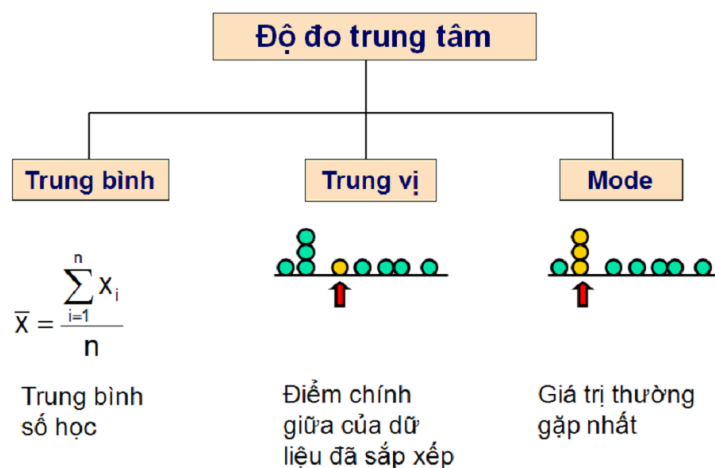
3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Các độ đo hướng tâm



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

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

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

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

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung bình (tt)

• Trung bình tổng thể:

Nếu một tổng thể có N phần tử được kí hiệu là x_1, x_2, \dots, x_N , thì **trung bình tổng thể** là

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

• Trung bình mẫu:

Nếu n quan sát của một mẫu được kí hiệu là x_1, x_2, \dots, x_n , thì **trung bình mẫu** là

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

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung bình (tt)

• Trường hợp dữ liệu có tần số như trong bảng sau

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

trong đó, $n_1 + n_2 + \dots + n_k = n$.

Khi đó, trung bình mẫu được tính theo công thức

$$\bar{x} = \frac{\sum_{i=1}^k n_i x_i}{n}$$

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung bình (tt)

Ví dụ: Bảng 3.1 Lương khởi điểm hàng tháng của 12 người tốt nghiệp ngành kinh doanh

TABLE 3.1 MONTHLY STARTING SALARIES FOR A SAMPLE OF 12 BUSINESS SCHOOL GRADUATES

Graduate	Monthly Starting Salary (\$)	Graduate	Monthly Starting Salary (\$)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

Mức lương khởi điểm hàng tháng trung bình của mẫu gồm 12 sinh viên tốt nghiệp ngành kinh doanh là :

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{3450 + 3550 + \dots + 3480}{12} = 3540$$

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung bình (tt)

Ví dụ khác : Lương tháng của 16 công nhân được chọn ngẫu nhiên (đv triệu đồng) trong một nhà máy như sau:

Lương tháng	0,8	1,0	1,2	1,3	1,5	1,7	2	2,3	2,5
Số công nhân	1	1	2	2	2	3	2	2	1

Lương trung bình hàng tháng của một công nhân từ mẫu gồm 16 công nhân trên là:

$$\bar{x} = \frac{\sum_{i=1}^k n_i x_i}{n} = \frac{\sum_{i=1}^9 n_i x_i}{16}$$

$$\bar{x} = \frac{0,8 * 1 + \dots + 2,3 * 2 + 2,5 * 1}{16} = 1,625$$

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

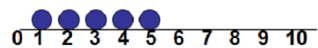
3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

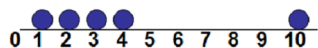
Trung bình (tt)

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

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

Ví dụ: Giả sử rằng sinh viên tốt nghiệp ở bảng 3.1 có mức lương khởi điểm cao nhất là 10000 USD/tháng không phải là 3925USD như trong bảng 3.1 thì trung bình mẫu thay đổi từ 3540 USD đến 4046 USD.

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

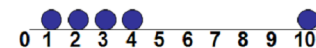
3.4 Phân tích dữ liệu thăm dò

Trung vị mẫu

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



Median = 3



Median = 3

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung vị mẫu (tt)

Cách tìm trung vị

Sắp xếp dữ liệu 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

$$i = (n + 1)/2, \text{ thì}$$

- Nếu n lẻ thì **trung vị** là giá trị thứ i hay x_i (trung vị là giá trị chính giữa);
- Nếu n chẵn thì **trung vị** là trung bình của hai giá trị thứ i và thứ $i + 1$ hay **trung vị** = $\frac{x_{[i]} + x_{[i]+1}}{2}$, với $[i]$ là phần nguyên của i (trung vị là trung bình của hai giá trị ở giữa).

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung vị (tt)

Ví dụ 1: Tìm trung vị về quy mô lớp cho mẫu của chín lớp đại học sau:

35; 34; 32; 56; 30; 54; 46; 38; 42. Giải:

- Sắp xếp dữ liệu theo thứ tự tăng dần : 30; 32; 34; 35; 38; 42; 46; 54; 56.
- $n = 9$ và $i = (n + 1)/2 = (9 + 1)/2 = 5$;
- Do $n = 9$ lẻ nên **trung vị** là giá trị thứ 5 hay $x_i = x_5 = 38$. Vậy trung vị về quy mô lớp học với mẫu trên là 38 sinh viên.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát tính đúng đắn phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung vị (tt)

Ví dụ 2: *Tìm trung vị mức lương khởi điểm của 12 sinh viên trong bảng 3.1.* Giải:

- Sắp xếp dữ liệu theo thứ tự tăng dần :
3310; 3355; 3450; 3480; 3480; 3490; 3520; 3540; 3550; 3650; 3730; 3925.
- $n = 12$ và $i = (n + 1)/2 = (12 + 1)/2 = 6.5$;
- Do $n = 12$ chẵn nên **trung vị** là trung bình của hai giá trị thứ 6 (x_6) và thứ 7 (x_7) hay
$$\text{trung vị} = \frac{x_{[6]} + x_{[6]+1}}{2} = \frac{3490 + 3520}{2} = 3505$$
. Vậy trung vị mức lương khởi điểm của 12 sinh viên tốt nghiệp đại học kinh doanh trong bảng 3.1 là 3505 USD.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát tính đúng đắn phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Trung vị (tt)

Trung vị không bị ảnh hưởng bởi các giá trị ngoại lai (outliers).

Ví dụ: Giả sử rằng sinh viên tốt nghiệp ở bảng 3.1 có mức lương khởi điểm cao nhất là 10000 USD/tháng không phải là 3925 USD/tháng như trong bảng 3.1 thì trung vị vẫn không thay đổi vì 3490 USD và 3520 USD vẫn là hai giá trị ở giữa như trên.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát tính đúng đắn phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Mode

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.

table2-7

- Mode không bị ảnh hưởng bởi các điểm ngoại lai (outlier);
- Mode có thể sử dụng cho cả dữ liệu số và dữ liệu phân loại.

Ví dụ: Dữ liệu ở bảng 3.1, mode là 3480 vì chỉ có mức lương khởi điểm hàng tháng có tần số cao nhất là 3480 USD.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát tính đúng đắn phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

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.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

So sánh trung bình, trung vị và mode (tt)

• nếu phân phối là lệch phải thì $\text{mode} < \text{trung vị} < \text{trung bình}$;

• nếu phân phối là lệch trái thì $\text{mode} > \text{trung vị} > \text{trung bình}$.

Lệch trái

Mean < Median

Đối xứng

Mean = Median

Lệch phải

Median < Mean

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phân vị

Phân vị thứ p là một giá trị mà ít nhất có $p\%$ các quan sát có giá trị nhỏ hơn hoặc bằng giá trị này và ít nhất có $(100 - p)\%$ các quan sát có giá trị lớn hơn hoặc bằng giá trị này.

Ví dụ: Các trường đại học thường báo cáo kết quả kiểm tra đầu vào dưới dạng phân vị. Giả sử, một sinh viên đạt được điểm của phần thi nói là 54 của một bài kiểm tra đầu vào.

• Làm thế nào để đánh giá sinh viên này trong mỗi liên hệ với các sinh viên khác cùng tham gia bài kiểm tra tương tự?

• TL: không dễ dàng trả lời câu hỏi này nếu không biết gì thêm về dữ liệu.

• Tuy nhiên, nếu số điểm 54 tương ứng với phân vị thứ 70, chúng ta biết rằng khoảng 70% số sinh viên đạt điểm thấp hơn so với sinh viên này và khoảng 30% số sinh viên có điểm số cao hơn sinh viên này.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phân vị (tt)

Cách tính phân vị thứ p :

• Bước 1: Sắp xếp dữ liệu theo thứ tự tăng dần.

• Bước 2: Tính chỉ số i

$$i = \left(\frac{p}{100}\right) * n$$

trong đó, p là phân vị cần tính và n là số quan sát.

• Bước 3:

• Nếu i là một số nguyên, phân vị thứ p là trung bình của hai giá trị ở vị trí thứ i và $i + 1$.

• Nếu i không phải là một số nguyên, làm tròn nó. Số nguyên tiếp theo lớn hơn i biểu thị vị trí của phân vị thứ p .

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phân vị (tt)

Ví dụ 1: Hãy xác định phân vị thứ 85 cho các dữ liệu mức lương khởi điểm trong bảng 3.1?

Giải :

• Bước 1: Sắp xếp dữ liệu theo thứ tự tăng dần:

3310; 3355; 3450; 3480; 3480; 3490; 3520; 3540; 3550; 3650; 3730; 3925.

• Bước 2: Tính

$$i = \left(\frac{p}{100}\right) * n = \left(\frac{85}{100}\right) * 12 = 10,2.$$

• Bước 3: Vì i vừa tính không phải là một số nguyên, làm tròn nó. Vị trí của phân vị thứ 85 là số nguyên kế tiếp lớn hơn 10,2 là vị trí thứ 11. Vậy, phân vị thứ 85 cho các dữ liệu mức lương khởi điểm trong bảng 3.1 là giá trị dữ liệu ở vị trí thứ 11 là 3730.

Ví dụ 2: Hãy xác định phân vị thứ 50 cho các dữ liệu mức lương khởi điểm trong bảng 3.1?

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

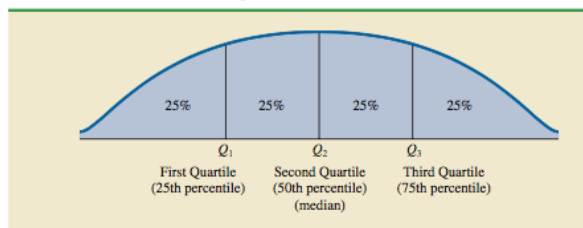
3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Tứ phân vị

Tứ phân vị chia dữ liệu thành bốn phần, mỗi phần chứa khoảng 25% số quan sát.

FIGURE 3.1 LOCATION OF THE QUARTILES



Hình 3.1 cho thấy một phân phối dữ liệu chia thành bốn phần. Các điểm chia đgl **Tứ phân vị** và được xác định như sau:

- Q_1 = tứ phân vị thứ nhất, hay là phân vị thứ 25.
- Q_2 = tứ phân vị thứ hai, hay là phân vị thứ 50 (Q_2 cũng được gọi là **trung vị**).
- Q_3 = tứ phân vị thứ ba, hay là phân vị thứ 75.

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ 2: Hãy xác định tứ phân vị cho các dữ liệu mức lương khởi điểm trong bảng 3.1?

Giải : Chúng ta cần tìm phân vị thứ 25 (Q_1), phân vị thứ 50 (Q_2) và phân vị thứ 75 (Q_3).

- Sắp xếp dữ liệu theo thứ tự tăng dần:
3310; 3355; 3450; 3480; 3480; 3490; 3520; 3540; 3550; 3650; 3730; 3925.
- Tìm Q_1 : Tính $i = \left(\frac{p}{100}\right) * n = \left(\frac{25}{100}\right) * 12 = 3$. Vì $i = 3$ là một số nguyên nên phân vị thứ 25 là trung bình của hai giá trị dữ liệu thứ ba và thứ tư hay $Q_1 = (3450 + 3480)/2 = 3465$.
- Tìm Q_2 : Tính $i = \left(\frac{50}{100}\right) * 12 = 6$. Vì $i = 6$ là một số nguyên nên trung vị là trung bình của hai giá trị dữ liệu thứ sáu và thứ bảy hay $Q_2 = (3490 + 3520)/2 = 3505$.
- Tìm Q_3 : Tính $i = \left(\frac{75}{100}\right) * 12 = 9$. Vì $i = 9$ là một số nguyên nên phân vị thứ 75 là trung bình của hai giá trị dữ liệu thứ chín và thứ mười hay $Q_3 = (3550 + 3650)/2 = 3600$.

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

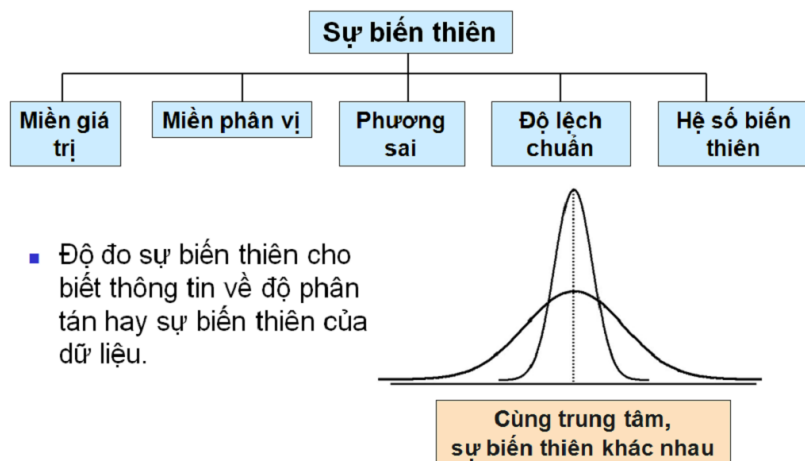
3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Độ đo sự biến thiên của dữ liệu (hay độ phân tán)



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

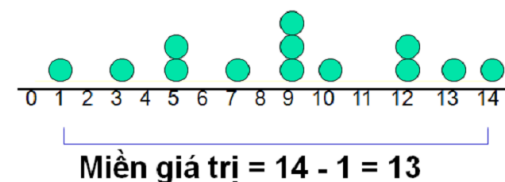
Khoảng biến thiên hay Miền giá trị mẫu (sample range)

Khoảng biến thiên = giá trị lớn nhất – giá trị nhỏ nhất. Ví dụ: Khoảng biến thiên trong bộ dữ liệu ở bảng 3.1 là $3925 - 3310 = 615$.

Hay **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, \dots, x_n thì **miền giá trị mẫu** là

$$r = \max(x_i) - \min(x_i)$$



XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Khoảng biến thiên (tt)

Khoảng biến thiên bị ảnh hưởng bởi các giá trị ngoại lai (hay giá trị đột biến).

Ví dụ: Giả sử rằng sinh viên tốt nghiệp ở bảng 3.1 có mức lương khởi điểm cao nhất là 10000 USD/tháng không phải là 3925 USD/tháng như trong bảng 3.1 thì khoảng biến thiên trong trường hợp này sẽ là $10000 - 3310 = 6690$ không phải là 615 như đã tính ở trên.

Ta thấy rõ là giá trị khoảng biến thiên lớn trong trường hợp này sẽ không mô tả tốt sự thay đổi trong bộ dữ liệu vì 11 trong 12 mức lương khởi điểm nằm trong khoảng 3310 USD và 3730 USD.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Độ trải giữa hay Khoảng tứ phân vị (interquartile range - IQR)

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

Ví dụ: Đối với mức lương khởi điểm hàng tháng trong bảng 3.1, độ trải giữa là $IQR = Q_3 - Q_1 = 3600 - 3465 = 135$.

- 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 ngoại lai (outlier).

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ

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

- Tìm các tứ phân vị của mẫu dữ liệu trên?
- Tìm khoảng tứ phân vị?

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phương sai và độ lệch chuẩn

Phương sai là trung bình bình phương độ lệch so với giá trị trung bình. Phương sai phản ánh mức độ phân tán các giá trị của các quan sát xung quanh giá trị trung bình.

- Nếu x_1, x_2, \dots, x_N là các phần tử của tổng thể và μ là trung bình tổng thể thì **phương sai tổng thể** là

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

- Độ lệch chuẩn tổng thể là $\sigma = \sqrt{\sigma^2}$.
- Nếu x_1, x_2, \dots, x_n là một mẫu có n quan sát và \bar{x} là trung bình mẫu thì **phương sai mẫu** là

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

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

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phương sai và độ lệch chuẩn (tt)

Ví dụ: Tính phương sai mẫu về dữ liệu lương khởi điểm trong bảng 3.1

TABLE 3.3

COMPUTATION OF THE SAMPLE VARIANCE FOR THE STARTING SALARY DATA

Monthly Salary (x_i)	Sample Mean (\bar{x})	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean ($(x_i - \bar{x})^2$)
3450	3540	-90	8,100
3550	3540	10	100
3650	3540	110	12,100
3480	3540	-60	3,600
3355	3540	-185	34,225
3310	3540	-230	52,900
3490	3540	-50	2,500
3730	3540	190	36,100
3540	3540	0	0
3925	3540	385	148,225
3520	3540	-20	400
3480	3540	-60	3,600
		0	301,850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Using equation (3.5),
$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301,850}{11} = 27,440.91$$

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Độ lệch tuyệt đối trung bình

Lưu ý: Đối với bất kỳ bộ dữ liệu nào, tổng các độ lệch so với giá trị trung bình sẽ luôn bằng không,

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Các độ lệch dương và các độ lệch âm bù trừ lẫn nhau, dẫn đến tổng các độ lệch so với giá trị trung bình bằng không.

Để tránh tất cả các độ lệch so với giá trị trung bình triệt tiêu lẫn nhau khi chúng ta cộng chúng lại với nhau, ta xét định nghĩa **độ lệch tuyệt đối trung bình** như sau:

$$\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}.$$

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

So sánh các độ lệch chuẩn

Dữ liệu A

Mean = 15.5
s = 3.338

Dữ liệu B

Mean = 15.5
s = 0.926

Dữ liệu C

Mean = 15.5
s = 4.570

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Hệ số biến thiên (Coefficient of Variation)

Hệ số biến thiên cho biết độ lệch chuẩn lớn bằng bao nhiêu lần so với trung bình,

$$CV = \frac{\text{Độ lệch chuẩn}}{\text{trung bình}} \times 100\%.$$

Hệ số biến thiên là một thống kê hữu ích để so sánh độ phân tán của các biến có độ lệch chuẩn khác nhau và trung bình khác nhau.

Ví dụ: Đối với bộ dữ liệu lương khởi điểm trong bảng 3.1, hệ số biến thiên là $[(165,65/3540) \times 100]\% = 4,7\%$. Hệ số biến thiên này cho chúng ta biết độ lệch chuẩn mẫu chỉ bằng 4,7% giá trị trung bình mẫu.

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ so sánh hệ số biến thiên

- Dữ liệu 1 có: trung bình $\bar{x}_1 = 50$ và độ lệch chuẩn $s_1 = 5$ nên

$$CV_1 = \frac{\bar{x}_1}{s_1} \times 100\% = \frac{5}{50} \times 100\% = 10\%.$$

- Dữ liệu 2 có: trung bình $\bar{x}_2 = 100$ và độ lệch chuẩn $s_2 = 5$ nên

$$CV_2 = \frac{\bar{x}_2}{s_2} \times 100\% = \frac{5}{100} \times 100\% = 5\%.$$

- Cả hai bộ dữ liệu có cùng độ lệch chuẩn nhưng dữ liệu 2 biến thiên ít hơn so với giá trị của nó.

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Hệ số bất đối xứng (Skewness)

Hệ số bất đối xứng (Skewness) là một đại lượng số quan trọng đo lường hình dáng của một phân phối.

Công thức tính Skewness cho dữ liệu mẫu:

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3.$$

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

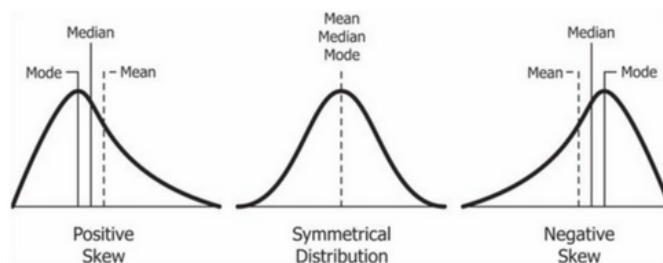
3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Hệ số bất đối xứng (Skewness) (tt)

- Khi phân phối đối xứng, **Skewness** có giá trị là 0, thì **trung bình, trung vị và mode bằng nhau**;
- Khi bộ dữ liệu có phân phối lệch phải, **Skewness** có giá trị **dương**, thì **mode < trung vị < trung bình**;
- Khi bộ dữ liệu có phân phối lệch trái, **Skewness** có giá trị **âm**, thì **mode > trung vị > trung bình**.



1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị
2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị
2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường khuynh hướng tập trung

3.2 Độ đo sự biến thiên

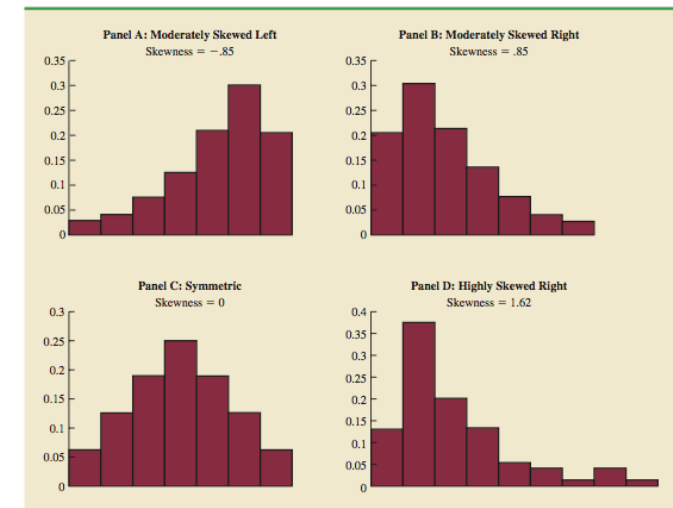
3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ:

Biểu đồ phân phối tần suất mô tả độ lệch của bốn phân phối:

FIGURE 3.3 HISTOGRAMS SHOWING THE SKEWNESS FOR FOUR DISTRIBUTIONS



XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Giá trị z (giá trị chuẩn hóa)

Giả sử chúng ta có một mẫu gồm n quan sát với các giá trị tương ứng x_1, x_2, \dots, x_n .

Giá trị z (giá trị chuẩn hóa) được tính cho mỗi x_i là:

$$z_i = \frac{x_i - \bar{x}}{s},$$

trong đó,

- z_i là giá trị z cho x_i ;
- \bar{x} là trung bình mẫu;
- s là độ lệch chuẩn mẫu.

Giá trị z cho bất kỳ quan sát nào có thể hiểu như là một thước đo vị trí tương đối của quan sát đó trong tập dữ liệu.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Quy tắc Chebyshev

Quy tắc Chebyshev Ít nhất là $(1 - 1/z^2)$ số lượng giá trị dữ liệu nằm trong khoảng z độ lệch chuẩn so với giá trị trung bình, trong đó z là giá trị bất kỳ lớn hơn 1.

Ý nghĩa của quy tắc này:

- Với $z = 2$: ít nhất 75% các giá trị dữ liệu nằm trong khoảng $z = 2$ độ lệch chuẩn so với giá trị trung bình.
- Với $z = 3$: ít nhất 89% các giá trị dữ liệu nằm trong khoảng $z = 3$ độ lệch chuẩn so với giá trị trung bình.
- Với $z = 4$: ít nhất 94% các giá trị dữ liệu nằm trong khoảng $z = 4$ độ lệch chuẩn so với giá trị trung bình.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Giá trị z (giá trị chuẩn hóa) (tt)

Ví dụ: Giá trị z của quy mô lớp học

TABLE 3.4 z-SCORES FOR THE CLASS SIZE DATA

Number of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score ($\frac{x_i - \bar{x}}{s}$)
46	2	$2/8 = .25$
54	10	$10/8 = 1.25$
42	-2	$-2/8 = -.25$
46	2	$2/8 = .25$
32	-12	$-12/8 = -1.50$

Trong bảng 3.4, ta thấy giá trị z là $-1,5$ của quan sát thứ năm cho thấy quan sát này ở xa so với trung bình; giá trị của quan sát này nhỏ hơn trung bình 1,5 lần độ lệch chuẩn.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ: Giả sử rằng các điểm kiểm tra giữ kỳ cho 100 sinh viên của khóa học thống kê có điểm trung bình là 70 và độ lệch chuẩn là 5. Có bao nhiêu sinh viên có điểm kiểm tra nằm giữa 60 và 80? Có bao nhiêu sinh viên có điểm kiểm tra nằm giữa 58 và 82?

- Chúng ta thấy 60 và 80 lần lượt ở dưới và ở trên trung bình 2 lần độ lệch chuẩn. Theo quy tắc Chebyshev, ta biết được có ít nhất 75% các quan sát có giá trị nằm trong khoảng $z = 2$ độ lệch chuẩn so với giá trị trung bình. Như vậy, ít nhất có 75% số sinh viên có điểm kiểm tra nằm giữa 60 và 80.
- Chúng ta thấy rằng $(58 - 70)/5 = -2,4$ cho thấy 58 ở dưới trung bình 2,4 lần độ lệch chuẩn và $(82 - 70)/5 = 2,4$ cho thấy 82 ở trên trung bình 2,4 lần độ lệch chuẩn. Áp dụng quy tắc Chebyshev với $z = 2,4$, ta có :

$$(1 - 1/z^2) = (1 - \frac{1}{2,4^2}) = 0,826.$$

Vậy có ít nhất 82,6% số sinh viên có điểm kiểm tra nằm giữa 58 và 82.

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Quy tắc thực nghiệm

Quy tắc Chebyshev áp dụng cho bất kỳ tập dữ liệu nào bất kể hình dáng của phân phối dữ liệu.

Quy tắc thực nghiệm áp dụng cho những tập dữ liệu được cho là xấp xỉ phân phối Gauss (hay phân phối hình chuông).

Ý nghĩa của quy tắc thực nghiệm

- Khoảng 68% của các giá trị dữ liệu sẽ nằm trong khoảng cộng và trừ 1 độ lệch chuẩn so với giá trị trung bình ($\bar{x} \pm 1s$).
- Khoảng 95% của các giá trị dữ liệu sẽ nằm trong khoảng cộng và trừ 2 độ lệch chuẩn so với giá trị trung bình ($\bar{x} \pm 2s$).
- Hầu như tất cả các giá trị dữ liệu sẽ nằm trong khoảng 3 độ lệch chuẩn so với giá trị trung bình ($\bar{x} \pm 3s$).

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Phát hiện các giá trị ngoại lai (hay bất thường)

Những giá trị quá lớn hoặc quá nhỏ trong một tập dữ liệu đgl giá trị ngoại lai (hay giá trị bất thường).

Giá trị z (giá trị chuẩn hóa) có thể được sử dụng để xác định giá trị ngoại lai (hay giá trị bất thường). Khi đó, bất kỳ giá trị dữ liệu nào với giá trị z nhỏ hơn -3 hoặc lớn hơn 3 thì được xem là giá trị ngoại lai (hay bất thường).

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Ví dụ: Hộp carton đựng nước giặt được tự động đóng gói trong một dây chuyền sản xuất. Trọng lượng sau khi đóng gói thường có phân phối hình chuông.

Nếu trọng lượng trung bình là $\bar{x} = 16$ ounce và độ lệch chuẩn là $s = 0,25$ ounce, áp dụng quy tắc thực nghiệm chúng ta có thể rút ra kết luận sau:

- Khoảng 68% các hộp có trọng lượng nằm giữa 15,75 và 16,25 ($\bar{x} \pm 1s$).
- Khoảng 95% các hộp có trọng lượng nằm giữa 15,50 và 16,50 ($\bar{x} \pm 2s$).
- Hầu như tất cả các hộp có trọng lượng nằm giữa 15,25 và 16,75 ($\bar{x} \pm 3s$).

XSTK
N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

3.4 Phân tích dữ liệu thăm dò

Ngoài kỹ thuật phân tích dữ liệu thăm dò bằng biểu đồ nhánh lá, chúng ta còn có thể dùng cách xem xét bộ tóm tắt năm trị số hoặc dùng biểu đồ hộp.

Bộ tóm tắt năm số gồm:

- Giá trị nhỏ nhất;
- Tứ phân vị thứ nhất (Q_1);
- Trung vị (Q_2);
- Tứ phân vị thứ ba (Q_3);
- Giá trị lớn nhất.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Bộ tóm tắt năm số

Cách xây dựng bộ tóm tắt năm số:

- Sắp xếp dữ liệu theo thứ tự tăng dần;
- Xác định giá trị nhỏ nhất; ba tứ phân vị (Q_1 , Q_2 và Q_3) và giá trị lớn nhất.

Ví dụ:

Bộ tóm tắt năm số cho các dữ liệu trong bảng 3.1 về mức lương khởi điểm hàng tháng là : 3310, 3465, 3505, 3600, 3925.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ hộp

Một biểu đồ hộp là một tóm tắt bằng hình vẽ của dữ liệu dự trên một bộ tóm tắt năm số.

Cách xây dựng biểu đồ hộp:

- Xác định ba tứ phân vị (Q_1 , Q_2 và Q_3) và độ trải giữa ($IQR = Q_3 - Q_1$);
- Xác định các giá trị bất thường: điểm ngoại lai (outlier) và cực ngoại lai (extreme outlier) (nếu có) và giá trị nhỏ nhất, giá trị lớn nhất.
- 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.
- Vẽ một đoạn thẳng (bằng đường đứt nét) từ Q_1 đến giá trị dữ liệu nhỏ nhất nhưng lớn hơn $Q_1 - 1,5IQR$. Vẽ một đoạn thẳng (bằng đường đứt nét) từ Q_3 đến giá trị dữ liệu lớn nhất nhưng nhỏ hơn $Q_3 + 1,5IQR$.
- Đánh dấu các điểm outlier và extreme outlier.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ hộp (tt)

Ví dụ: Biểu đồ hộp cho các dữ liệu trong bảng 3.1 về mức lương khởi điểm hàng tháng được xây dựng như sau:

- Một hộp được vẽ với các cạnh của hộp nằm ở $Q_1 = 3465$ và $Q_3 = 3600$;
- Đường thẳng được vẽ trong hộp ở vị trí trung vị $Q_2 = 3505$;
- Với $IQR = Q_3 - Q_1 = 135$, ta có các giới hạn cho biểu đồ hộp là $Q_1 - 1,5IQR = 3465 - 1,5 * 135 = 3262,5$ và $Q_3 + 1,5IQR = 3600 + 1,5 * 135 = 3802,5$. Dữ liệu nằm ngoài các giới hạn này được xem là các giá trị bất thường.
- Vẽ các râu bằng đường đứt nét từ các cạnh của hộp đến giá trị tiền lương 3310 và 3730.
- Dùng biểu tượng dấu sao * để đánh dấu điểm ngoại lai 3925.

XSTK

N.T. M. Ngọc

1. Một số khái niệm thường dùng trong thống kê

2. Thống kê mô tả: trình bày dữ liệu

2.1 Tóm tắt dữ liệu định tính: phương pháp bảng và đồ thị

2.2 Tóm tắt dữ liệu định lượng: phương pháp bảng và đồ thị

2.3 Phương pháp nhánh lá

3. Mô tả dữ liệu định lượng

3.1 Các đặc trưng đo lường xu hướng tập trung

3.2 Độ đo sự biến thiên

3.3 Khảo sát hình dáng phân phối của dữ liệu

3.4 Phân tích dữ liệu thăm dò

Biểu đồ hộp (tt)

Biểu đồ hộp về mức lương khởi điểm hàng tháng

FIGURE 3.5 BOX PLOT OF THE STARTING SALARY DATA WITH LINES SHOWING THE LOWER AND UPPER LIMITS