THỐNG KÊ MÔ TẢ

Nguyễn Văn Thìn

BỘ MÔN THỐNG KÊ TOÁN HỌC
KHOA TOÁN - TIN HỌC
ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP.HCM

Tháng 2 năm 2016

Outline

Outline

1 Giới thiệu về thống kê

2 Mô tả dữ liệu một biến bằng phương pháp đồ thị
- Dữ liệu của biến định tính
- Dữ liệu của biến định lượng

3 Mô tả dữ liệu một biến bằng phương pháp số
- Các độ đo hướng tâm
- Các độ đo sự biến thiên của dữ liệu

4 Mô tả dữ liệu nhiều biến

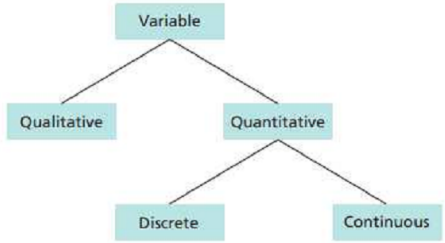
Giới thiệu về **thống kê**

Biến và dữ liệu

- **Biến (variable)**: một đặc trưng mà thay đổi từ người (vật, hiện tượng) này sang người (vật, hiện tượng) khác. Biến gồm hai loại: biến định tính (qualitative variable) và biến định lượng (quantitative variable).
- **Biến định tính**: biểu diễn tính chất của đặc trưng mà nó thể hiện, có tác dụng phân loại; ví dụ: nhóm máu (A, B, AB, O), giới tính (nam, nữ). màu mắt (đen, nâu, xanh),. . .
- **Biến định lượng**: biểu diễn độ lớn của đặc trưng mà nó thể hiện; ví dụ: chiều cao, cân nặng, thời gian,. . .
- Biến định lượng bao gồm biến rời rạc (discrete variable) và biến liên tục (continuous variable).

Biến và dữ liệu

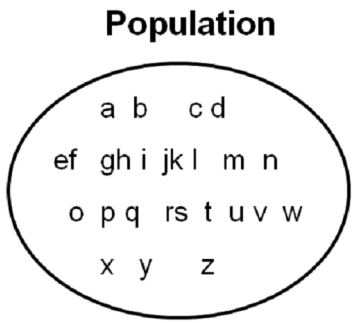
- Thông thường biến rời rạc liên quan đến bài toán đếm số các phần tử của một tổng thể; ví dụ: số sản phẩm hỏng trong 1 lô hàng, số con trong 1 gia đình, số cuộc điện thoại đến tổng đài trong 1 giờ, . . . trong khi biến liên tục liên quan đến sự đo đạc; ví dụ: cân nặng của 1 sản phẩm, chiều cao của 1 cây, cường độ dòng điện, nhiệt độ,. . .
- **Dữ liệu (data):** các giá trị của một biến. Tập hợp tất cả những quan trắc cho một biến cụ thể được gọi là một tập dữ liệu (data set).



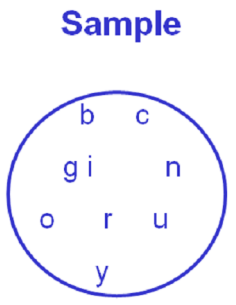
Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- **Mẫu (sample):** là một tập con được chọn ra từ tổng thể. Ta thường kí hiệu N để chỉ số phần tử của tổng thể và n để chỉ cỡ mẫu.
- **Tham số (parameter):** là một đặc trưng cụ thể của một tổng thể.
- **Thống kê (statistic):** là một đặc trưng cụ thể của một mẫu.

Tổng thể và mẫu



Những giá trị tính từ dữ liệu của tổng thể gọi là các tham số



Những giá trị được tính từ dữ liệu của mẫu gọi là các thống kê

Ví dụ về tổng thể

- Số cử tri đăng kí đi bầu cử.
- Thu nhập của các hộ gia đình trong thành phố.
- Điểm trung bình của tất cả các sinh viên trong một trường đại học.
- Trọng lượng của các sản phẩm trong một nhà máy.

Thông thường, ta không thể chọn hết được tất cả các phần tử của tổng thể để nghiên cứu bởi vì:

- Số phần tử của tổng thể rất lớn.
- Thời gian và kinh phí không cho phép.
- Có thể làm hư hại các phần tử của tổng thể.

Do đó, ta chỉ thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

Chọn mẫu ngẫu nhiên

Mẫu ngẫu nhiên hoàn hảo rất khó đạt được trong thực tế. Nếu tổng thể có kích thước N nhỏ, ta có thể viết N số lên các phiếu nhỏ, trộn đều các phiếu và chọn một mẫu gồm n phiếu. Các số mà ta chọn tương ứng với n số đo xuất hiện trong mẫu.

Bởi vì phương pháp này không thực tế cho lắm, phương pháp đáng tin cậy và đơn giản hơn là sử dụng **các số ngẫu nhiên**– các số được sinh ra sao cho các giá trị 0 đến 9 xuất hiện ngẫu nhiên và với tần số bằng nhau. Các số này có thể được sinh ra bằng máy tính hoặc có sẵn trên máy tính bỏ túi.

Một cách khác, ta có thể dùng bảng các số ngẫu nhiên để chọn một *mẫu ngẫu nhiên*.

Chọn mẫu ngẫu nhiên

Ví dụ 1

Giả sử ta muốn chọn một mẫu kích thước $n = 2$ từ một tổng thể chứa $N = 4$ đối tượng. Nếu 4 đối tượng được xác định bởi các kí hiệu x_1, x_2, x_3 và x_4, có 6 cặp khác nhau có thể được chọn là $(x_1, x_2), (x_1, x_3), (x_1, x_4), (x_2, x_3), (x_2, x_4), (x_3, x_4)$. Nếu mẫu 2 quan sát được chọn sau cho mỗi trong 6 mẫu này có cùng khả năng được chọn, bằng 1/6, thì mẫu kết quả được gọi là **mẫu ngẫu nhiên đơn giản**, hoặc ngắn gọn là **mẫu ngẫu nhiên**.

Định nghĩa 2

Nếu một mẫu gồm n phần tử được chọn từ một tổng thể có N phần tử bằng cách sử dụng một cách lấy mẫu sao cho mỗi mẫu bất kỳ đều có cùng khả năng được chọn như nhau, thì mẫu này được gọi là **ngẫu nhiên** và mẫu kết quả là **mẫu ngẫu nhiên đơn giản**.

Ví dụ về chọn mẫu ngẫu nhiên đơn giản

Ví dụ 3

Chọn một mẫu gồm $n = 10$ phần tử từ tập hợp có 200 phần tử. Sử dụng chương trình thống kê $\mathbb{R}$: dùng lệnh sample

- Đánh số từ 1 đến 200:

```
P <- 1:200
```

- Chọn mẫu lần thứ nhất:

```
S1 <- sample(P, 10, rep = TRUE)
S1
```

- Chọn mẫu lần thứ hai:

```
S2 <- sample(P, 10, rep = TRUE)
S2
```

Thống kê mô tả

Thống kê mô tả (Descriptive statistics): là quá trình thu thập, tổng hợp và xử lí dữ liệu để biến đổi dữ liệu thành thông tin.

- Thu thập dữ liệu: khảo sát, đo đạc,...
- Biểu diễn dữ liệu: dùng bảng và đồ thị,
- Tổng hợp dữ liệu: tính các thống kê mẫu như trung bình mẫu, phương sai mẫu, trung vị,...

Thống kê suy luận

- Suy luận là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- **Thống kê suy luận (Inferential statistics):** xử lí các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở để dự đoán (predictions), dự báo (forecasts) và ước lượng (estimations).
- Một số ví dụ về thống kê suy luận:
  - Ước lượng tỉ lệ sản phẩm kém chất lượng trong 1 nhà máy; ước lượng trọng lượng trung bình sử dụng trung bình mẫu.
  - Kiểm định giả thuyết cho rằng trọng lượng trung bình của 1 sản phẩm là 20 kg.

Outline

Mô tả dữ liệu một biến bằng phương pháp đồ thị

Dữ liệu của biến định tính

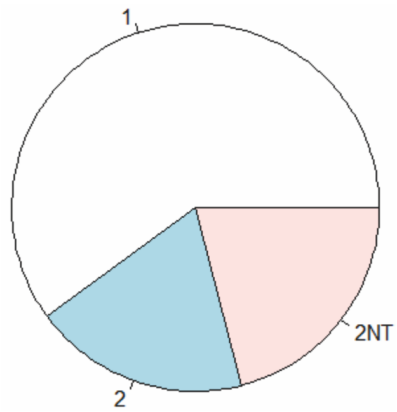
Biểu đồ hình quạt

Dữ liệu được mô tả bằng một hình tròn và mỗi một lớp được mô tả bằng một phần của hình tròn (hình quạt). Độ lớn phần hình quạt mô tả một lớp tương ứng với phần trăm số liệu của lớp đó so với tổng thể.

Ví dụ 4

Với dữ liệu của biến định tính là biến khu vực (KV trong [5]), ta có bảng tần số

Khu vực	Tần số	Tỷ lệ (%)
1	60	60%
2	19	19%
2NT	21	21%



Nhận xét 5

- Biểu đồ hình quạt dùng để hiển thị tỷ lệ các lớp của dữ liệu,
- Để đồ thị được sáng sủa, người ta có thể chia các lớp trên hình tròn theo thứ tự phần trăm tăng hay giảm dần.
- Ta có thể tạo ra dữ liệu của một biến *định tính* từ dữ liệu của một biến *định lượng* bằng cách phân loại dữ liệu định lượng thành các lớp sao cho mỗi một dữ liệu được đưa vào đúng một lớp. Xem ví dụ sau,

Ví dụ 6

Liên quan đến việc đánh giá điểm số học sinh, người ta có thể phân lớp các điểm số thành các hạng. Chẳng hạn, với phân lớp

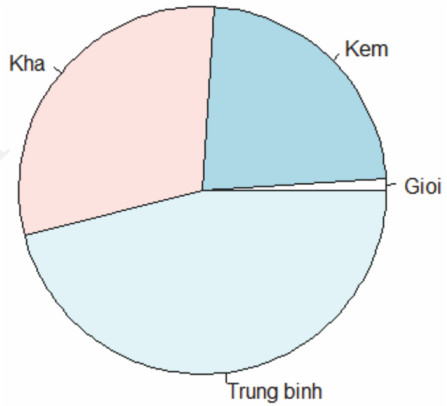
Đánh giá	Điểm số
Kém	nhỏ hơn 5
Trung bình	từ 5 đến cận 7
Khá	từ 7 đến cận 8
Giỏi	từ 8 trở lên

mỗi một số liệu (điểm số học sinh) được đưa vào đúng một lớp. Như vậy, ta đã có dữ liệu của biến định tính mới là biến "Đánh giá" và biểu đồ hình quạt lúc này đã có thể dùng cho dữ liệu này.

Ví dụ 7

Với dữ liệu về điểm số môn Toán học kỳ 1 (T1, xem [5]), ta thành lập được bảng tần số sau

Đánh giá	Tần số	Tỷ lệ (%)
Kém	23	23%
Trung bình	46	46%
Khá	30	30%
Giỏi	1	1%

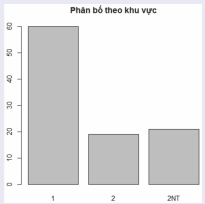


Biểu đồ hình thanh (biểu đồ cột)

Gán nhãn số liệu cho một trục và gán nhãn các lớp cho trục còn lại; vẽ một hình chữ nhật trên nhãn mỗi lớp với chiều dài tương ứng với tần số của nó; các hình chữ nhật này có cùng chiều rộng và chừa khoảng trắng giữa các hình chữ nhật nhằm làm rõ sự khác biệt giữa các lớp.

Ví dụ 8

Biểu đồ hình thanh sau biểu diễn số liệu học sinh phân lớp theo loại khu vực (biến KV trong [5])



Dữ liệu của biến định lượng

Biểu đồ cành lá (Stem-Leaf)

- Biểu đồ stem-leaf cung cấp một cái nhìn trực quan về bộ dữ liệu $x_1, x_2, \ldots, x_n$, 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.
- 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.

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

Biểu đồ Stem-Leaf
Một ví dụ khác

Sử dụng đơn vị hàng trăm cho stem

Data:
613, 632, 658, 717,
722, 750, 776, 827,
841, 859, 863, 891,
894, 906, 928, 933,
955, 982, 1034,
1047, 1056, 1140,
1169, 1224

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

Biểu đồ Stem-Leaf
Một ví dụ khác

Ví dụ 9

Vẽ đồ thị stem - leaf cho tập dữ liệu sau:

61	63	70	71	71	81	83	84	64	65
65	66	84	87	73	75	92	93	77	78
78	88	88	95	79					

Biểu đồ Stem-Leaf
Chọn stem phù hợp là điều quan trọng

Ví dụ 10

Biểu đồ stem-leaf cho 25 quan sát về các sản lượng từ một quá trình hóa học.

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

(a)

Stem	Leaf
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

(b)

Stem	Leaf
6z	1
6t	3
6f	4 5 5
6s	6
6e	
7z	0 1 1
7t	3
7f	5
7s	7
7e	8 8 9
8z	1
8t	3
8f	4 4
8s	7
8e	8 8
9z	
9t	2 3
9f	5
9s	
9e	

(c)

Biểu đồ Stem-Leaf

Chọn stem phù hợp là điều quan trọng

Nhận xét 11

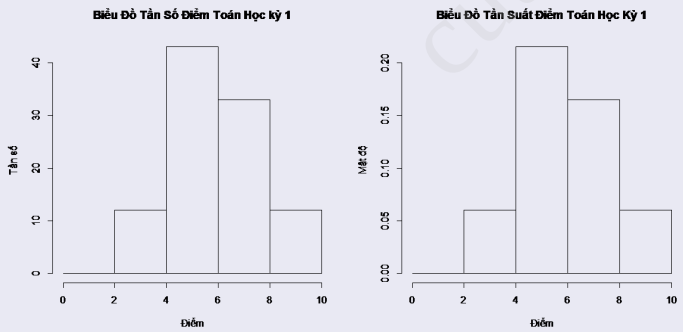
- Trong hình (a) ta sử dụng 6, 7, 8, và 9 là các stem. Điều này khiến cho có quá ít stem, và đồ thị stem-leaf không cung cấp nhiều thông tin về dữ liệu.
- Trong hình (b) ta chia mỗi stem thành hai phần và đồ thị mô tả dữ liệu tốt hơn.
- Trong hình (c) mỗi stem được chia thành 5 phần. Có quá nhiều stem trong đồ thị này, điều này khiến đồ thị không nói cho ta nhiều về hình dạng của đồ thị.

Biểu đồ tần số và biểu đồ tần suất (histogram)

- Dữ liệu định lượng được phân thành lớp bằng cách chia khoảng dữ liệu (khoảng xác định từ số liệu nhỏ nhất đến số liệu lớn nhất) thành một số các khoảng con, thường từ 5 đến 20 khoảng con. Từ đó ta thành lập được bảng tần số hay tần suất.
- Sau khi dữ liệu được phân lớp bằng bảng tần số hay tần suất, ta xây dựng biểu đồ tần số hay tần suất bằng cách gán nhãn trục hoành dữ liệu định lượng và trục tung cho tần số hay mật độ và vẽ các hình chữ nhật trên từng lớp trên các khoảng con xác định lớp đó với chiều cao chính là tần số hay mật độ của lớp đó. Trong đó, mật độ của lớp được tính bằng tần suất của lớp chia cho độ rộng của lớp đó.

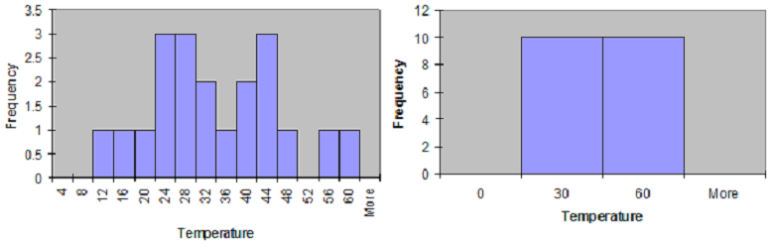
Ví dụ 12

Dữ liệu điểm Toán học kỳ 1 (T1 trong [5])



Chia dữ liệu thành bao nhiêu khoảng là tốt?

- Là quá trình "thử" và "sai",
- Mục tiêu là tạo được 1 phân phối không quá "lởm chởm" (jagged), có nhiều đỉnh và không có dạng "khối" (blocky),
- Mục tiêu là chỉ ra được sự biến thiên trong dữ liệu.
- Trong hầu hết mọi trường hợp, người ta thường chọn số khoảng từ 5-20. Trong thực tế, số các khoảng có thể lấy xấp xỉ là căn bậc hai của số quan sát.



Ví dụ 13

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 tần số và vẽ biểu đồ tần số cho số liệu này.

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 miền dữ liệu (range): $58 - 12 = 46$
- Chọn số khoảng cần chia: 5
- 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.

Dữ liệu được sắp tăng:

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

Bảng phân phối tần số:

Khoảng	Tần số	Tần số quan hệ	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

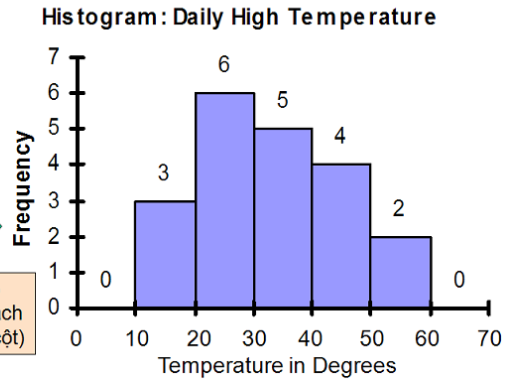
Đồ thị tổ chức tần số (histogram) là một hình ảnh hiển thị của phân phối tần số. Các bước để xây dựng một đồ thị tần số như sau:

Xây dựng đồ thị tổ chức tần số

1. Đánh nhãn các khoảng trên trục hoành
2. Đánh nhãn trục tung bằng tần số hoặc tần suất
3. Trên mỗi khoảng, vẽ một hình chữ nhật với chiều cao bằng với tần số (hoặc tần suất) tương ứng với khoảng đó.

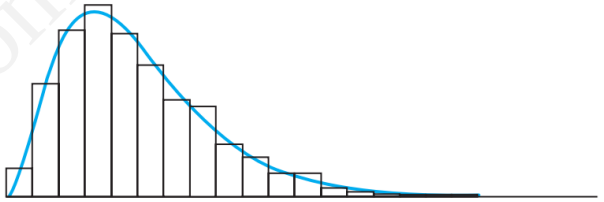
Khoảng	Tần số
[10, 20)	3
[20, 30)	6
[30, 40)	5
[40, 50)	4
[50, 60)	2

(Không có khoảng cách giữa các cột)



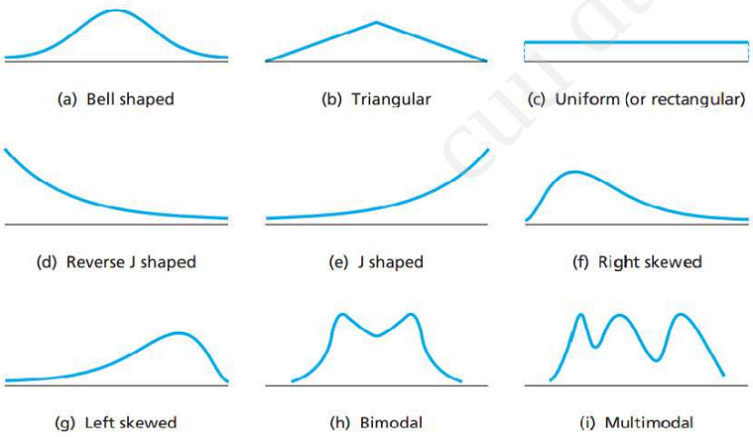
Hình dạng tổng thể có thể được nhận biết từ histogram

Khi kích thước mẫu lớn, đồ thị tổ chức tần số phản ánh hình dạng của phân phối tổng thể. Hình dạng của phân phối có thể được xác định bởi một đường cong trơn xấp xỉ đồ thị tổ chức tần số như trong hình sau.



Dưới đây là một số hình dạng phân phối thường gặp.

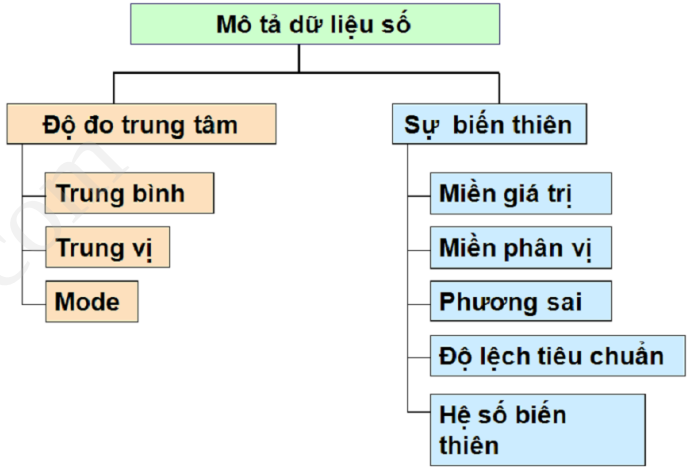
Một số hình dạng của phân phối tổng thể



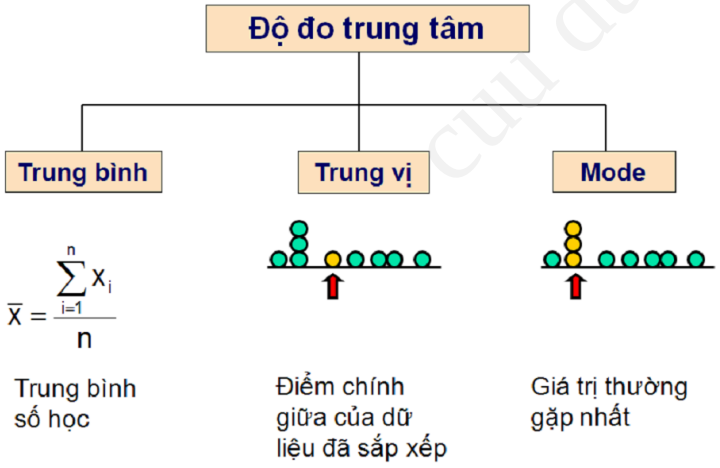
Outline

Mô tả dữ liệu **một biến** bằng phương pháp **số**

Giới thiệu



Các độ đo hướng tâm



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).

Định nghĩa 14

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

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

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

Nhận xét 15

Trường hợp dữ liệu có tần số như trong bảng sau

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

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

$$\frac{\sum_{i=1}^{k} n_i x_i}{n} \qquad (2)$$

Nhận xét 16

Khi dữ liệu được trình bày dưới dạng khoảng như sau

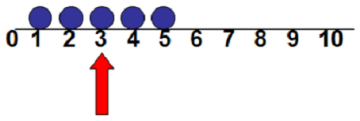
Giá trị dữ liệu	$< a_1$	$[a_1, b_1[$	...	$[a_k, b_k[$	$\geq b_k$
Tần số tương ứng	n_1	n_2	...	n_{k+1}	n_{k+2}

Bảng 1: Dữ liệu dưới dạng khoảng

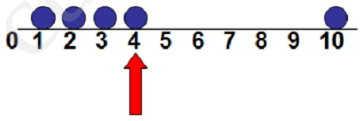
Giả sử rằng độ rộng các khoảng là như nhau, tức là $b_i - a_i = c$ với mọi i. Khi đó, mỗi khoảng ta thay bằng điểm chính giữa của khoảng, riêng hai khoảng đầu và cuối ta thay bằng $a_1 - c/2$ và $b_k + c/2$. Sau đó, dùng công thức (2) để tính trung bình.

Trung bình

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



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



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

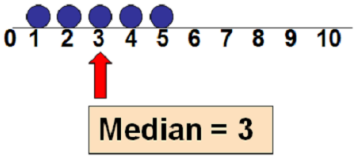
Trung vị mẫu

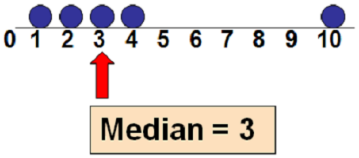
Định nghĩa 17

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

Nhận xét 18

Trung vị không bị ảnh hưởng bởi các điểm outlier.





Trung vị mẫu

Cách tìm trung vị

Sắp xếp mẫu theo thứ tự tăng dần.

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

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

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

Đối với dữ liệu dạng khoảng (xem bảng 1)

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

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

Mode

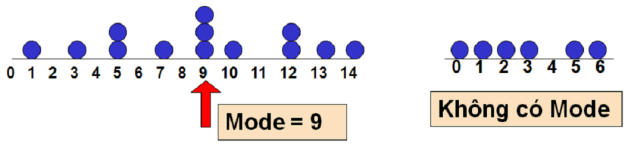
Định nghĩa 19

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

Nhận xét 20

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

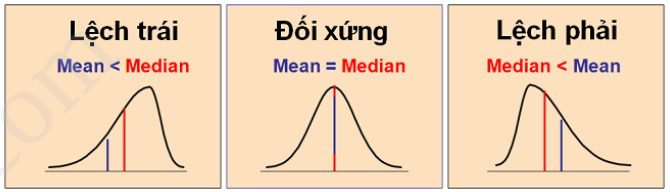
Mode



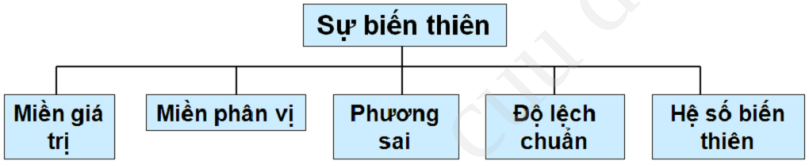
So sánh trung bình, trung vị và mode

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

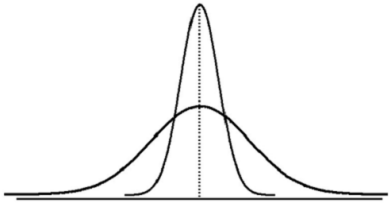
So sánh trung bình, trung vị và mode



Độ đo sự biến thiên của dữ liệu



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



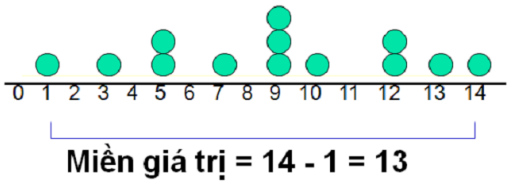
Miền giá trị mẫu (sample range)

Định nghĩa 21

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

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

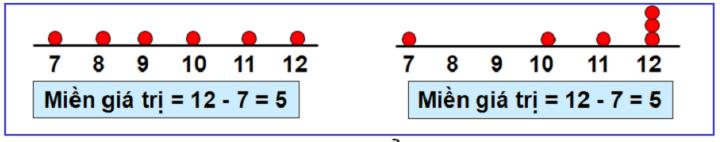
$$r = \max(x_i) - \min(x_i) \tag{3}$$



Miền giá trị mẫu

Nhược điểm

- Bỏ qua phân bố của dữ liệu



- Bị ảnh hưởng bởi các điểm oulier

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Miền giá trị = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Miền giá trị = 120 - 1 = 119

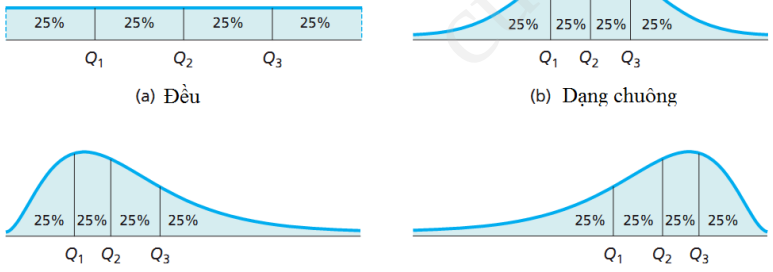
Tứ phân vị

Định nghĩa 22

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

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

Các tứ phân vị cho một số phân phối



Tứ phân vị

Cách tìm tứ phân vị

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

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

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

Khi đó,

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

Khoảng tứ phân vị (interquartile range - IQR)

Định nghĩa 23

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

Nhận xét 24

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

Ví dụ

Ví dụ 25

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

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

(a) Tìm các tứ phân vị của dữ liệu trên?
(b) Tìm khoảng tứ phân vị?

Dữ liệu outlier

Định nghĩa 26

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

Nguyên nhân xuất hiện dữ liệu outlier

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

Dữ liệu outlier

Nhận xét 27

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

Ví dụ 28

Xét dữ liệu về thời gian xem phim hàng tuần trong Ví dụ 25. Xác định các giá trị outlier (nếu có)?

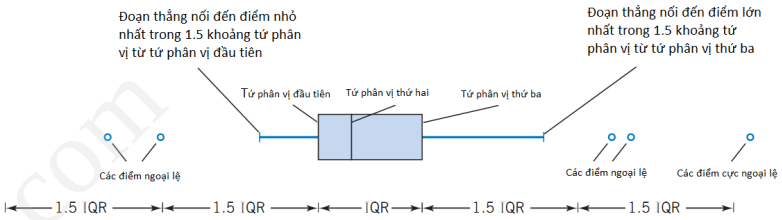
Đồ thị dạng hộp (boxplot)

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

Xây dựng đồ thị dạng hộp

B1. Xác định Q_1, Q_2, Q_3 và $IQR = Q_3 - Q_1$
B2. Xác định các điểm outlier và extreme outlier (nếu có)
B3. Vẽ một trục tọa độ ngang (hoặc dọc), và vẽ các đoạn thẳng tại Q_1, Q_2 và Q_3. Đóng khung các đoạn thẳng này trong một hộp.
B4. Vẽ một đoạn thẳng từ Q_1 đến giá trị dữ liệu nhỏ nhất nhưng lớn hơn $Q_1 - 1.5IQR$. Vẽ một đoạn thẳng từ Q_3 đến giá trị dữ liệu lớn nhất nhưng nhỏ hơn $Q_3 + 1.5IQR$.
B5. Đánh dấu các điểm outlier và extreme outlier.

Đồ thị dạng hộp (boxplot)



Chú ý

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

Đồ thị dạng hộp (boxplot)

Ví dụ 29

Vẽ đồ thị dạng hộp cho dữ liệu thời gian xem ti vi hàng tuần trong Ví dụ 25.

Giải

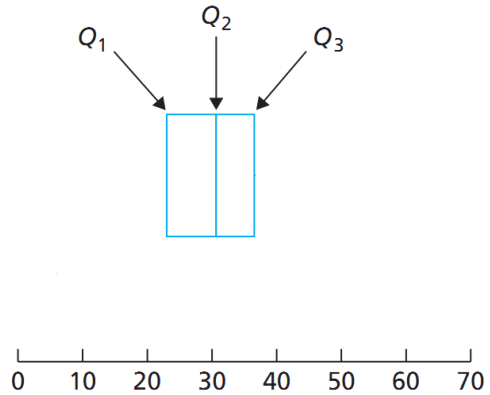
Sắp xếp dữ liệu theo thứ tự tăng dần

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

B1. **Xác định Q_1, Q_2, Q_3 và $IQR = Q_3 - Q_1$.**
$Q_1 = 23, Q_2 = 30.5, Q_3 = 36.5$, và $IQR = 13.5$
B2. **Xác định các điểm outlier và extreme outlier (nếu có)**
66

Đồ thị dạng hộp (boxplot)
Ví dụ 29 (tt)

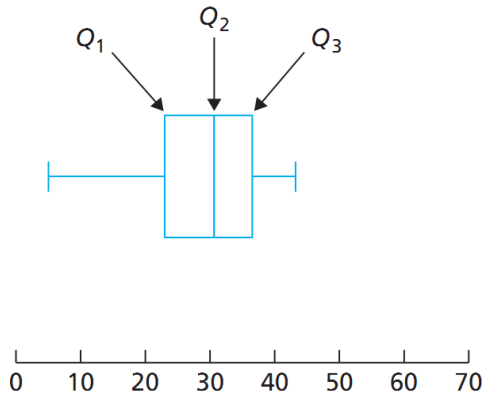
B3. **Vẽ một trục tọa độ ngang (hoặc dọc), và vẽ các đoạn thẳng tại Q_1, Q_2 và Q_3. Đóng khung các đoạn thẳng này trong một hộp.**



Đồ thị dạng hộp (boxplot)

Ví dụ 29 (tt)

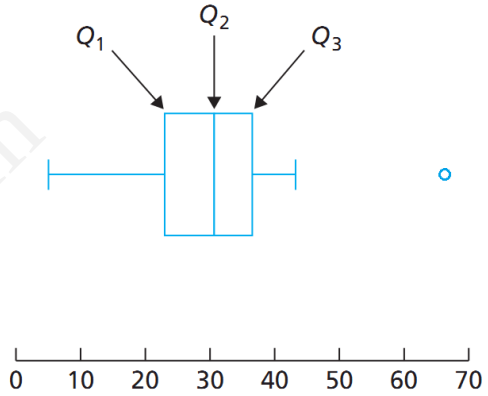
B4. **Vẽ một đoạn thẳng từ Q_1 đến giá trị dữ liệu nhỏ nhất nhưng lớn hơn $Q_1 - 1.5IQR$. Vẽ một đoạn thẳng từ Q_3 đến giá trị dữ liệu lớn nhất nhưng nhỏ hơn $Q_3 + 1.5IQR$.**



Đồ thị dạng hộp (boxplot)

Ví dụ 29 (tt)

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



Đồ thị dạng hộp (boxplot)

Nhận xét 30

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

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

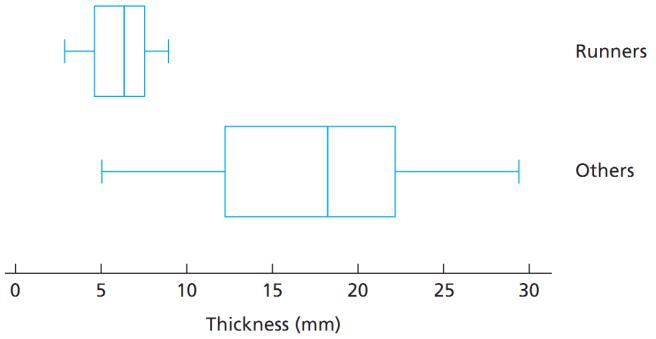
Bảng 2: Độ dày nếp gấp da

Đồ thị dạng hộp (boxplot)

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

Ví dụ 31 (Độ dày nếp gấp da (skinfold thickness))

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



Nhận xét 32

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

Đồ thị dạng hộp (boxplot)

Hình dạng của phân phối

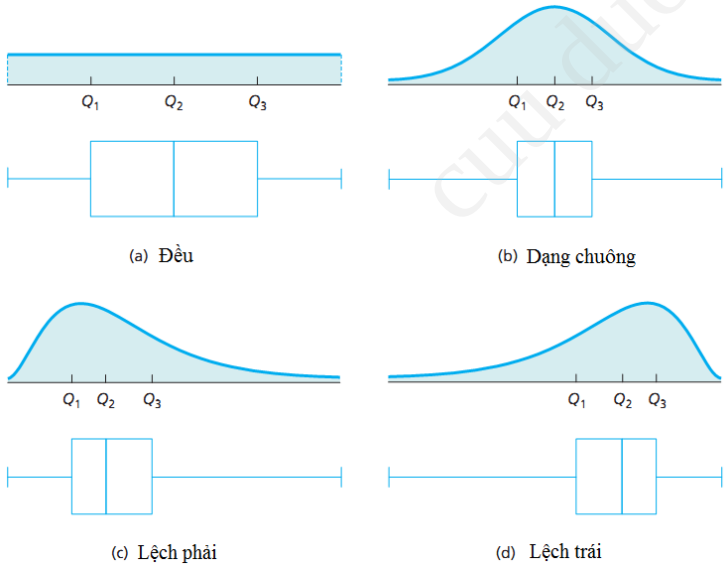
THỐNG KÊ MÔ TẢ
Nguyễn Văn Thìn
Giới thiệu về thống kê
Mô tả dữ liệu một biến bằng phương pháp đồ thị
Dữ liệu của biến định tính
Dữ liệu của biến định lượng
Mô tả dữ liệu một biến bằng phương pháp số
Các độ đo hướng tâm
Các độ đo sự biến thiên của dữ liệu
Mô tả dữ liệu nhiều biến

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

Đồ thị dạng hộp (boxplot)

Hình dạng của phân phối

THỐNG KÊ MÔ TẢ
Nguyễn Văn Thìn
Giới thiệu về thống kê
Mô tả dữ liệu một biến bằng phương pháp đồ thị
Dữ liệu của biến định tính
Dữ liệu của biến định lượng
Mô tả dữ liệu một biến bằng phương pháp số
Các độ đo hướng tâm
Các độ đo sự biến thiên của dữ liệu
Mô tả dữ liệu nhiều biến



Phương sai

THỐNG KÊ MÔ TẢ
Nguyễn Văn Thìn
Giới thiệu về thống kê
Mô tả dữ liệu một biến bằng phương pháp đồ thị
Dữ liệu của biến định tính
Dữ liệu của biến định lượng
Mô tả dữ liệu một biến bằng phương pháp số
Các độ đo hướng tâm
Các độ đo sự biến thiên của dữ liệu
Mô tả dữ liệu nhiều biến

Định nghĩa 33

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

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

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

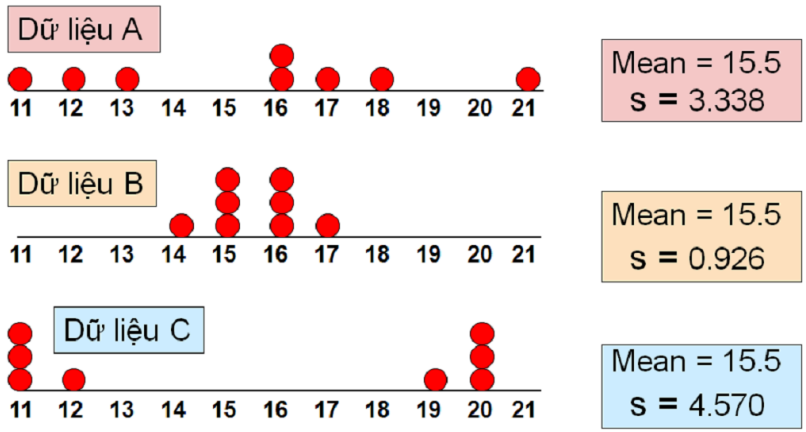
Định nghĩa 34

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

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

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

So sánh các độ lệch chuẩn



Mean = 15.5
s = 3.338

Mean = 15.5
s = 0.926

Mean = 15.5
s = 4.570

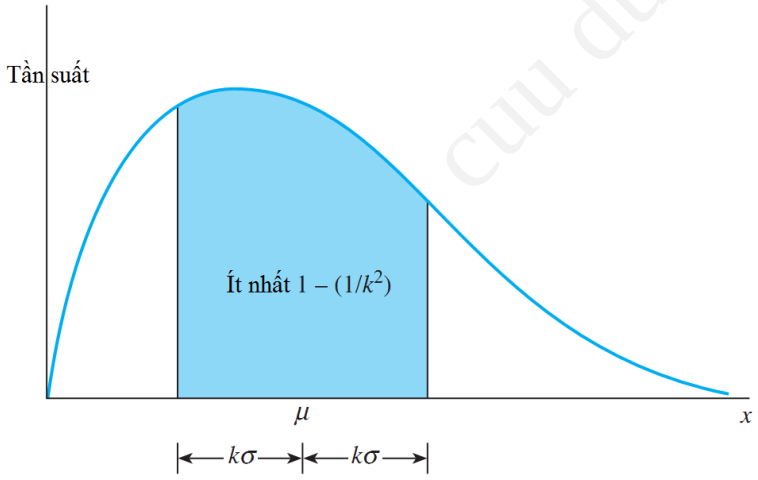
Định lý Chebyshev

Định lí 35 (Chebyshev)

Cho trước số k lớn hơn hoặc bằng 1 và một tập n số đo. Khi đó, có ít nhất $[1 - (1/k^2)]$ *các số đo sẽ nằm trong khoảng k độ lệch chuẩn so với trung bình.*

Nhận xét 36

Định lý Chebyshev áp dụng cho một tập bất kỳ các số đo của một tổng thể hoặc của một mẫu.



Quy tắc thực nghiệm (The Empirical Rule)

- Bởi vì Định lý Chebyshev áp dụng với bất kỳ phân phối nào, nên nó rất thô. Đó là tại sao ta nhấn mạnh đến "ít nhất $1 - (1/k^2)$" trong định lý này.
- Một quy tắc khác được dùng để mô tả sự biến thiên của một tập dữ liệu mặc dù không áp dụng được cho mọi tập dữ liệu, nhưng nó áp dụng rất tốt cho các dữ liệu có phân phối xấp xỉ chuẩn. Phân phối dữ liệu càng xấp xỉ chuẩn thì quy tắc áp dụng càng chính xác.
- Bởi vì phân phối chuẩn khá phổ biến trong tự nhiên, nên quy tắc này thường được áp dụng trong thực tế. Đó là lí do ta gọi là quy tắc thực nghiệm

Quy tắc thực nghiệm (The Empirical Rule)

Quy tắc thực nghiệm

Biết rằng phân phối của các số đo xấp xỉ chuẩn. Khi đó,

- Khoảng 68% các số đo nằm trong khoảng một độ lệch chuẩn so với trung bình.
- Khoảng 95% các số đo nằm trong khoảng hai độ lệch chuẩn so với trung bình.
- Khoảng 99.7% các số đo nằm trong khoảng ba độ lệch chuẩn so với trung bình.

Ví dụ 37

Các sinh viên sư phạm được dạy cách viết giáo án. Trong một nghiên cứu đánh giá mối quan hệ giữa giáo án và sự thành công của họ trong một lớp học, 25 giáo án được chấm theo thang điểm từ 0 đến 34 như trong bảng bên dưới. Sử dụng Định lý Chebyshev và Quy tắc thực nghiệm (nếu có thể) để mô tả phân phối của các điểm đánh giá này.

26.1	26.0	14.5	29.3	19.7
22.1	21.2	26.6	31.9	25.0
15.9	20.8	20.2	17.8	13.3
25.6	26.5	15.7	22.1	13.8
29.0	21.3	23.5	22.1	10.2

Ta tìm được $\bar{x} =$, $s =$ và

k	Khoảng $\bar{x} \pm ks$	Tần số	Tần suất	Chebyshev	QTTN
1				≥ 0.00	68%
2				≥ 0.75	95%
3				≥ 0.89	99.7%

Kiểm tra dữ liệu có phân phối xấp xỉ chuẩn hay không? (Dùng đồ thị stem - leaf hoặc histogram)

Ứng dụng quy tắc thực nghiệm

Để xác định các quan sát hiếm (bất thường, cực trị)

Nếu một quan sát nằm ngoài khoảng hai độ lệch chuẩn so với trung bình, nó chỉ có 5% khả năng xuất hiện. Do đó, nó được xem như là một trường hợp hiếm hay bất thường.

Nhận xét 38

Đây là một công cụ **đơn giản** và **mạnh** để xác định các số liệu ngoại lai (outliers), cực trị (extremes), hoặc bất thường, hoặc hiếm.

Ví dụ 39

Xét dữ liệu điểm giáo án trong Ví dụ 37. (Các) điểm giáo án nào được gọi là bất thường?

Ứng dụng quy tắc thực nghiệm

Để ước lượng độ lệch chuẩn từ đồ thị tổ chức tần số

- Ước lượng 2 điểm đầu của một khoảng có tâm là trung bình và chứa 95% dữ liệu.
- Quy tắc thực nghiệm nói rằng khoảng này xấp xỉ $(\bar{y} - 2s, \bar{y} + 2s)$. Do đó, chiều dài khoảng này xấp xỉ 4 lần độ lệch chuẩn.
- Ước lượng độ lệch chuẩn s = chiều dài khoảng / 4.

Ví dụ 40

Dựa vào đồ thị histogram trong Ví dụ 37 hãy ước lượng s?

Ứng dụng quy tắc thực nghiệm

Nhận xét 41

- Ước lượng từ đồ thị cho khoảng giữa chứa 95% dữ liệu có thể không đúng. Hơn nữa, quy tắc thực nghiệm áp dụng tốt nhất cho phân phối xấp xỉ chuẩn (nhưng nó cũng áp dụng tốt khi phân phối bị lệch một chút). Do đó, phương pháp ước lượng độ lệch chuẩn này sẽ chỉ cho một ước lượng chung chung, rất thô.
- Tuy nhiên, nó rất hữu ích trong việc kiểm tra các sai số lớn khi tính toán s. Chẳng hạn như thiếu chia tổng các bình phương các độ lệch cho $(n-1)$ hoặc thiếu lấy căn bậc hai của s^2. Nếu ta phạm các sai sót này thì kết quả tìm được sẽ khác biệt rất lớn so với xấp xỉ của s.

Hệ số biến thiên (coefficient of variation)

Định nghĩa 42

- **Hệ số biến thiên** (coefficient of variation) là độ lệch chuẩn được biểu diễn theo tỷ lệ phần trăm của trung bình.
- Công thức tính: $CV = \frac{s}{\bar{y}} \times 100\%$.

Nhận xét 43

- Hệ số biến thiên không bị tác động bởi các thay đổi thang đo.
- Vì thế nó là một độ đo hữu ích để so sánh các độ phân tán của hai hay nhiều biến được đo trên các thang đo khác nhau.

Ví dụ 44

Chiều cao (theo cm) và trọng lượng (theo kg) của 13 bé gái 2 tuổi được đo đạc. Chiều cao trung bình là 86.6 cm và độ lệch chuẩn là 2.9 cm. Do đó, hệ số biến thiên của chiều cao là

Trọng lượng trung bình là 12.6 kg và độ lệch chuẩn là 1.4 kg. Do đó, hệ số biến thiên của trọng lượng là

Nhận xét:

Outline

Mô tả dữ liệu **nhiều biến**

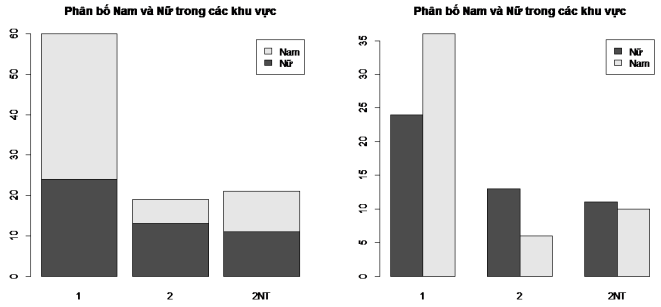
Mô tả dữ liệu của hai biến định tính

Bảng dữ liệu đồng thời [1] được thành lập, gọi là bảng ngẫu nhiên [2]. Chẳng hạn, ta có bảng ngẫu nhiên sau cho các biến giới tính (GT) và khu vực (KV)

Giới tính \ Khu vực	1	2	2NT	Tổng cộng
Nữ	24	13	11	48
Nam	36	6	10	52
Tổng cộng	60	19	21	100

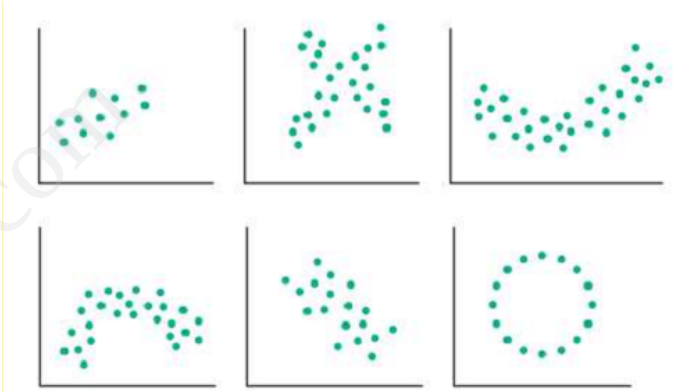
Ta có thể trình bày bảng trên dưới dạng biểu đồ cột chồng [3] hoặc biểu đồ cộ kề như sau:

[1] Cross-tabulation
[2] Contigency table
[3] Stacked bar graph



Mô tả dữ liệu của hai biến định lượng

Đồ thị phân tán [4] được sử dụng để mô tả sự quan hệ giữa hai biến định lượng.



[4] Scatter plot