THỐNG KÊ ỨNG DỤNG

ĐỖ LÂN

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 5 tháng 9 năm 2018

Nội dung môn học

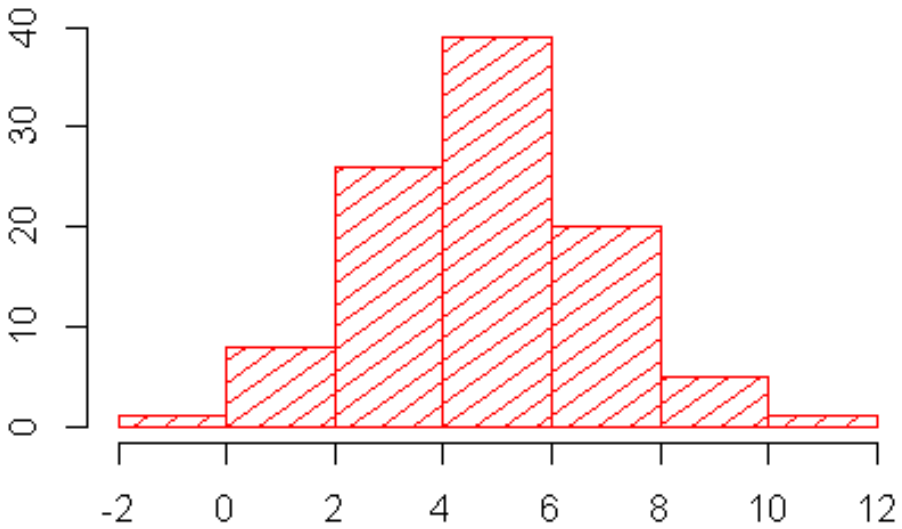
1. Tổng quan về Thống kê
2. Thu thập dữ liệu
3. Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị
4. **Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả**
5. Xác suất căn bản và biến ngẫu nhiên
6. Phân phối của tham số mẫu và ước lượng tham số tổng thể
7. Kiểm định giả thuyết về tham số một tổng thể
8. Kiểm định giả thuyết về tham số hai tổng thể
9. Phân tích phương sai
10. Kiểm định phi tham số
11. Kiểm định chi - bình phương

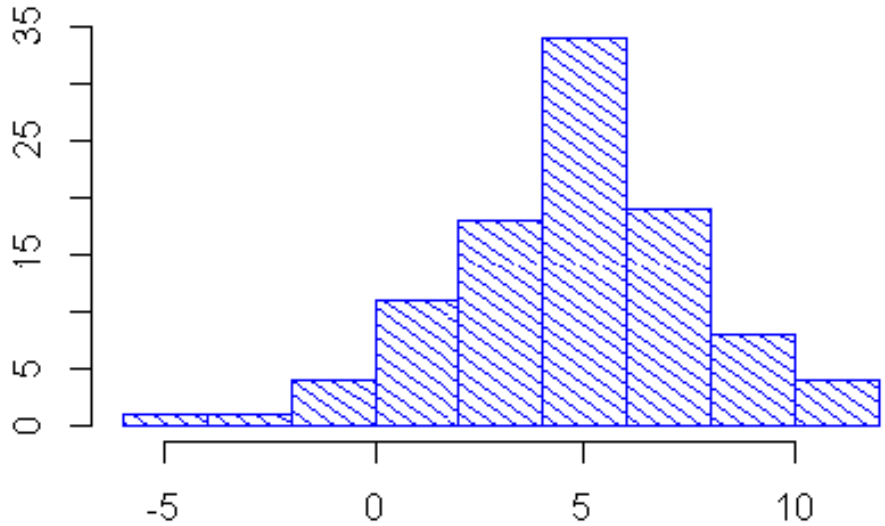
Phần IV

Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả

Mục lục

So sánh bằng cách nào?





$\rightarrow$ Ta phải có các chỉ số cụ thể mới so sánh được . . .

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

Trung bình

Định nghĩa

Trung bình cộng đơn giản *được tính bằng cách cộng tất cả các giá trị quan sát của tập dữ liệu rồi chia cho số quan sát của tập dữ liệu đó.*

Trung bình cộng đơn giản $\overline{x}$ của các giá trị $x_1, x_2, \ldots, x_n$ được cho bởi công thức:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Trung bình

Định nghĩa

Trung bình cộng đơn giản *được tính bằng cách cộng tất cả các giá trị quan sát của tập dữ liệu rồi chia cho số quan sát của tập dữ liệu đó.*

Trung bình cộng đơn giản $\overline{x}$ của các giá trị $x_1, x_2, \ldots, x_n$ được cho bởi công thức:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Ví dụ

Ví dụ: Giả sử số tiền (đơn vị nghìn VND) dùng cho chi tiêu thực phẩm trong một tuần là: 120, 150, 125, 100, 180, 140, 200. Khi đó số tiền trung bình một ngày trong tuần dành cho chi tiêu thực phẩm là:

$$\overline{x} = \frac{120 + 150 + 125 + 100 + 180 + 140 + 200}{7} = 145.$$

Trung bình cộng có trọng số

Định nghĩa

Trung bình cộng có trọng số được tính bằng cách cộng các tích của giá trị quan sát của tập dữ liệu với trọng số tương ứng rồi chia cho tổng các trọng số của tập dữ liệu đó.

Trung bình cộng có trọng số

Định nghĩa

Trung bình cộng có trọng số được tính bằng cách cộng các tích của giá trị quan sát của tập dữ liệu với trọng số tương ứng rồi chia cho tổng các trọng số của tập dữ liệu đó.

Công thức tính trung bình cộng có trọng số

- Trung bình có trọng số $\overline{x}$ của các giá trị x_i với trọng số tương ứng w_i, $\forall i = 1, \ldots, n$ được tính bởi công thức.

$$\overline{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_k w_k}{w_1 + w_2 + \ldots + w_k},$$

Ví dụ

Điểm thi kết thúc học kì của một sinh viên được cho trong bảng dưới đây:

Môn học	**Điểm**	**Số đ.v tín chỉ**
TK Ứng dụng	*7.0*	*3*
Học máy	*5.0*	*2*
Trí tuệ nhân tạo	*8.0*	*2*
Hệ điều hành	*4.0*	*2*
Mạng máy tính	*6.0*	*2*

Ví dụ

Điểm thi kết thúc học kì của một sinh viên được cho trong bảng dưới đây:

Môn học	**Điểm**	**Số đ.v tín chỉ**
TK Ứng dụng	*7.0*	*3*
Học máy	*5.0*	*2*
Trí tuệ nhân tạo	*8.0*	*2*
Hệ điều hành	*4.0*	*2*
Mạng máy tính	*6.0*	*2*

Khi đó điểm trung bình các môn học của sinh viên trên trong kì này sẽ là:

$$\overline{x} = \frac{7 \times 3 + 5 \times 2 + 8 \times 2 + 4 \times 2 + 6 \times 2}{4 + 2 + 2 + 2 + 2} = 6.17.$$

Trung bình của dữ liệu phân tổ

Ví dụ

Dữ liệu cho theo bảng sau:

Giá trị	(0, 10]	(10,20]	(20,30]	(30,40]	(40,50]
Tần số	2	1	1	6	10

Trung bình của dữ liệu phân tổ

Ví dụ

Dữ liệu cho theo bảng sau:

Giá trị	(0, 10]	(10,20]	(20,30]	(30,40]	(40,50]
Tần số	2	1	1	6	10

Khi đó

$$\bar{x} = \frac{2 \times 5 + 1 \times 15 + 1 \times 25 + 6 \times 35 + 10 \times 45}{2 + 1 + 1 + 6 + 10} = 35.5$$

Ví dụ 2

Giả sử điểm thi cuối kỳ môn TKUD được cho bởi bảng sau đây, tính điểm trung bình môn TKUD của K58.

Khoảng điểm	**Điểm đại diện** (x_i^*)	**Tần số** (f_i)
[0.0, 1.5]	0.75	13
(1.5, 3.0]	2.25	27
(3.0, 4.5]	3.75	41
(4.5, 6.0]	5.25	31
(6.0, 7.5]	6.75	18
(7.5, 9.0]	8.25	13
(9.5, 10]	9.75	3

Ưu nhược điểm của trung bình cộng

Ưu điểm

- Khái niệm trung bình cộng là quen thuộc và dễ hình dung.
- Trung bình cộng chịu sự tác động của tất cả giá trị trong tập dữ liệu, mỗi dữ liệu chỉ có 1 trung bình cộng duy nhất.
- Có nhiều tính chất tốt, thuận lợi cho thống kê suy diễn.

Nhược điểm

- Nếu trong tập dữ liệu có giá trị ngoại biên thì TBC sẽ chịu tác động của giá trị ngoại biên đó và không phản ánh sát với xu hướng chung của tập dữ liệu.
- Không dùng được cho thang đo định danh. TBC chỉ có ý nghĩ nhất với dữ liệu định lượng theo thang đo tỉ lệ.

Trung vị

Định nghĩa

Trung vị *của tập dữ liệu đã được sắp thứ tự là giá trị mà có không quá* 50% *số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị và không quá* 50% *số quan sát của tập dữ liệu có giá trị lớn hơn trung vị.*

Trung vị

Định nghĩa

Trung vị *của tập dữ liệu đã được sắp thứ tự là giá trị mà có không quá* 50% *số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị và không quá* 50% *số quan sát của tập dữ liệu có giá trị lớn hơn trung vị.*

Để tìm trung vị của tập dữ liệu có n quan sát ta làm như sau:

- Sắp xếp lại tập dữ liệu theo một chiều tăng hoặc giảm.
- Nếu số quan sát n là số lẻ thì trung vị là quan sát ở vị trí thứ $(n+1)/2$.
- Nếu số quan sát n là số chẵn thì trung vị giá trị trung bình cộng của hai quan sát ở vị trí chính giữa của tập dữ liệu, tức là hai quan sát ở vị trí thứ $n/2$ và $(n+2)/2$.

Trung vị

Ví dụ 1

Tính trung vị của tập dữ liệu về tuổi của 9 người như sau:

$$5, 11, 9, 12, 10, 20, 15, 30, 25.$$

Trung vị

Ví dụ 1

Tính trung vị của tập dữ liệu về tuổi của 9 người như sau:

$$5, 11, 9, 12, 10, 20, 15, 30, 25.$$

Đáp án: 12

Trung vị

Ví dụ 1

Tính trung vị của tập dữ liệu về tuổi của 9 người như sau:

$$5, 11, 9, 12, 10, 20, 15, 30, 25.$$

Đáp án: 12
Lí do: Sau khi sắp xếp tập dữ liệu theo chiều tăng dần, ta được tập dữ liệu như sau:

$$5 \quad 9 \quad 10 \quad 11 \quad 12 \quad 15 \quad 20 \quad 25 \quad 30$$

Với 9 dữ liệu như vậy, số trung vị là số chính giữa, là số thứ 5, tức là 12. Như vậy có nghĩa là, có khoảng một nửa số người được điều tra có tuổi nhỏ hơn hoặc bằng 12, còn khoảng một nửa có tuổi lớn hơn hoặc bằng 12.

Ví dụ 2

Tính trung vị của tập dữ liệu về tuổi của 8 người như sau:

$$5, 11, 9, 12, 10, 20, 15, 30.$$

Ví dụ 2

Tính trung vị của tập dữ liệu về tuổi của 8 người như sau:

$$5, 11, 9, 12, 10, 20, 15, 30.$$

Đáp án: 11.5

Ví dụ 2

Tính trung vị của tập dữ liệu về tuổi của 8 người như sau:

$$5, 11, 9, 12, 10, 20, 15, 30.$$

Đáp án: 11.5
Lí do: Sau khi sắp xếp tập dữ liệu theo chiều tăng dần, ta được tập dữ liệu như sau:

$$5 \quad 9 \quad 10 \quad 11 \quad 12 \quad 15 \quad 20 \quad 30$$

Với dữ liệu như vậy, số trung vị là trung bình cộng của 2 số chính giữa, là các số thứ 4 và 5, tức là $\frac{11+12}{2} = 11.5$.

Ưu nhược điểm của trung vị

Ưu điểm

- Trung vị của tập dữ liệu là duy nhất.
- Không bị tác động của giá trị ngoại biên.
- Có thể tìm trung vị cho những tập dữ liệu định lượng bằng thang đo khoảng trở lên.

Nhược điểm

- Trung vị không được tác động bởi toàn bộ giá trị của tập dữ liệu.
- Không có nhiều tính chất tốt để phục vụ thống kê suy diễn.

Mode

Định nghĩa

Mode *(yếu vị) của một tập dữ liệu là giá trị xuất hiện nhiều nhất trong tập dữ liệu.*

Mode

Định nghĩa

Mode *(yếu vị) của một tập dữ liệu là giá trị xuất hiện nhiều nhất trong tập dữ liệu.*

Để tìm mode của một tập dữ liệu ta làm như sau:

- Lập bảng tần số cho tập dữ liệu;
- Tìm giá trị lớn nhất trong các tần số;
- Tìm các giá trị của tập dữ liệu tương ứng với tần số lớn nhất; và mode của tập dữ liệu là các giá trị này.

Mode

Định nghĩa

Mode *(yếu vị) của một tập dữ liệu là giá trị xuất hiện nhiều nhất trong tập dữ liệu.*

Để tìm mode của một tập dữ liệu ta làm như sau:

- Lập bảng tần số cho tập dữ liệu;
- Tìm giá trị lớn nhất trong các tần số;
- Tìm các giá trị của tập dữ liệu tương ứng với tần số lớn nhất; và mode của tập dữ liệu là các giá trị này.

Ví dụ

Mode của tập dữ liệu:

- $0, 1, 3, 1, 5, 2, 6, 2, 9, 2$ *là 2.*
- $2, 3, 2, 5, 7, 8, 7, 15$ *là 2 và 7.*
- $0, 1, 2, 3, 4, 5, 6$ *là tất cả các phần tử của tập.*

Example

Tìm các số đo trung tâm của mỗi cột trong tập dữ liệu sau:

Thứ tự	Xếp loại	Giới tính	Lương
1	Kha	Nu	14
2	Kha	Nu	10
3	Kha	Nam	14
4	TrungBinh	Nu	12
5	TrungBinh	Nu	14
6	Kha	Nam	5
7	Gioi	Nam	11
8	Kha	Nu	6
9	Gioi	Nam	12
11	TrungBinh	Nu	10

1. So sánh lương trung bình của nhóm nam và nhóm nữ.
2. Trong các nhóm xếp loại tốt nghiệp khá, giỏi, trung bình, nhóm nào có lương trung bình lớn nhất?

Ưu nhược điểm của Mode

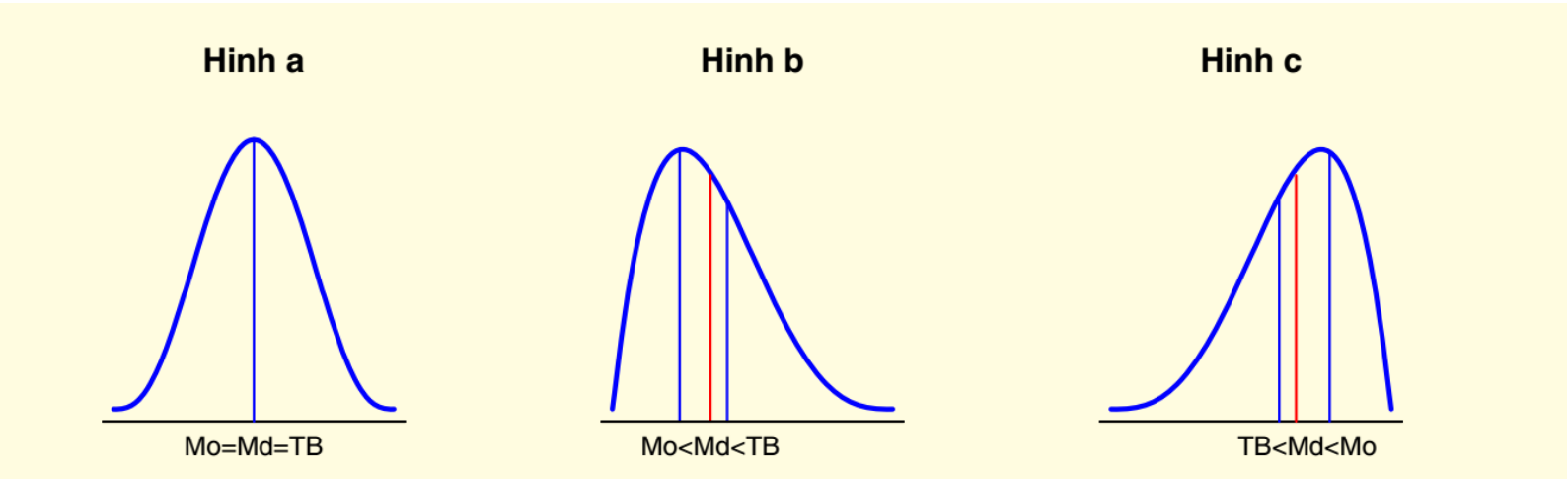
Ưu điểm

- Số mode có thể dùng với mọi loại dữ liệu định tính cũng như định lượng.
- Không bị ảnh hưởng của giá trị ngoại biên.

Nhược điểm

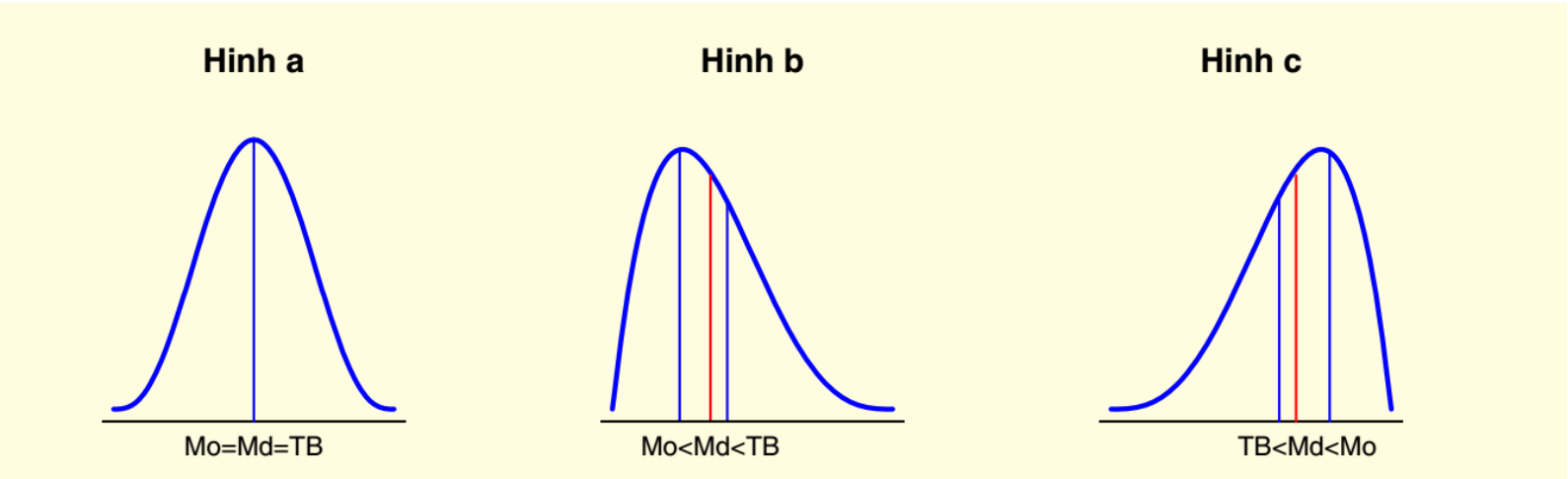
- Nếu có nhiều số Mode thì khó lí giải và so sánh.

So sánh trung bình, trung vị, mode



Hình dáng của phân phối của một tập dữ liệu cũng được phản ánh qua mối quan hệ hơn kém giữa Trung bình, Trung vị của tập dữ liệu đó.
Trong hình (a), hình dáng của đa giác tần số đối xứng khi $\mu = Mode = M_0$.

So sánh trung bình, trung vị, mode



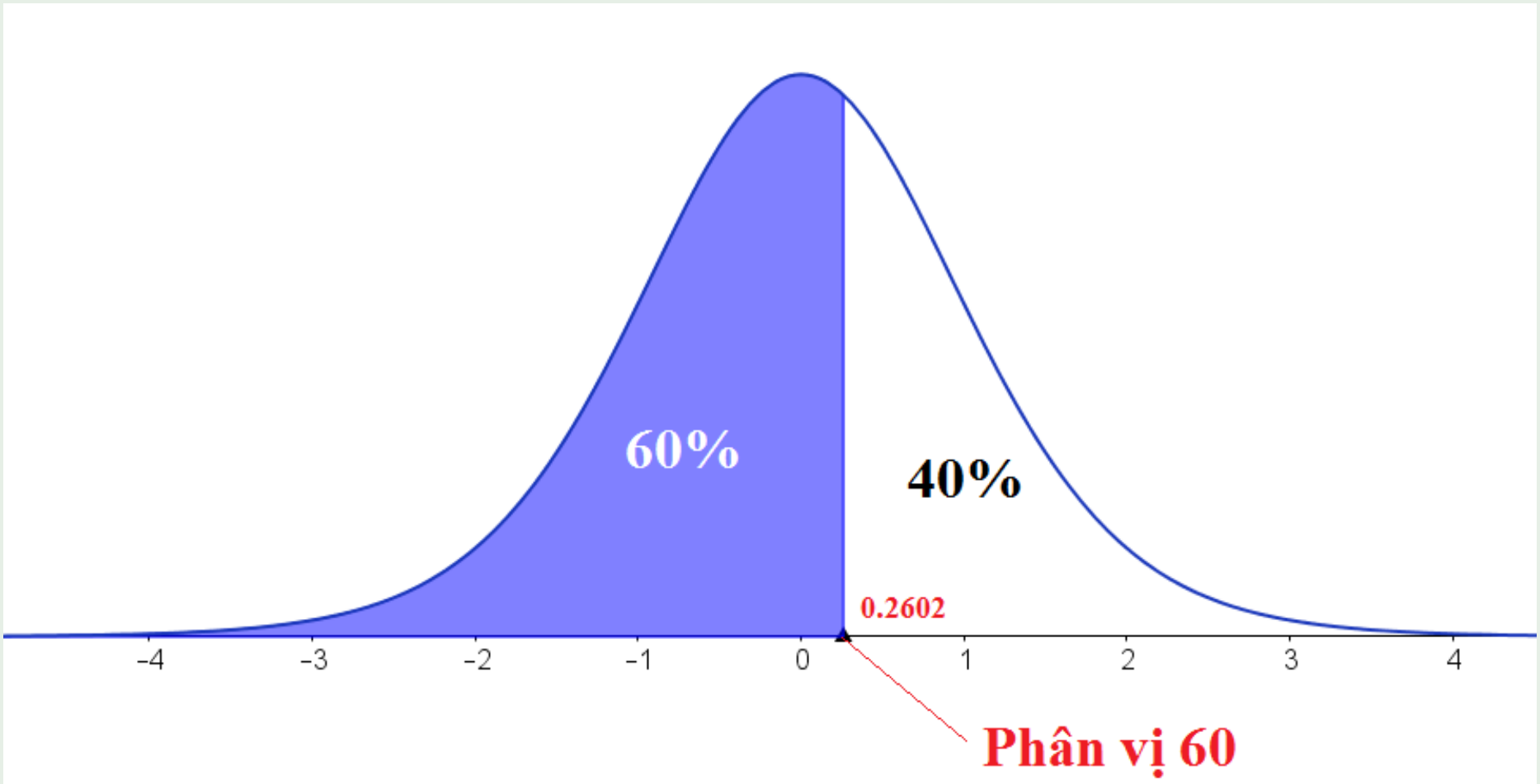
Trong hình (b), hình dáng của đa giác tần số lệch (nghiêng) phải với một cái "đuôi" kéo dài về bên phải khi $\mu > Mode > M_0$. Trong hình (c), hình dáng của đa giác tần số lệch (nghiêng) trái với một cái "đuôi" kéo dài về bên trái khi $\mu < Mode < M_0$.

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

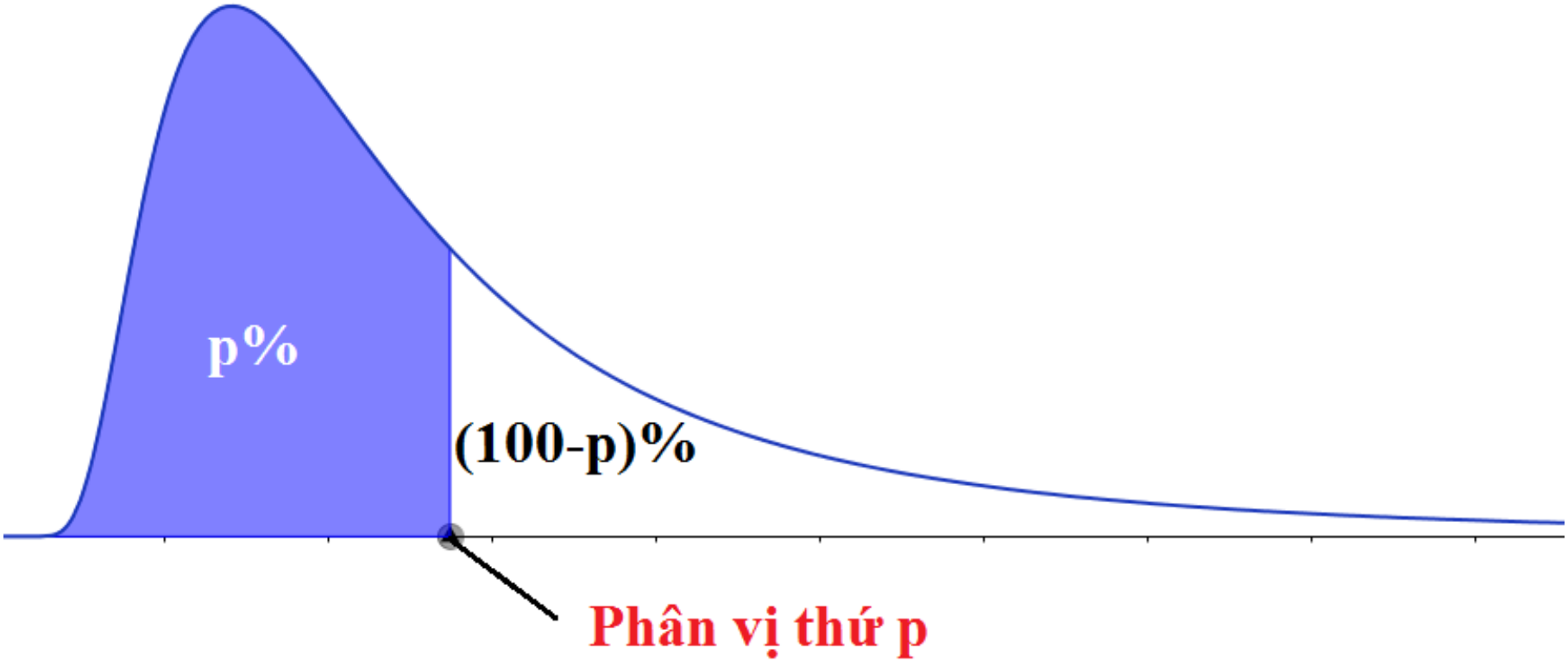
1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

Phân vị 60

Example



Phân vị p



Phân vị

Definition

Phân vị thứ p của một tập dữ liệu đã được sắp thứ tự là giá trị chia tập dữ liệu thành hai phần, một phần không quá $p\%$ số quan sát có giá trị nhỏ hơn phân vị thứ p, phần còn lại có không quá $(100 - p)\%$ số quan sát lớn hơn phân vị thứ p.

Phân vị

Definition

Phân vị thứ p của một tập dữ liệu đã được sắp thứ tự là giá trị chia tập dữ liệu thành hai phần, một phần không quá $p\%$ số quan sát có giá trị nhỏ hơn phân vị thứ p, phần còn lại có không quá $(100 - p)\%$ số quan sát lớn hơn phân vị thứ p.

Có thể sử dụng một cách **ước lượng** phân vị thứ p của tập dữ liệu có n phần tử như sau:

1. Sắp xếp dữ liệu theo chiều tăng dần.
2. Phân vị thứ p là giá trị có vị trí được xác định bởi công thức $i = \frac{p}{100}(n+1)$.

Ví dụ

Tìm phân vị thứ 25, thứ 50, thứ 60 và thứ 75 của tập dữ liệu sau:
9 9 10 11 13 13 13 15 16 20 20 24

Tứ phân vị

Định nghĩa

Tứ phân vị là 3 giá trị gồm phân vị thứ 25, thứ 50 và thứ 75. Ba giá trị này lần lượt được kí hiệu là Q_1, Q_2, Q_3, chúng chia tập dữ liệu đã sắp xếp theo trật tự thành bốn phần có số quan sát bằng nhau.

Tứ phân vị

Định nghĩa

Tứ phân vị là 3 giá trị gồm phân vị thứ 25, thứ 50 và thứ 75. Ba giá trị này lần lượt được kí hiệu là Q_1, Q_2, Q_3, chúng chia tập dữ liệu đã sắp xếp theo trật tự thành bốn phần có số quan sát bằng nhau.

Example

Tứ phân vị của tập dữ liệu: $10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70$ tính trên phần mềm R lần lượt là 12.25, 22.50, 36.25.

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

Khoảng biến thiên

Định nghĩa

Khoảng biến thiên của một tập dữ liệu được tính bởi công thức:

$$R = X_{max} - X_{min}$$

Khoảng biến thiên

Định nghĩa

Khoảng biến thiên của một tập dữ liệu được tính bởi công thức:

$$R = X_{max} - X_{min}$$

- Nhược điểm của khoảng biến thiên là chỉ phụ thuộc vào hai giá trị lớn nhất và nhỏ nhất của tập dữ liệu nên nó thay đổi nhạy cảm với các quan sát ngoại lệ. Ngoài ra, nó bỏ qua các thông tin về cách phân bố nội bộ tập dữ liệu.

Khoảng biến thiên

Định nghĩa

Khoảng biến thiên của một tập dữ liệu được tính bởi công thức:

$$R = X_{max} - X_{min}$$

- Nhược điểm của khoảng biến thiên là chỉ phụ thuộc vào hai giá trị lớn nhất và nhỏ nhất của tập dữ liệu nên nó thay đổi nhạy cảm với các quan sát ngoại lệ. Ngoài ra, nó bỏ qua các thông tin về cách phân bố nội bộ tập dữ liệu.
- **Ví dụ**
  1. Cho tập dữ liệu: 1 4 3 6 7 thì $R = 7 - 1 = 6$.

Khoảng biến thiên

Định nghĩa

Khoảng biến thiên của một tập dữ liệu được tính bởi công thức:

$$R = X_{max} - X_{min}$$

- Nhược điểm của khoảng biến thiên là chỉ phụ thuộc vào hai giá trị lớn nhất và nhỏ nhất của tập dữ liệu nên nó thay đổi nhạy cảm với các quan sát ngoại lệ. Ngoài ra, nó bỏ qua các thông tin về cách phân bố nội bộ tập dữ liệu.
- **Ví dụ**
  1. Cho tập dữ liệu: 1 4 3 6 7 thì $R = 7 - 1 = 6$.
  2. Cho tập dữ liệu: 1 4 3 6 7 100 thì $R = 100 - 1 = 99$.

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

Độ trải giữa

Định nghĩa

Độ trải giữa còn được gọi là khoảng tứ phân vị được tính bằng hiệu giữa phân vị thứ ba và phân vị thứ nhất.

Công thức

$$R_Q = Q_3 - Q_1$$

Độ trải giữa

Định nghĩa

Độ trải giữa còn được gọi là khoảng tứ phân vị được tính bằng hiệu giữa phân vị thứ ba và phân vị thứ nhất.
Công thức

$$R_Q = Q_3 - Q_1$$

- **Ví dụ** Xét tập dữ liệu:
  10 15 20 25 30 40 80 90
  Khi đó $R_Q == Q_3 - Q_1 = 70 - 16.26 = 53.75$.

Độ trải giữa

Định nghĩa

Độ trải giữa còn được gọi là khoảng tứ phân vị được tính bằng hiệu giữa phân vị thứ ba và phân vị thứ nhất.
Công thức

$$R_Q = Q_3 - Q_1$$

- **Ví dụ** Xét tập dữ liệu:
  10 15 20 25 30 40 80 90
  Khi đó $R_Q == Q_3 - Q_1 = 70 - 16.26 = 53.75$.
- Độ trải giữa cũng khắc phục được nhược điểm của khoảng biến thiên nhưng nó ko xem xét đến cách thức phân bố của tất cả các quan sát trong tập dữ liệu.

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

Phương sai

Định nghĩa

- **Phương sai của một tập dữ liệu tổng thể**, *kí hiệu là* σ^2, *được xác định bởi công thức:* $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$, *ở đây* μ *là trung bình của tổng thể và N là số phần tử trong tổng thể.*

Phương sai

Định nghĩa

- **Phương sai của một tập dữ liệu tổng thể**, *kí hiệu là* σ^2, *được xác định bởi công thức:* $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$, *ở đây* μ *là trung bình của tổng thể và* N *là số phần tử trong tổng thể.*
- **Phương sai của một tập dữ liệu mẫu**, *kí hiệu là* s^2, *được xác định bởi công thức:* $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$, *ở đây* $\overline{x}$ *là trung bình của mẫu và* n *là số quan sát trong mẫu.*

Độ lệch chuẩn

Definition

- **Độ lệch chuẩn của một tập dữ liệu tổng thể**, kí hiệu là σ, là căn bậc hai của phương sai của tổng thể:

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

Độ lệch chuẩn

Definition

- **Độ lệch chuẩn của một tập dữ liệu tổng thể**, kí hiệu là σ, là căn bậc hai của phương sai của tổng thể:

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

- **Độ lệch chuẩn của một tập dữ liệu mẫu**, kí hiệu là s, là căn bậc hai của phương sai mẫu:

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

Example

Cho tập dữ liệu

$$10, 15, 32, 18, 25, 65, 30, 38.$$

Ta coi nó như là một mẫu thì phương sai là 297.2679 và độ lệch chuẩn là 17.24146.

Example

Tính phương sai và độ lệch chuẩn mẫu của tập dữ liệu

$$1, 3, 3, 5, 8.$$

Example

Tính phương sai và độ lệch chuẩn của lương của mỗi nhóm nam và nữ trong tập dữ liệu sau, lương của nhóm nào đồng đều hơn?

Thứ tự	Xếp loại	Giới tính	Lương
1	Kha	Nu	14
2	Kha	Nu	10
3	Kha	Nam	14
4	TrungBinh	Nu	12
5	TrungBinh	Nu	14
6	Kha	Nam	5
7	Gioi	Nu	11
8	Kha	Nu	6
9	Kha	Nu	13
10	Gioi	Nam	12
11	TrungBinh	Nu	10
12	Kha	Nam	6

Tính các đại lượng thống kê mô tả trong **R**

mean(x, trim = 0, na.rm = FALSE, ...) Trung bình
median(x, na.rm = FALSE) Trung vị
range(x, na.rm = FALSE) khoảng biến thiên
var(x, na.rm = FALSE) Phương sai mẫu
sd(x, na.rm = FALSE) độ lệch chuẩn mẫu
quantile(x, probs = seq(0, 1, 0.25),na.rm=FALSE,...) phân vị
fivenum(x, na.rm = TRUE)
summary(x)
boxplot(x,horizontal=FALSE,...) biểu đồ hộp và râu

Tính các đại lượng thống kê mô tả trong **R**

mean(x, trim = 0, na.rm = FALSE, ...) Trung bình
median(x, na.rm = FALSE) Trung vị
range(x, na.rm = FALSE) khoảng biến thiên
var(x, na.rm = FALSE) Phương sai mẫu
sd(x, na.rm = FALSE) độ lệch chuẩn mẫu
quantile(x, probs = seq(0, 1, 0.25),na.rm=FALSE,...) phân vị
fivenum(x, na.rm = TRUE)
summary(x)
boxplot(x,horizontal=FALSE,...) biểu đồ hộp và râu

trong đó

x: tập dữ liệu
na.rm: tham số logic TRUE/FALSE không bỏ qua/bỏ qua những giá trị trống trong tính toán

trong đó

x: tập dữ liệu
na.rm: tham số logic TRUE/FALSE không bỏ qua/bỏ qua những giá trị trống trong tính toán

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

Biểu đồ hộp và râu

Định nghĩa

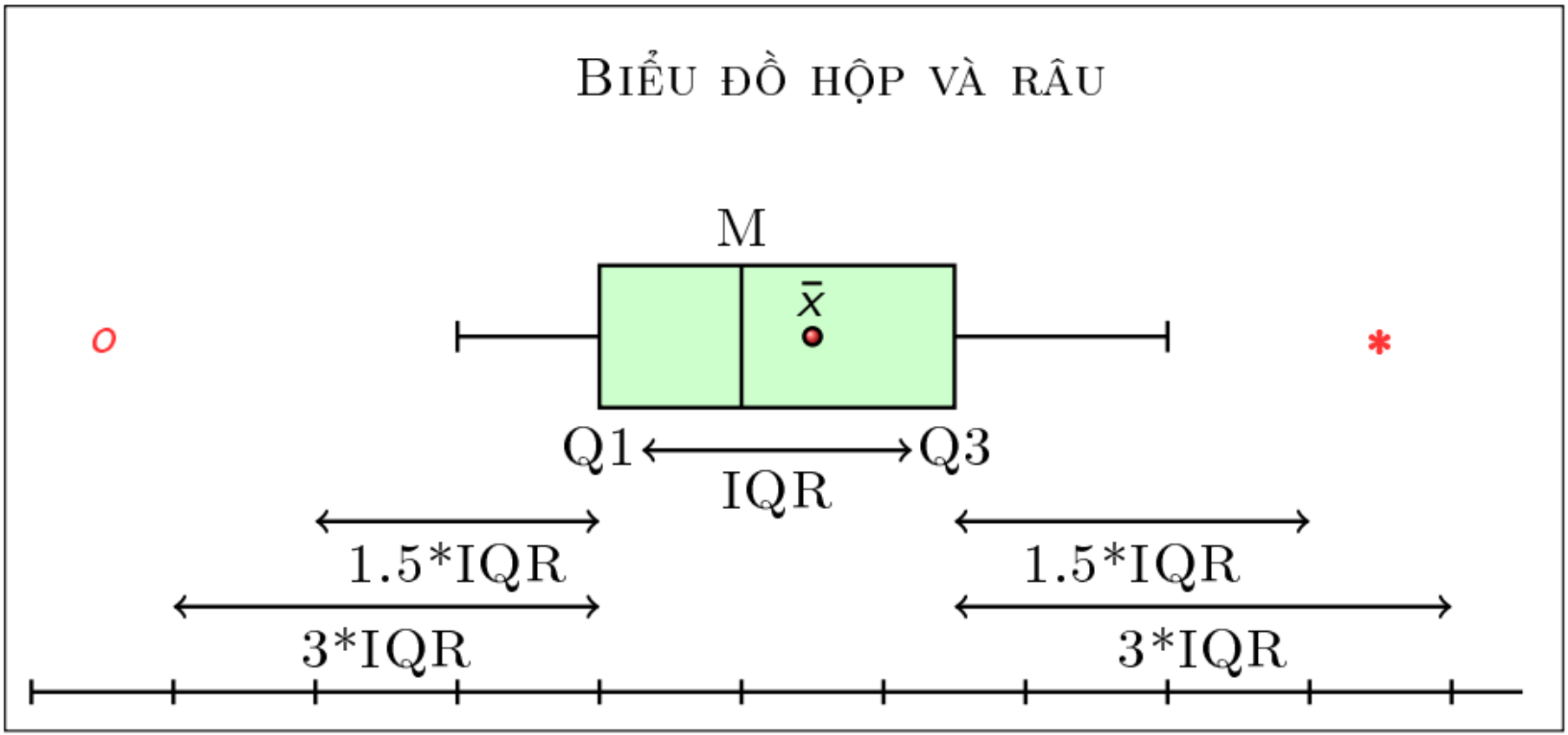
Biểu đồ hộp và râu là một biểu đồ nghiên cứu về độ tập trung, độ phân tán và phân phối của một tập dữ liệu định lượng. Trong biểu đồ hộp và râu thể hiện đồng thời các thông tin sau: Giá trị lớn nhất, giá trị nhỏ nhất, tứ phân vị, đôi khi cả các giá trị ngoại lệ.

Để vẽ biểu đồ hộp và râu của một tập dữ liệu, ta phải tính các đại lượng sau:

- Trung bình, trung vị
- Tứ phân vị thứ nhất, tứ phân vị thứ 3, độ trải giữa.
- Giá trị lớn nhất, giá trị nhỏ nhất và các giá trị ngoại biên.

Hộp hình chữ nhật thể hiện 50% các quan sát ở giữa tập dữ liệu, chiều rộng của hôp bằng độ trải giữa R_Q, hai cạnh bên đi qua giá trị tứ phân vị thứ nhất và thứ 3. Đường thẳng đứng trong hộp đi qua giá trị tứ phân vị thứ hai. Hai râu của biểu đồ biểu diễn 25% quan sát phía dưới Q_1 và 25% quan sát phía trên Q_3. Râu trái đi từ Q_1 đến giá trị nhỏ nhất, râu phải đi từ Q_3 đến giá trị lớn nhất.

Biểu đồ hộp và râu



Tác dụng của biểu đồ

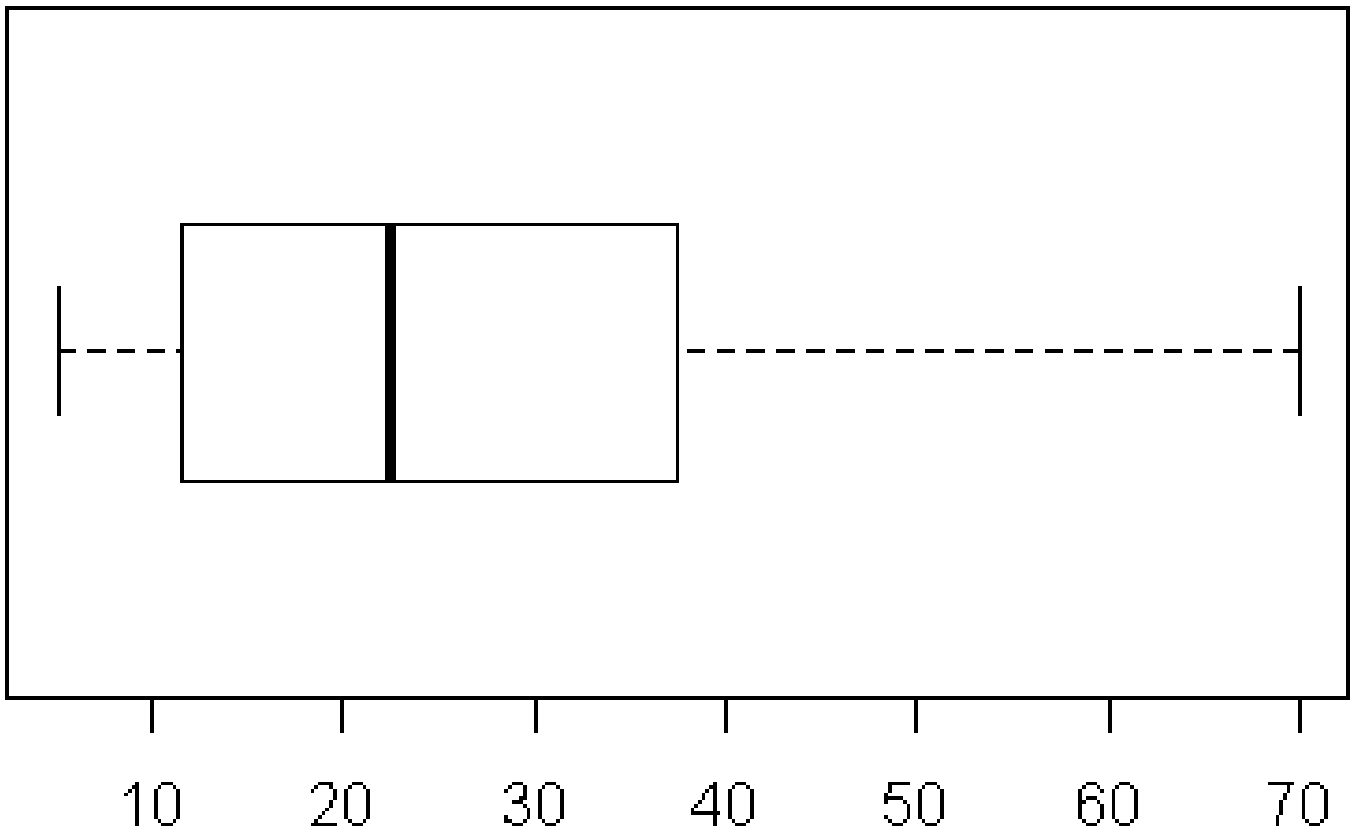
- Cho ta cái nhìn tổng quát về sự phân tán của tập dữ liệu.
- Cho ta biết tính chất đối xứng hay nghiêng của tập dữ liệu.
- Cho ta biết các giá trị ngoại biên.
- Dễ dàng so sánh nhiều tập dữ liệu khi vẽ các biểu đồ cạnh nhau.

Example

Tập dữ liệu

$$10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70$$

có giá trị nhỏ nhất là 5, lớn nhất là 70 và tứ phân vị là: 12.25 22.50 36.25.



Example

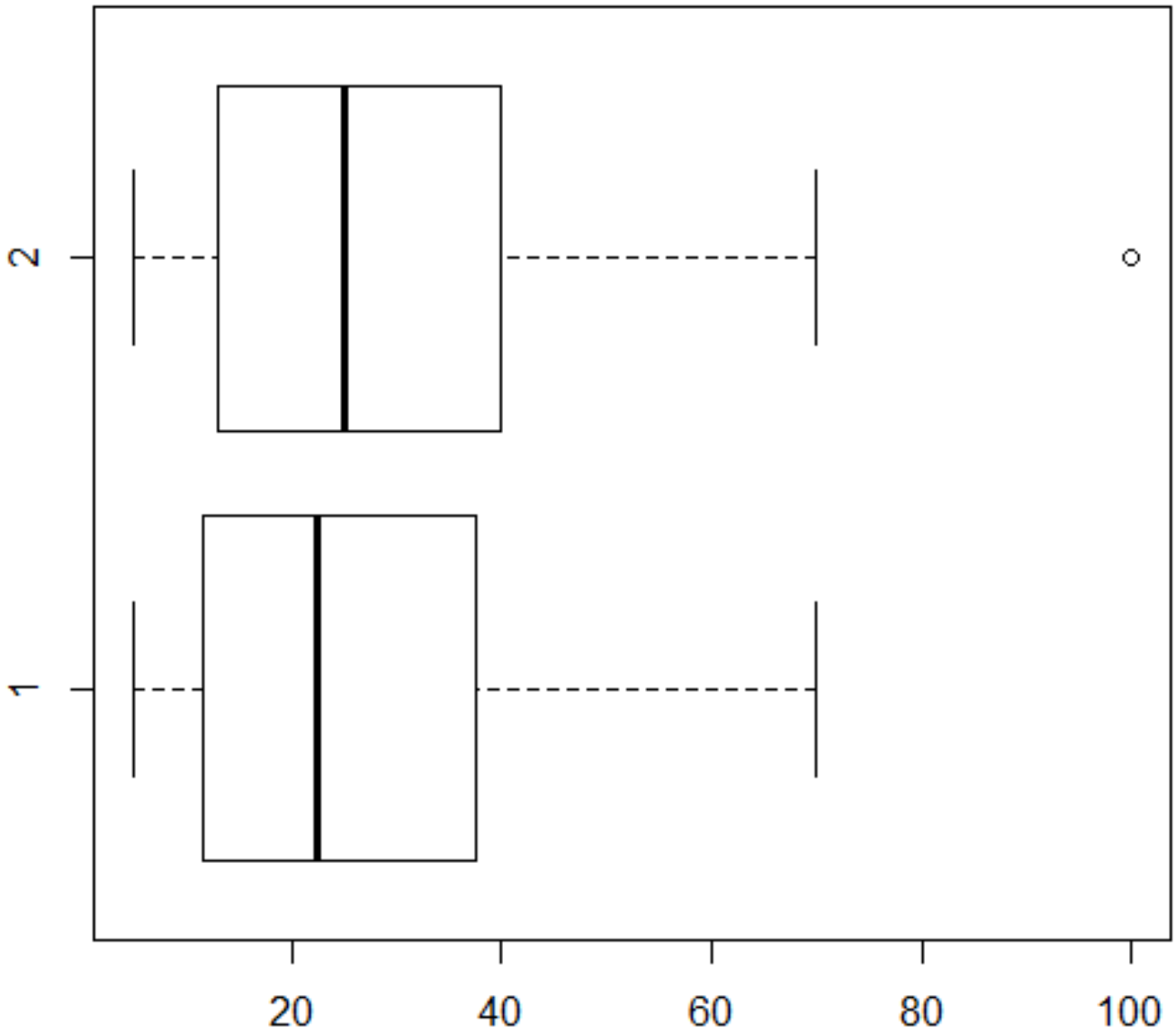
tập dữ liệu

$$10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70$$

có giá trị nhỏ nhất là 5, lớn nhất là 70 và tứ phân vị là: 12.25 22.50 36.25.
và tập dữ liệu

$$10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70, 100$$

có có giá trị nhỏ nhất là 5, lớn nhất là 100 và tứ phân vị là: 13, 25, 40.
Ta có thể vẽ được biểu đồ hộp và râu của hai tập dữ liệu trên như sau:


2
1
20
40
60
80
100

Vẽ biểu đồ hộp và râu trong **R**

hàm boxplots

boxplot(x, border = "", col = "", horizontal=FALSE)

trong đó

- x: vector dữ liệu số cần vẽ biểu đồ;
- border: tham số chỉ màu của râu, đường biên của hộp và giá trị ngoại biên;
- col: màu của hộp;
- horizontal: tham số logic chỉ cách vẽ biểu đồ là đứng hay ngang (=FALSE thì biểu đồ được vẽ đứng, =TRUE thì biểu đồ được vẽ ngang).

Vẽ biểu đồ hộp và râu trong **R**

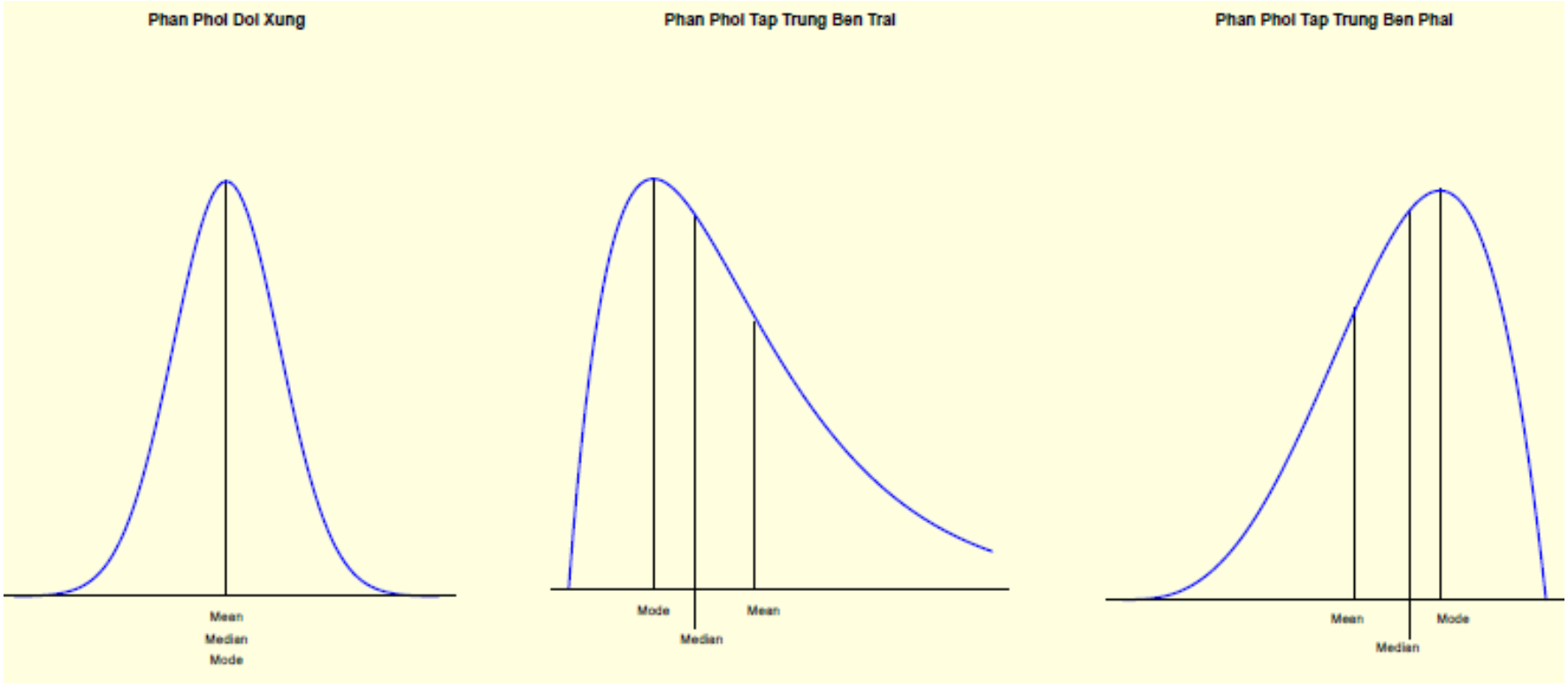
hàm boxplots

boxplot(x, border = "", col = "", horizontal=FALSE)

trong đó

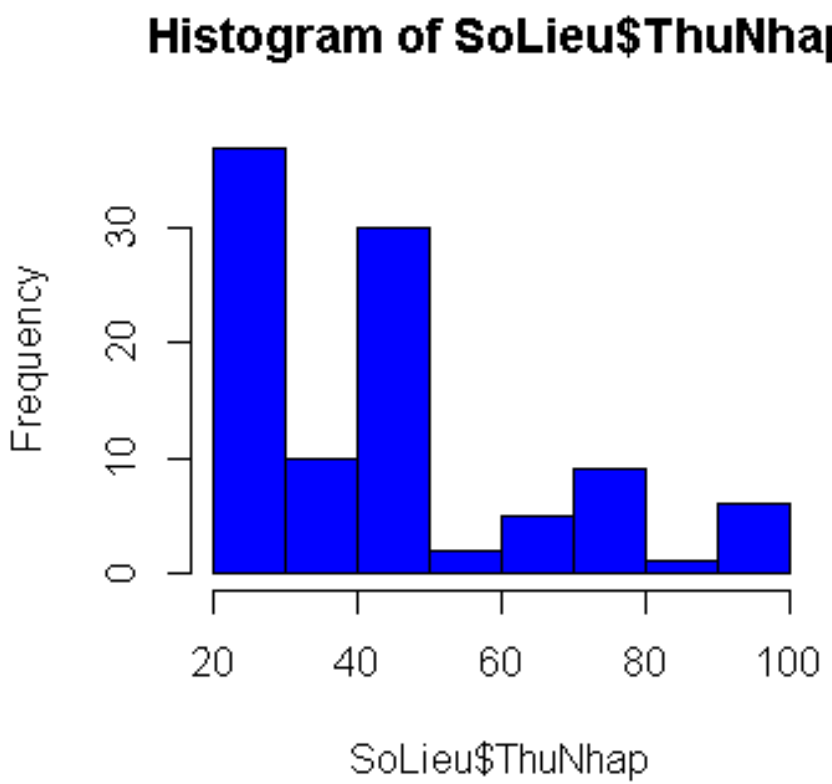
- x: vector dữ liệu số cần vẽ biểu đồ;
- border: tham số chỉ màu của râu, đường biên của hộp và giá trị ngoại biên;
- col: màu của hộp;
- horizontal: tham số logic chỉ cách vẽ biểu đồ là đứng hay ngang (=FALSE thì biểu đồ được vẽ đứng, =TRUE thì biểu đồ được vẽ ngang).

Hình dáng của tập dữ liệu

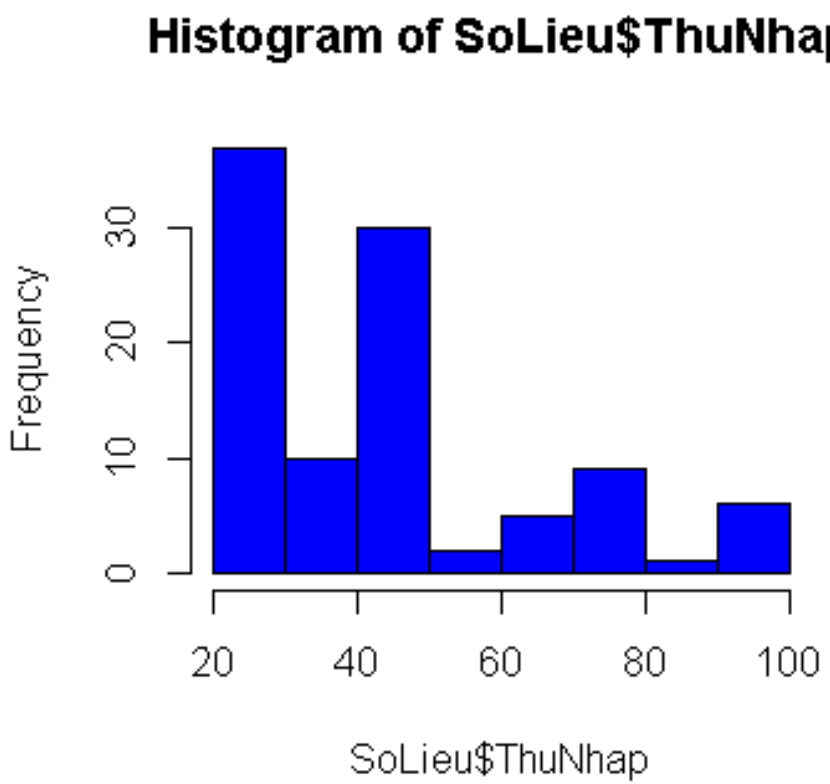


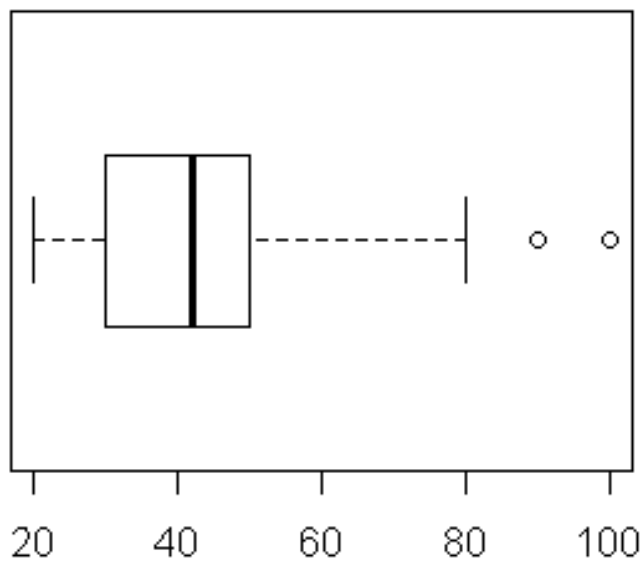
2	00000000025555555599
3	00000000000000000000122223557
4	02222222268888888888888
5	00000000058
6	
7	0000058888
8	0000
9	0
10	000000

```
 2 | 00000000025555555599
 3 | 00000000000000000000122223557
 4 | 02222222268888888888888
 5 | 00000000058
 6 |
 7 | 0000058888
 8 | 0000
 9 | 0
10 | 000000
```



```
 2 | 00000000025555555599
 3 | 00000000000000000000122223557
 4 | 02222222268888888888888
 5 | 00000000058
 6 |
 7 | 0000058888
 8 | 0000
 9 | 0
10 | 000000
```





Tính các đại lượng thống kê mô tả trong **R**

mean(x, trim = 0, na.rm = FALSE, ...) Trung bình
median(x, na.rm = FALSE) Trung vị
range(x, na.rm = FALSE) khoảng biến thiên
var(x, na.rm = FALSE) Phương sai mẫu
sd(x, na.rm = FALSE) độ lệch chuẩn mẫu
quantile(x, probs = seq(0, 1, 0.25),na.rm=FALSE,...) phân vị
fivenum(x, na.rm = TRUE)
summary(x)
boxplot(x,horizontal=FALSE,...) biểu đồ hộp và râu

Tính các đại lượng thống kê mô tả trong **R**

mean(x, trim = 0, na.rm = FALSE, ...) Trung bình
median(x, na.rm = FALSE) Trung vị
range(x, na.rm = FALSE) khoảng biến thiên
var(x, na.rm = FALSE) Phương sai mẫu
sd(x, na.rm = FALSE) độ lệch chuẩn mẫu
quantile(x, probs = seq(0, 1, 0.25),na.rm=FALSE,...) phân vị
fivenum(x, na.rm = TRUE)
summary(x)
boxplot(x,horizontal=FALSE,...) biểu đồ hộp và râu

trong đó

x: tập dữ liệu
na.rm: tham số logic TRUE/FALSE không bỏ qua/bỏ qua những giá trị trống trong tính toán

trong đó

x: tập dữ liệu
na.rm: tham số logic TRUE/FALSE không bỏ qua/bỏ qua những giá trị trống trong tính toán

1. Các số đo hướng tâm của tập dữ liệu
   - Trung bình cộng, trung vị và mode
2. Các đại lượng mô tả sự phân bố của tập dữ liệu
   - Tứ phân vị và phân vị thứ p
3. Các đại lượng đo lường độ phân tán
   - Khoảng biến thiên
   - Độ trải giữa
   - Phương sai, độ lệch chuẩn
   - Biểu đồ hộp và râu
4. Ví dụ quanh ta

Chọn nghề nào?

Example

Giả sử có hai nghề A và B. Trên toàn thế giới có 6 người làm nghề A, 8 người làm nghề B với mức lương như sau:

Người	A1	A2	A3	A4	A5	A6
Lương (nghìn đô la)	4	5	6	4	5	6

Người	B1	B2	B3	B4	B5	B6	B7	B8
Lương (Nghìn đô la)	1	9	5	2	8	5	1	9

Nếu chỉ dựa vào lương để chọn nghề thì nên chọn nghề nào?

Bollinger Bands

Bollinger Bands

Thống kê quanh ta - Số liệu ngày 31/3/2018

Largest Audience

YAN News

Total fans

13 860 390

MC Trấn Thành

Total fans

10 898 405

Trần Khởi My

Total fans

10 732 842

Fastest-Growing Pages in Vietnam

Last Day

Beta Cineplex

+49 452 Fans ↑

Clip Vui

+31 915 Fans ↑

VNlady

+23 468 Fans ↑

Blogtamsu...

+21 982 Fans ↑

WSS Beauty

+21 049 Fans ↑

Thống kê quanh ta - Cập nhật ngày 31/3/2017

Largest Audience

Tiếng Anh là chuyện nhỏ

Total fans

1 955 259

Tuyensinh247.com - Học trực tuyến

Total fans

1 145 610

Hocmai.vn Online

Total fans

906 551

Fastest-Growing Education Pages in Vietnam

Last Day

Tiếng Anh là…

+893 Fans ↑

Hocmai.vn…

+843 Fans ↑

Nói thành thạo…

+785 Fans ↑

Tiếng Anh Cho…

+690 Fans ↑

Ms Hoa Toeic

+550 Fans ↑

https://www.socialbakers.com/statistics/facebook/pages/total/vietnam/soci

So sánh mức phổ biến của các phần mềm thống kê

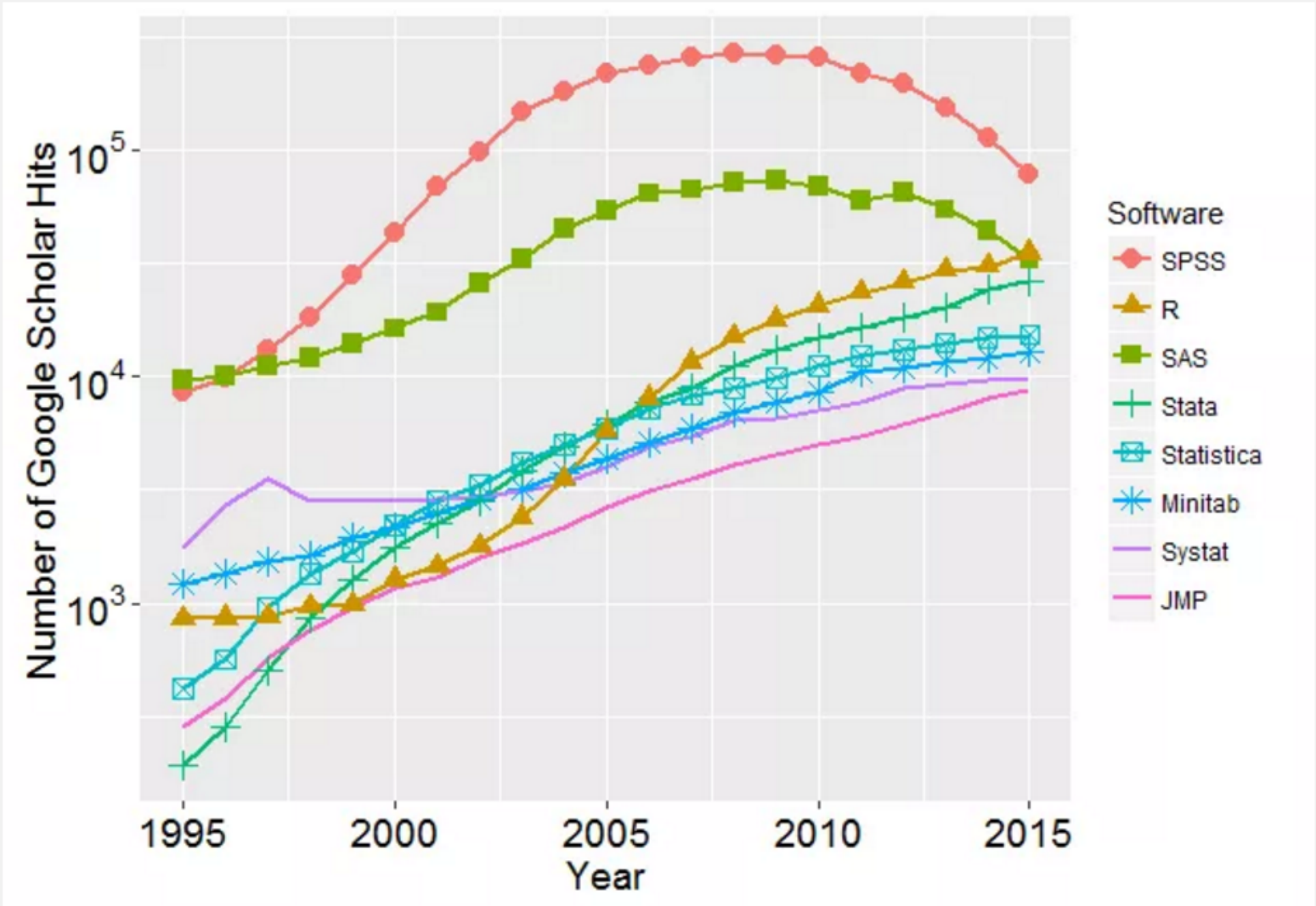


Figure 2f. A logarithmic view of the number of scholarly articles found in each year by Google Scholar. This combines the previous two figures into one by compressing the y-axis with a base 10 logarithm.