Bài giảng Xác suất Thống kê và Ứng dụng

Phan Thanh Hồng

Bộ môn Toán-Đại học THĂNG LONG

Ngày 21 tháng 11 năm 2009

Phần IV

Các đại lượng thống kê mô tả

Phần IV

Phần IV

Phần IV

Phần IV

1. Các đại lượng mô tả độ tập trung (trung tâm)
2. Các đại lượng mô tả độ phân tán
3. Phân vị, tứ phân vị, độ trải giữa, biểu đồ hộp và râu
4. Sử dụng kết hợp trung bình và độ lệch chuẩn

Các đại lượng thống kê mô tả

Phần này trình bày về các đại lượng thống kê mô tả của một tổng thể các quan sát và mẫu những quan sát được rút ra một cách ngẫu nhiên từ một tổng thể. Và vì phần lớn các trường hợp, ta khó có thể tính toán các đại lượng này cho tập dữ liệu tổng thể nên chủ yếu ta làm việc với các mẫu và dùng các tính toán trên mẫu để ước lượng cho tổng thể.
Phần này cũng sẽ trình bày cách tính các đại lượng thống kê mô tả cho hai trường hợp tập dữ liệu mẫu gốc và tập dữ liệu mẫu đã được tóm tắt trong bảng tần số.

Các đại lượng mô tả độ tập trung (trung tâm)

1. Trung bình
2. Trung vị
3. Mode

Các đại lượng mô tả độ tập trung (trung tâm)

1. Trung bình
2. Trung vị
3. Mode

Các đại lượng mô tả độ tập trung (trung tâm)

1. Trung bình
2. Trung vị
3. Mode

Trung bình tổng thể (Mean population)

Định nghĩa

Trung bình tổng thể , ký hiệu là μ là trung bình của tất cả các quan sát trong tổng thể.

Ví dụ: Xét tổng thể gồm 5 sinh viên có chiều cao là: 1.75m, 1.68m, 1.59m, 1.80m, 1.74m, thì chiều cao trung bình của tổng thể 5 sinh viên là

$$\mu = \frac{1.75 + 1.68 + 1.59 + 1.80 + 1.74}{5} = 1.712m$$

Trung bình mẫu (Mean sample)

Định nghĩa

Giả sử một mẫu những quan sát gồm n giá trị (n gọi là cỡ mẫu) $x_1, x_2, \ldots, x_n$ *thì trung bình mẫu, ký hiệu* $\overline{x}$ *được tính:*

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

Ví dụ: Chọn ngẫu nhiên 5 sinh viên đã thi môn thống kê, điểm thi môn này của họ như sau: 7.5, 5.0, 8.5, 9.0, 4.0 thì trung bình của mẫu gồm 5 điểm thi này là:

$$\overline{x} = \frac{7.5 + 5.0 + 8.5 + 9.0 + 4.0}{5} = 6.8$$

Trung bình mẫu

Nhận xét: Trong trường hợp tập dữ liệu được tóm tắt trong bảng phân phối tần số không phân tổ,

Quan sát	x_1	x_2	$\cdots$	x_k
Tần số	f_1	f_2	$\cdots$	f_k

thì từ định nghĩa trên ta có công thức tính trung bình mẫu như sau

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{n}$$

Trong đó $x_1, x_2, \ldots, x_k$ là các giá trị của bảng tần số, $f_1, f_2, \ldots, f_k$ là tần số tương ứng và $n = f_1 + f_2 + \cdots + f_k$.

Trung bình mẫu

Khi tập dữ liệu mẫu đã được tóm tắt trong bảng tần số có phân tổ, công thức tính trung bình mẫu tương tự trường hợp không phân tổ trong đó các giá trị x_i được lấy là điểm giữa của tổ thứ i và là giá trị đại diện cho tổ đó.

Ví dụ: Xét lại ví dụ về 65 thời gian thanh toán ở phần 2

Ta có

$$\overline{x} = \frac{1}{65}\Big(3 \cdot 11.5 + 14 \cdot 14.5 + 23 \cdot 17.5 + 12 \cdot 20.5 + 8 \cdot 23.5 + 4 \cdot 26.5 + 1 \cdot 29.5\Big) = 18.61$$

Trung bình mẫu

Khi tập dữ liệu mẫu đã được tóm tắt trong bảng tần số có phân tổ, công thức tính trung bình mẫu tương tự trường hợp không phân tổ trong đó các giá trị x_i được lấy là điểm giữa của tổ thứ i và là giá trị đại diện cho tổ đó.

Ví dụ: Xét lại ví dụ về 65 thời gian thanh toán ở phần 2

Ta có

$$\overline{x} = \frac{1}{65}\Big(3 \cdot 11.5 + 14 \cdot 14.5 + 23 \cdot 17.5 + 12 \cdot 20.5 + 8 \cdot 23.5 + 4 \cdot 26.5 + 1 \cdot 29.5\Big) = 18.61$$

Trung bình mẫu

Thời gian(ngày)	Tần số	Tần suất(%)	Điểm giữa
10-13	3	4.62	11.5
13-16	14	21.54	14.5
16-19	23	35.38	17.5
19-22	12	18.46	20.5
22-25	8	12.31	23.5
25-28	4	6.15	26.5
28-31	1	1.54	29.5

Bảng: Bảng tần số 65 thời gian thanh toán

Trung vị (Median)

Định nghĩa

Xét một tổng thể hay một mẫu các quan sát mà các giá trị đã được xếp theo thứ tự tăng dần. Trung vị, ký hiệu là M_d được xác định như sau

1. *Nếu số các quan sát là số lẻ, trung vị là quan sát ở vị trí chính giữa tức là vị trí thứ* $\frac{n+1}{2}$.
2. *Nếu số quan sát là số chẵn, trung vị là trung bình cộng của hai quan sát ở vị trí thứ* $\frac{n}{2}$ *và* $\frac{n+2}{2}$.

Như vậy trung vị là giá trị chia tập dữ liệu thành hai phần bằng nhau.

Trung vị (Median)

Ví dụ: Một nhà sản xuất máy nghe nhạc chọn ngẫu nhiên 20 khách hàng đã dùng sản phẩm được một năm. Mỗi khách hàng trong mẫu được yêu cầu cho điểm về mức hài lòng của mình về sản phẩm theo thang điểm 10. Giả sử sau khi sắp xếp các điểm số của các khách hàng ta được kết quả như sau

1 3 5 5 7 8 8 8 8 **8** **8** 9 9 9 9 9 10
10 10 10

Vì số khách hàng trong mẫu là một số chẵn ($n = 20$) nên trung vị của mẫu này là trung bình cộng của hai điểm số ở vị trí giữa (vị trí thứ 10 và 11) và bằng 8. Và ta lấy 8 là một ước lượng điểm cho trung vị của điểm số đánh giá bởi tất cả các khách hàng.

Trung vị (Median)

Ví dụ: Chọn ngẫu nhiên 15 sinh viên và ghi lại điểm số môn thống kê xã hội học, sau khi xếp các điểm số theo thứ tự tăng dần ta được dãy sau:

2 3 3 5 6 6 6 **6** 7 7 8 8 8 9 10

Vì số sinh viên trong mẫu là một số lẻ ($n = 15$) nên trung vị của mẫu là số đo ở vị trị thứ 7 và bằng 6.

Mode

Định nghĩa

Mode, ký hiệu là M_0, của một tổng thể hay một mẫu các quan sát là quan sát có tần số lớn nhất.

Ví dụ: Trong cuộc khảo sát 200 người về số tờ báo mà họ đọc trong một tuần, người ta thu được bảng tổng kết như sau

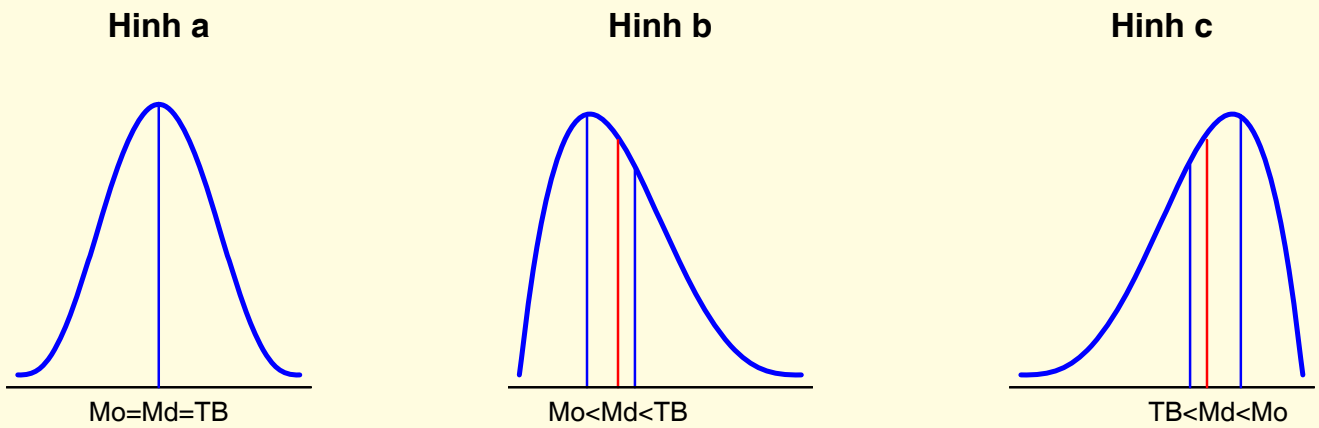
Số báo đọc	Tần số
0	44
1	24
2	18
3	16
4	20
5	22
6	26
7	30

Mode

Vậy giá trị mode $M_0 = 0$, có nghĩa là phổ biến nhất là trường hợp không đọc báo.

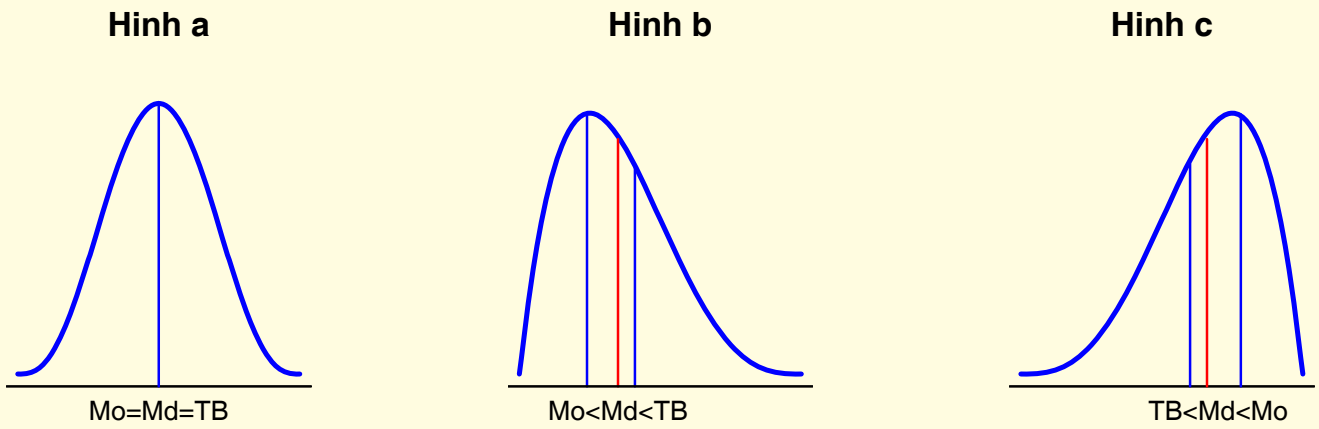
Chú ý: Khi tần số lớn nhất xảy ra ở hai số đo khác nhau thì ta nói tập dữ liệu nhị mode. Cũng có thể xảy ra trường hợp nhiều mode cùng tồn tại.

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 và trung vị của tập dữ liệu đó. Trong hình a, hình dáng của đa giác tần số đối xứng khi $\mu = M_d = M_0$.

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

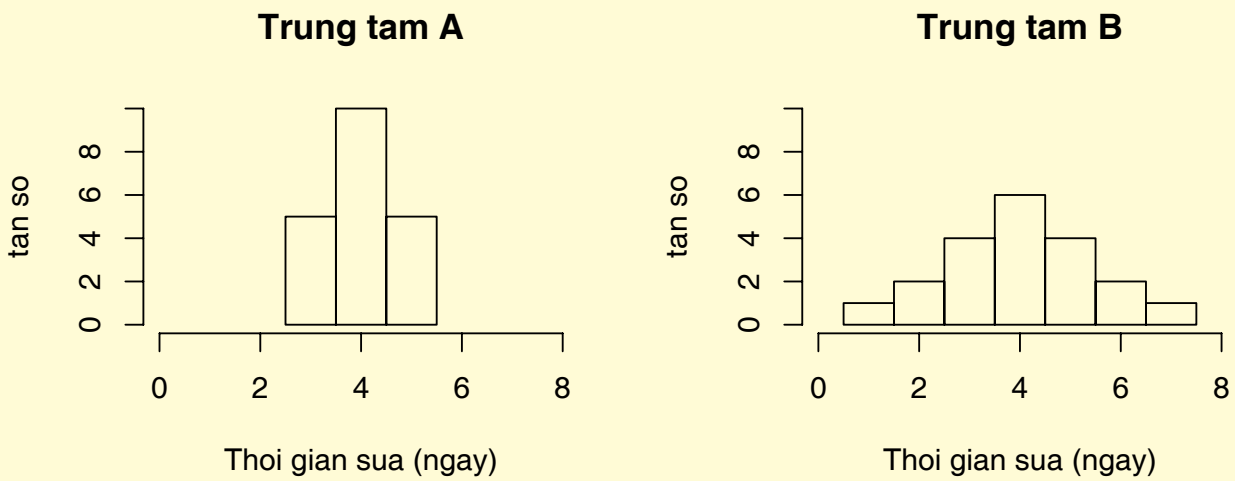
Các đại lượng mô tả độ tập trung (trung tâm)

Trung bình, trung vi, mode phản ánh độ tập trung hay trung tâm của tập dữ liệu. Tùy thuộc vào hình dáng phân phối của tập dữ liệu, người ta sử dụng những đại lượng thích hợp:

1. Sử dụng trung bình cho tập dữ liệu (định lượng) đối xứng
2. Sử dụng trung vị cho tập dữ liệu (định lượng) không đối xứng (lệch trái, lệch phải)
3. Sử dụng mode cho tập dữ liệu nhị mode

Các đại lượng đo độ phân tán

Để đánh giá xu hướng chính của một tổng thể, việc đánh giá mức độ biến đổi của các giá trị trong tổng thể là rất quan trọng.
Xét hai biểu đồ mô tả hai tập dữ liệu về thời gian sửa chữa máy của hai trung tâm dịch vụ khác nhau sau đây:



Các đại lượng đo độ phân tán

Có thể thấy, trung bình, trung vị và mode của hai mẫu đều là 4 nhưng thời gian sửa của trung tâm B thay đổi trong một phạm vi rộng hơn. Nên ta cần đo độ biến thiên để thấy sự khác biệt về phân phối của hai tập dữ liệu.

Các đại lượng đo độ phân tán

1. Khoảng biến thiên
2. Phương sai và độ lệch chuẩn

Các đại lượng đo độ phân tán

1. Khoảng biến thiên
2. Phương sai và độ lệch chuẩn

Khoảng biến thiên

Định nghĩa

Xét một tổng thể hay một mẫu các quan sát , **khoảng biến thiên**, *ký hiệu là R, là đại lượng được tính bằng cách lấy quan sát lớn nhất trừ đi quan sát nhỏ nhất.*

Trong trường hợp trung tâm A, khoảng biến thiên là $5 - 3 = 2$ còn trung tâm B là $7 - 1 = 6$.
Trong ví dụ này phạm vi biến thiên của của tập dữ liệu ứng với trung tâm B lớn hơn trung tâm A.

Phương sai và độ lệch chuẩn tổng thể

Định nghĩa

Phương sai tổng thể, *kí hiệu là* σ^2, *là giá trị trung bình cộng của các bình phương hiệu của các số đo và trung bình tổng thể.*
Độ lệch chuẩn tổng thể, σ *là căn bậc hai (số học) của phương sai tổng thể.*

Phương sai và độ lệch chuẩn tổng thể

Ví dụ: Xét lại ví dụ tổng thể gồm 5 sinh viên có chiều cao là: 1.75m, 1.68m, 1.59m, 1.80m, 1.74m.
Ta đã có trung bình tổng thể $\mu = 1.712m$.
Theo định nghĩa trên ta có phương sai tổng thể

$$\sigma^2 = \frac{1}{5}\Big[(1.75-1.712)^2 + (1.68-1.712)^2 + (1.59-1.712)^2 + (1.80-1.712)^2 + (1.74-1.712)^2\Big] = 0.0052$$

và do đó độ lệch chuẩn $\sigma = \sqrt{0.0052} = 0.072$.

Phương sai và độ lệch chuẩn mẫu

Phương sai và độ lệch chuẩn phản ánh mức độ phân tán của các giá trị quan sát trong tổng thể.
Khi tổng thể quá lớn, việc tính giá trị của hai đại lượng này rất khó khăn.
Thay vào đo ta sẽ ước lượng hai tham số này bởi phương sai và độ lệch chuẩn mẫu.

Phương sai và độ lệch chuẩn mẫu

Định nghĩa

Phương sai mẫu, *kí hiệu là* s^2, *được xác định bởi*

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Độ lệch chuẩn mẫu, $s = \sqrt{s^2}$ *là ước lượng điểm của độ lệch chuẩn tổng thể.*

Trong đó

- x_i *là giá trị quan sát thứ* i *của mẫu những quan sát*
- $\bar{x}$ *là trung bình mẫu*
- n *là cỡ mẫu*

Phương sai và độ lệch chuẩn mẫu

Chú ý:

1. Trường hợp mẫu những quan sát được tóm tắt trong bảng tần số không phân tổ, công thức tính phương sai như sau

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n-1} = \frac{(x_1 - \bar{x})^2 f_1 + \cdots + (x_k - \bar{x})^2 f_k}{n-1}$$

   Trong đó x_i là giá trị thứ i trong bảng tần số với f_i là tần số tương ứng, $f_1 + f_2 + \cdots + f_k = n$.

2. Trường hợp mẫu những quan sát được tóm tắt trong bảng tần số phân tổ, công thức tính phương sai tương tự như trên, trong đó x_i là điểm giữa của tổ thứ i.

Phương sai và độ lệch chuẩn mẫu

Ví dụ: Xét ví dụ về mẫu 65 thời gian thanh toán trong phần 1, bảng tần số sau khi phân tổ:

Thời gian(ngày)	Tần số	Tần suất(%)	Điểm giữa
10-13	3	4.62	11.5
13-16	14	21.54	14.5
16-19	23	35.38	17.5
19-22	12	18.46	20.5
22-25	8	12.31	23.5
25-28	4	6.15	26.5
28-31	1	1.54	29.5

Bảng: Bảng tần số 65 thời gian thanh toán

Phương sai và độ lệch chuẩn mẫu

Ta tính được trung bình mẫu $\overline{x} = 18.61$ và phương sai $s^2 = 15.91$ do đó độ lệch chuẩn $s = 3.99$.
(Trong khi nếu tính trực tiếp trên tập dữ liệu ban đầu thì $\overline{x} = 18.18$, phương sai $s^2 = 16.53$ và độ lệch chuẩn $s = 4.06$)

Phân vị, tứ phân vị, độ trải giữa, biểu đồ hộp và râu

1. Phân vị
2. Tứ phân vị
3. Độ trải giữa
4. Biểu đồ hộp và râu

Phân vị

Định nghĩa

Phân vị thứ p $(0 \leqslant p \leqslant 100)$ *trong một tập các quan sát được xếp theo thứ tự tăng dần, là giá trị sao cho* $p\%$ *số các quan sát nhỏ hơn hay bằng phân vị thứ p và* $(100 - p)\%$ *số các quan sát lớn hơn hay bằng phân vị thứ p.*

Phân vị

Có nhiều cách tìm phân vị thứ p trong một tập dữ liệu có n quan sát. Ta có thể tìm bằng cách sau đây:

1. Sắp xếp các giá trị quan sát theo thứ tự tăng dần (có thể dùng biểu diễn thân và lá)
2. Tính $i = \frac{p}{100} n$.
3. Nếu i không là một số nguyên thì $\lceil i \rceil$ là vị trí của phân vị thứ p trong dãy đã sắp xếp. Nếu i là một số nguyên thì phân vị thứ p là trung bình cộng của hai quan sát ở vị trí i và $i + 1$.

Phân vị

Ta có biểu diễn thân và lá của 65 thời gian thanh toán như sau

10	0
11	
12	00
13	000
14	0000
15	000000**0**
16	000000000
17	00000000
18	000000
19	00000
20	000
21	000
22	000
23	00
24	000
25	00
26	00
27	0
28	
29	0

Để tính phân vị thứ 25 ta tính
$i = (25/100)65 = 16.25$, do 16.25 không là số nguyên nên phân vị thứ 25 là quan sát ở vị trí 17 và bằng 15.

Phân vị

Để tính phân vị thứ 80 ta tính $i = (80/100)65 = 52$, do 52 là số nguyên nên phân vị thứ 80 là trung bình cộng của hai quan sát ở vị trí 52 và 53 và bằng 22. Do đó ta có thể ước lượng khoảng 80% trong tất cả các các thời gian thanh toán không quá 22 ngày.

10	0
11	
12	00
13	000
14	0000
15	0000000
16	000000000
17	00000000
18	000000
19	00000
20	000
21	000
22	000
23	00
24	000
25	00
26	00
27	0
28	
29	0

Tứ phân vị

Định nghĩa

- *Tứ phân vị thứ nhất, ký hiệu là Q_1, là phân vị thứ 25.*
- *Tứ phân vị thứ hai (hay là trung vị), ký hiệu là Q_2, là phân vị thứ 50.*
- *Tứ phân vị thứ ba, ký hiệu là Q_3, là phân vị thứ 75.*

Tứ phân vị

Ví dụ: Từ $i = (50/100).65 = 32.5$, tứ phân vị thứ hai của mẫu 65 thời gian thanh toán là $Q_2 = Md = 17$.
Và $i = (75/100).65 = 48.75$ ta có $Q_3 = 21$.

Độ trải giữa (khoảng tứ phân vị)

Định nghĩa

Đại lượng **độ trải giữa**, *ký hiệu là* R_Q *được tính bằng chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất.*

$$R_Q = Q_3 - Q_1$$

Với ví dụ trên ta có $R_Q = 21 - 15 = 6$. Vậy ta ước lượng rằng khoảng 50% quan sát ở giữa của tổng thể các thời gian thanh toán rơi vào khoảng thời gian từ 15 đến 21 ngày (có chiều dài là 6 ngày).

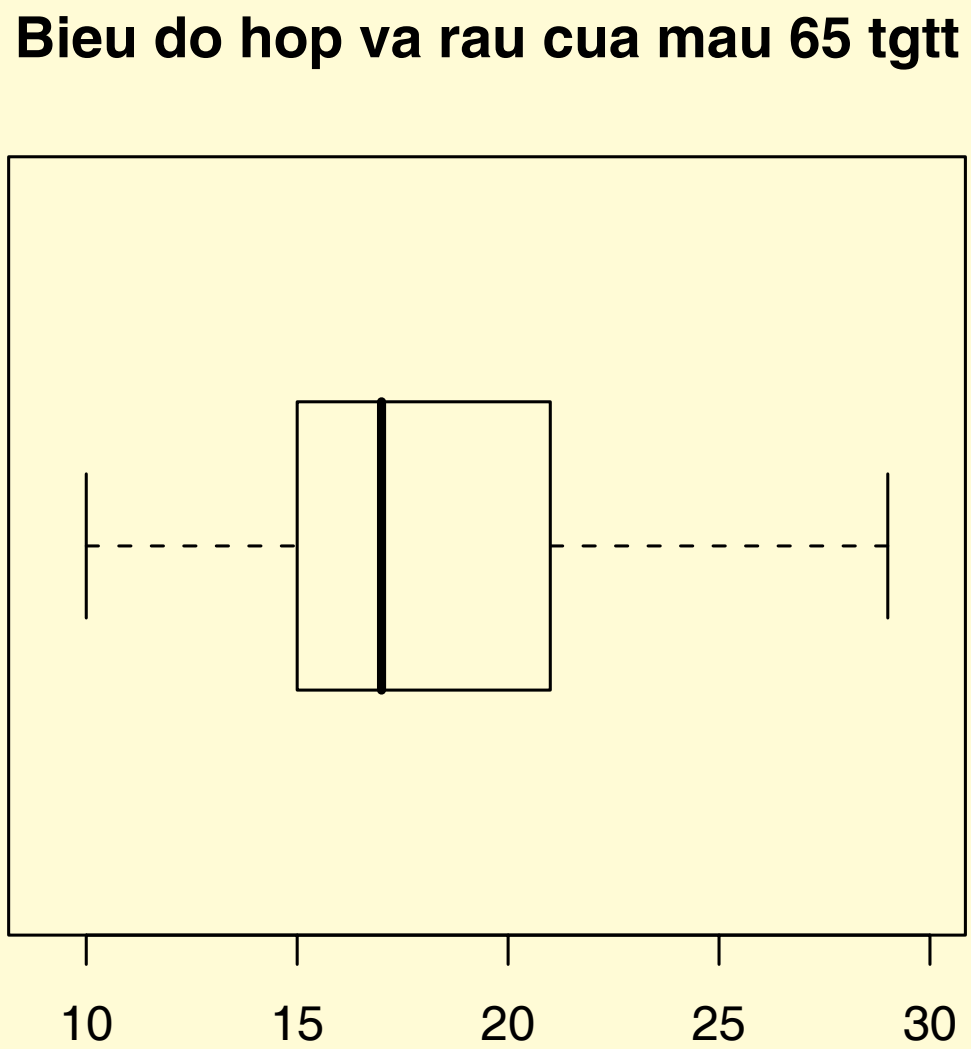
Các đại lượng đo độ phân tán

Khi muốn mô tả độ phân tán của một tập dữ liệu, ta có thể dùng một số gợi ý sau

1. Sử dụng độ lệch chuẩn khi dùng trung bình.
2. Sử dụng phân vị, độ trải giữa khi dùng trung vị.
3. Độ trải giữa được dùng khi muốn mô tả 50% số quan sát ở trung tâm của tập dữ liệu.
4. Khoảng biến thiên được dùng khi muốn nhấn mạnh vào các giá trị lớn nhất, nhỏ nhất của tập dữ liệu.

Biểu đồ hộp và râu (box and plot)

Đây là biểu đồ hộp và râu của mẫu 65 thời gian thanh toán



Biểu đồ hộp và râu (Box and plot)

Biểu đồ hộp và râu biểu diễn 5 số đo: giá trị nhỏ nhất, giá trị lớn nhất và các tứ phân vị thứ nhất, thứ hai, thứ ba.
Hộp hình chữ nhật thể hiện 50% các quan sát ở giữa tập dữ liệu, chiều rộng của hộp bằng độ trải giữa R_Q, hai cạnh bên đi qua giá trị tứ phân vị thứ nhất và thứ ba. Đường thẳng đứng bên 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.

Tác dụng của biểu đồ:

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

Biểu đồ hộp và râu (Box and plot)

Biểu đồ hộp và râu biểu diễn 5 số đo: giá trị nhỏ nhất, giá trị lớn nhất và các tứ phân vị thứ nhất, thứ hai, thứ ba.
Hộp hình chữ nhật thể hiện 50% các quan sát ở giữa tập dữ liệu, chiều rộng của hộp bằng độ trải giữa R_Q, hai cạnh bên đi qua giá trị tứ phân vị thứ nhất và thứ ba. Đường thẳng đứng bên 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.
Tác dụng của biểu đồ:

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

Sử dụng kết hợp trung bình và độ lệch chuẩn

1. Hệ số biến thiên
2. Quy tắc thực nghiệm
3. Quy tắc Chebyshev
4. Chuẩn hóa dữ liệu

Hệ số biến thiên

Để đo lường mức độ biến động tương đối của những tập dữ liệu có giá trị trung bình khác nhau người ta sử dụng **hệ số biến thiên**.
Hệ số biến thiên của một tổng thể các quan sát:

$$CV = \left(\frac{\sigma}{\mu}\right) \cdot 100\%$$

Hệ số biến thiên của một mẫu những quan sát:

$$CV = \left(\frac{s}{\overline{x}}\right) \cdot 100\%$$

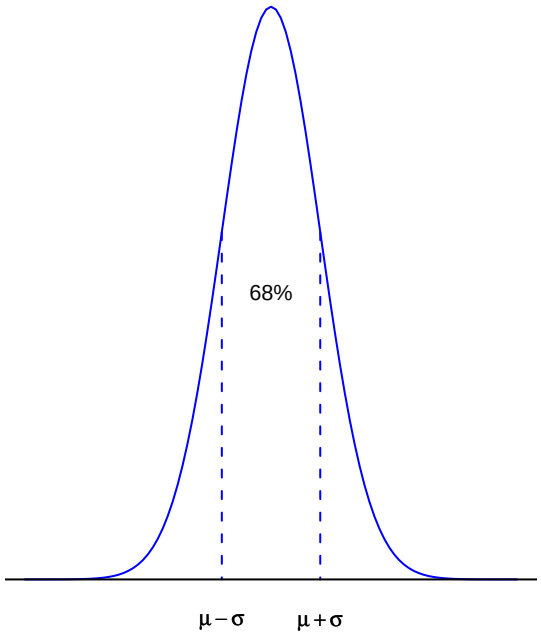
Khi so sánh hai tập dữ liệu, tập dữ liệu nào có hệ số biến thiên lớn hơn thì biến động nhiều hơn.

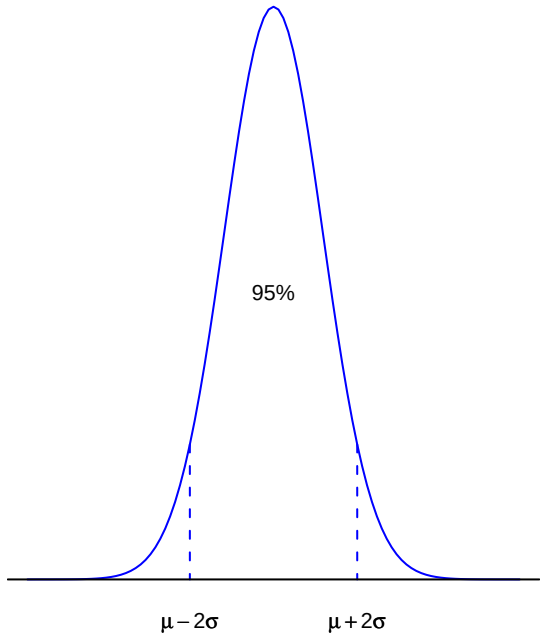
Quy tắc thực nghiệm

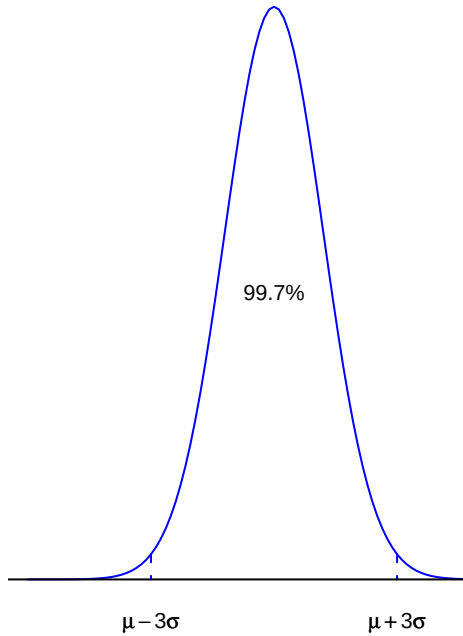
Nếu một tổng thể có trung bình μ và độ lệch chuẩn σ mà phân phối có dạng hình chuông cân đối thì

- Khoảng 68% số quan sát của tổng thể thuộc vào khoảng $[\mu - \sigma, \mu + \sigma]$.
- Khoảng 95% số quan sát của tổng thể thuộc vào khoảng $[\mu - 2\sigma, \mu + 2\sigma]$.
- Khoảng 99.7% số quan sát của tổng thể thuộc vào khoảng $[\mu - 3\sigma, \mu + 3\sigma]$.

Quy tắc thực nghiệm







Quy tắc thực nghiệm

Xét mẫu 49 lượng xăng tiêu thụ trong phần III, biểu diễn thân và lá của tập dữ liệu này cho thấy phân phối của mẫu có dạng hình chuông cân đối.

29	8
30	1344
30	5666889
31	001233444
31	55566777889
32	0001122344
32	556788
33	3

Quy tắc thực nghiệm

Ta tính được $\overline{x} = 31.5531$ và $s = 0.7992$. Ta lấy hai giá trị này ước lượng cho trung bình và độ lệch chuẩn tổng thể các lượng xăng tiêu thụ trên 500 km. Áp dụng quy tắc thực nghiệm ta có:

1. $[\overline{x} - s, \overline{x} + s] = [30.8, 32.4]$, ta ước lượng rằng khoảng 68% số xe tiêu thụ từ 30.8 đến 32.4 lít xăng trên 500 km.
2. $[\overline{x} - 2s, \overline{x} + 2s] = [30.0, 33.2]$, ta ước lượng rằng khoảng 95% số xe tiêu thụ từ 30.0 đến 33.2 lít xăng trên 500 km.
3. $[\overline{x} - 3s, \overline{x} + 3s] = [29.2, 34.0]$, ta ước lượng rằng khoảng 99.7% số xe tiêu thụ từ 29.2 đến 34 lít xăng trên 500 km.

Quy tắc Chebyshev

Định lí

(Định lý Chebyshev)
Xét một tổng thể có trung bình μ *và độ lệch chuẩn* σ*. Thì với mọi số* $k > 1$*, ít nhất* $100(1 - 1/k^2)\%$ *số các quan sát nằm trong khoảng* $[\mu - k\sigma, \mu + k\sigma]$.

Chuẩn hóa dữ liệu

Ta có thể xác định vị trí tương đối của một giá trị trong tổng thể hay mẫu bằng cách sử dụng trung bình và độ lệch chuẩn để tính giá trị chuẩn hóa z.

Cho x là một giá trị trong tổng thể hay mẫu, giá trị chuẩn hóa z của x được tính như sau

$$z = \frac{x - \text{trung bình}}{\text{độ lệch chuẩn}}$$

z còn được gọi là giá trị chuẩn hóa của x.

Nếu giá trị chuẩn hóa z của x là 3.1 thì x lớn hơn 3.1 lần độ lệch chuẩn so với trung bình. Nếu giá trị chuẩn hóa z của x là -1.7 thì x nhỏ hơn 1.7 lần độ lệch chuẩn so với trung bình.

Chuẩn hóa dữ liệu

Ta có thể xác định vị trí tương đối của một giá trị trong tổng thể hay mẫu bằng cách sử dụng trung bình và độ lệch chuẩn để tính giá trị chuẩn hóa z. Cho x là một giá trị trong tổng thể hay mẫu, giá trị chuẩn hóa z của x đươc tính như sau

$$z = \frac{x - \text{trung bình}}{\text{độ lệch chuẩn}}$$

z còn được gọi là giá trị chuẩn hóa của x.

Nếu giá trị chuẩn hóa z của x là 3.1 thì x lớn hơn 3.1 lần độ lệch chuẩn so với trung bình. Nếu giá trị chuẩn hóa z của x là -1.7 thì x nhỏ hơn 1.7 lần độ lệch chuẩn so với trung bình.

Chuẩn hóa dữ liệu

Ta có thể xác định vị trí tương đối của một giá trị trong tổng thể hay mẫu bằng cách sử dụng trung bình và độ lệch chuẩn để tính giá trị chuẩn hóa z. Cho x là một giá trị trong tổng thể hay mẫu, giá trị chuẩn hóa z của x đươc tính như sau

$$z = \frac{x - \text{trung bình}}{\text{độ lệch chuẩn}}$$

z còn được gọi là giá trị chuẩn hóa của x.
Nếu giá trị chuẩn hóa z của x là 3.1 thì x lớn hơn 3.1 lần độ lệch chuẩn so với trung bình. Nếu giá trị chuẩn hóa z của x là -1.7 thì x nhỏ hơn 1.7 lần độ lệch chuẩn so với trung bình.

Chuẩn hóa dữ liệu

Ví dụ: Xét điểm thi môn thống kê của hai nhóm sinh viên. Nhóm 1 có trung bình 65, độ lệch chuẩn 10. Nhóm 2 có trung bình 80, độ lệch chuẩn 5. Một sinh viên nhóm 1 có điểm thi là 85, và một sinh viên nhóm 2 có điểm thi là 90 sẽ có vị trí tương đối trong nhóm của mình như nhau vì chuẩn hóa z của 85 là $z_1 = (85 - 65)/10 = 2$
và chuẩn hóa z của 90 là $z_2 = (90 - 80)/5 = 2$.

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

Trong đó

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