Table of contents

Biến,tổng thể, mẫu

Biến và dữ liệu

- Biến (Variable): một đặc trưng mà thay đổi từ người hay vật, hiện tượng này sang người hay vật, hiện tượng khác. Biến gồm hai loại: biến định tính (quanlitative variable) và biến định lượng (quantitative variable).
- Biến định tính: biểu diễn tính chất của đặc trưng 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,tổng thể, mẫu

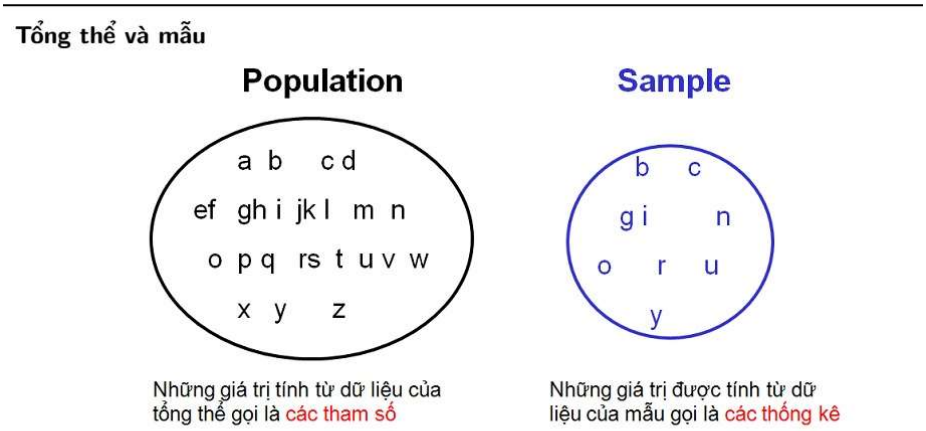
- Thông thường, biến rời rạc liên quan đến bài toán đếm số các phần tử của một tổng thể (số sản phẩm hỏng của một lô hàng, số cuộc điện thoại gọi đến tổng đài trong 1 giờ, số con trong gia đình . . .), trong khi biến liên tục liên quan đến sự đo đạc (cân nặng của 1 sản phẩm, chiều cao của 1 cây, cường 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).

Biến,tổng thể, mẫu

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 ký 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 mẫu.

Một số khái niệm cơ bản- Ví dụ về tổng thể



Một số khái niệm cơ bản- Ví dụ về tổng thể

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

Thông thường, ta không thể chọn hết 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

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).
- Mọi mẫu cỡ n có cùng khả năng được chọn từ tổng thể.

Phương pháp chọn mẫu ngẫu nhiên đơn giản (simple random sampling)

- Đánh số các phần tử của tổng thể từ 1 đến N. Lập các phiếu cũng đánh số như vậy.
- Trộn đều các phiếu, sau đó chọn có hoàn lại n phiếu. Các phần tử của tổng thể có số thự tự trong phiếu lấy ra sẽ được chọn làm mẫu.

Chọn mẫu ngẫu nhiên- Ví dụ

Ví dụ 1

Chọn một mẫu gồm $n = 15$ phần tử từ tập hợp 300 phần tử. Sử dụng chương trình thống kê R: dùng lệnh sample

Đánh số từ 1 đến 300:

```
P<-1:300
```

Chọn mẫu lần thứ nhất :

```
S1<-sample(P,15,rep=TRUE)
S1
```

Chọn mẫu lần thứ hai :

```
S2<-sample(P,15,rep=TRUE)
S2
```

Thống kê mô tả

Thống kê mô tả

Thống kê mô tả (Discriptive statisstics): là quá trình thu thập, tổng hợp và xử lý dữ liệu để 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 tham số mẫu như trung bình mẫu, phương sai mẫu, trung vị, ...

Thống kê suy luận

- Suy luận là một quá trình rút ra kết luận hoặc đưa ra quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- Thống kê suy luận (Inferential statistics): xử lý 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).
  - Ước lượng: ví dụ ước lượng tỷ lệ sản phẩm kém chất lượng trong 1 nhà máy; ước lượng trọng lượng trung bình sử dụng trung bình mẫu ...
  - Kiểm định gỉa thuyết : ví dụ cần kiểm định trọng lượng trung bình của 1 sản phẩm là 30kg.

Giới thiệu

- Dữ liệu dạng ban đầu (sơ khai, thô) rất khó để quan sát, nhận dạng, đánh giá. Do vậy, ta cần phải tổ chức lại dữ liệu.
- Các dạng tổ chức dữ liệu:
  - Bảng
  - Đồ thị
- Các dạng đồ thị được sử dụng sẽ phụ thuộc vào biến được tổng hợp.
- Các dạng đồ thị quan trọng thường dùng: biểu đồ thường(line chart), đồ thị stem-leaf, đồ thị tổ chức tần số (histogram), đồ thị phân tán (scatter plot).

Mô tả dữ liệu bằng đồ thị Stem-Leaf

Đồ thị Stem-Leaf

- Đồ thị 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.
- Đồ thị 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.

Mô tả dữ liệu bằng đồ thị Stem-Leaf (tt)

Đồ thị Stem-Leaf

Để xây dựng một đồ thị 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 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 lại các đơn vị cho các stem và leaf lên đồ thị.

Mô tả dữ liệu bằng đồ thị Stem-Leaf(tt)

Ví dụ 2 (Lập đồ thị stem-leaf)

- *Sắp xếp dữ liệu:*

21,24,24,26,27,27,30,32,38,41

- *Sử dụng đơn vị hàng chục cho đơn vị của stem*

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

Mô tả dữ liệu bằng đồ thị Stem-Leaf(tt)

Ví dụ 3

Lập đồ thị stem-leaf cho bộ dữ liệu sau
613,722,841,894,955,1047,1169,1224,1056,
982,1034,1140,933,928,906,863,859,891,827, 717,750,658,776,632

Sắp xếp dữ liệu tăng dần:

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

Mô tả dữ liệu bằng đồ thị Stem-Leaf(tt)

Sử dụng đơn vị hàng trăm cho stem

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

Ví dụ 4

Vẽ đồ thị stem-leaf cho tập dữ liệu 25 quan sát về các sản lượng từ một quá trình hóa học:

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

Bài giải 1

stem	leaves
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

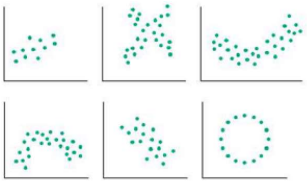
Bài giải 2 (tt)

Stem	Leaves
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
9L	2 3
9U	5

Stem	Leaves
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

Đồ thị phân tán

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



Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng.
- chứa các khoảng được phân nhóm theo dữ liệu quan trắc.
- và các tần số tương ứng của dữ liệu nằm trong từng khoảng.

Tại sao sử dụng phân phối tần số ?

- Phân phối tần số sử dụng để tổng hợp dữ liệu,
- biến đổi dữ liệu thô thành các thông tin có ích hơn,
- cho ta một cái nhìn trực quan để giải thích dữ liệu.

Phân phối tần số

Cách xây dựng

Các bước thực hiện:

- Đầu tiên, ta chia miền dữ liệu thành các khoảng(class intervals, cells, hoặc bins).
- Độ rộng của các khoảng bằng nhau.
- Số lượng các khoảng phụ thuộc vào số quan sát và độ phân tán của dữ liệu.
- Một phân phối tần số có quá ít hoặc quá nhiều khoảng đều không mang lại nhiều thông tin.
- Trong hầu hết mọi trường hợp, người ta thường chọn số khoảng từ 5-20. Trong thực tế, số lượng các khoảng có thể lấy xấp xỉ là căn bậc hai của số quan sát.

Ví dụ: Cách xây dựng phân phối tần số

Ví dụ 5

Chọn ngẫu nhiên 20 ngày mùa đông có nhiệt độ cao và đo nhiệt độ (đv: Độ F) được số liệu 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 phối tần số cho số liệu này.

Bài giải 3

Các bước thực hiện:

1. *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*
2. *Xác định miền dữ liệu (range): 58-12 =46*
3. *Chọn số khoảng cần chia: 5*
4. *Xác định độ rộng của khoảng: 10 (làm tròn 46/5)*
5. *Xác định biên của các khoảng: từ 10 đến 20; từ 20 đến 30; . . . , từ 50 đến 60.*
6. *Đếm số gía trị dữ liệu nằm trong mỗi khoảng.*

Bài giải

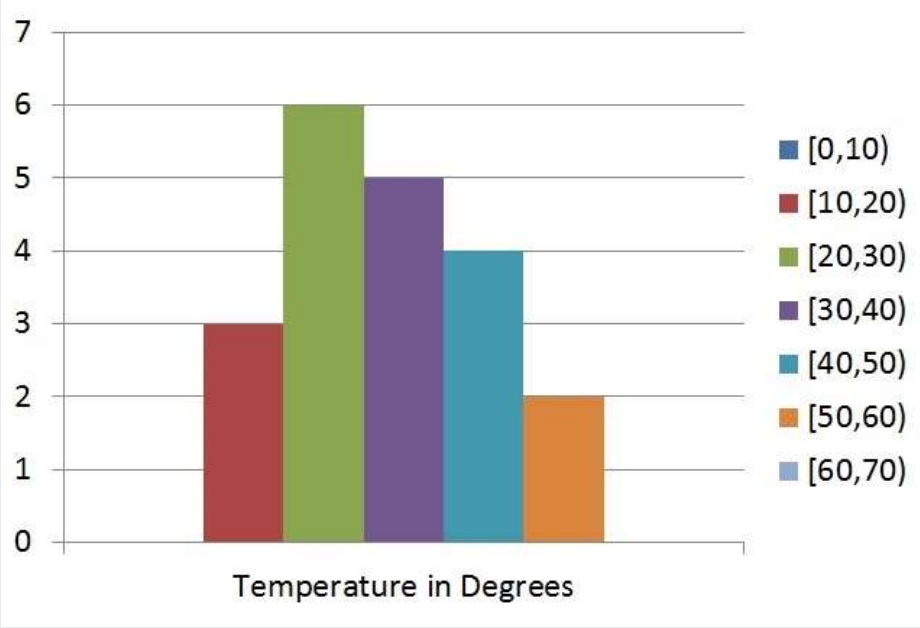
Dữ liệu được sắp xếp tăng dần:
12,13,17,21,24,24,26,27,27,30,32,35,37,38,41,43,44,46,53,58
Bảng phân phối tần số

Khoảng	Tần số	Tần số quan hệ	Phần trăm
[10;20)	3	0.15	15
[20;30)	6	0.3	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

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:

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



Hình : Đồ thị tần số

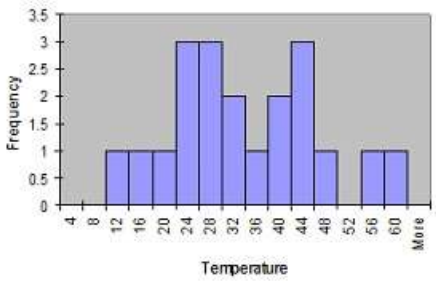
Nhận xét: Khi kích thước mẫu lớn, đồ thị tổ chức tần số phản ánh hình dạng của phân phối tổng thể. Hình dạng của phân phối có thể được xác định bởi đường cong trơn xấp xỉ đồ thị tổ chức tần số.

Xây dựng một phân phối tần số

Xây dựng một phân phối tần số

Câu hỏi: Chia dữ liệu thành bao nhiêu khaonrg là tốt? Chọn điểm cắt như thế nào cho phù hợp?

- Là quá trình "thử" và "sai",
- Mục tiêu lá tạo thành một phân phối không quá "lởm chơrm",có nhiều đỉnh và không có dạng "khối".
- Mục tiêu là chỉ ra được sự biến thiên trong dữ liệu



Đồ thị xác suất

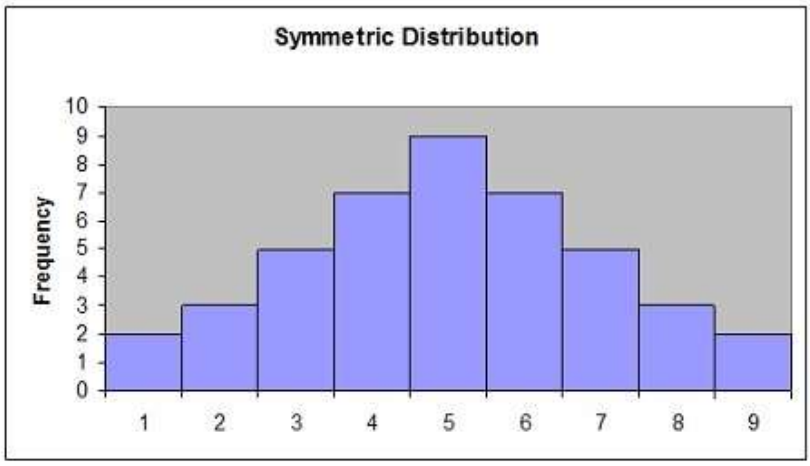
Bài toán 1

Nhiều kĩ thuật thống kê dựa trên giả sử rằng phân phối của tổng thể có một dạng cụ thể nào đó (chẳng hạn phân phối chuẩn). Vậy thì làm cách nào ta biết được dữ liệu thu được có phù hợp với giả sử này hay không?

- Các đồ thị mà ta đã học, như đồ thị tổ chức tần số, có thể cho ta biết được hình dạng của phân phối tổng thể.
- Tuy nhiên, thông thường, đồ thị tổ chức tần số chỉ xác định được rõ phân phối tổng thể nếu kích thước mẫu khá lớn. Đây là một khuyết điểm.

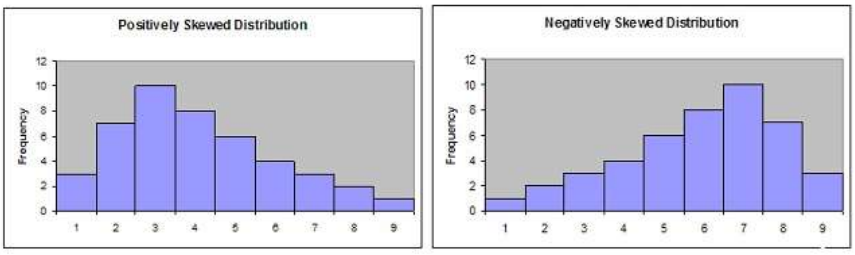
Hình dạng của phân phối

Hình dạng của phân phối (shape of the distribution) gọi là đối xứng (symmetric) nếu dữ liệu quan trắc cân bằng xung quanh trung tâm.



Hình dạng của phân phối

Hình dạng của phân phối (shape of the distribution) gọi là bất đối xứng (skewed) nếu dữ liệu quan trắc không phân bố đối xứng xung quanh trung tâm.



Đồ thị xác suất

Cách giải quyết

- Đồ thị xác suất là một dạng đồ thị khắc phục được nhược điểm này. Nó đáng tin cậy hơn đồ thị tổ chức tần số đối với các mẫu có kích thước nhỏ hoặc trung bình.
- Đồ thị xác suất thường sử dụng các trục đặc biệt với thang đo phù hợp với phân phối giả sử. Có nhiều phân phối như chuẩn, log chuẩn, Weibull, Chi bình phương, gamma. Tuy nhiên, ta chỉ tập trung chủ yếu vào đồ thị xác suất chuẩn vì nhiều kĩ thuật thống kê chỉ phù hợp khi tổng thể có phân phối chuẩn (hoặc ít nhất là xấp xỉ phân phối chuẩn).

Cách xây dựng

- Đầu tiên các quan sát trong mẫu được sắp xếp theo thứ tự tăng dần. Các quan sát được sắp x_j khi đó được vẽ cùng tần số tích lũy quan sát $(j - 0.5)/n$ trên một giấy xác suất phù hợp.
- Nếu phân phối được gỉa sử phù hợp với dữ liệu, thì các điểm được vẽ sẽ xấp xỉ một đường thẳng; nếu các điểm cách khỏi một đường thẳng, thì mô hình giả sử không phù hợp. Thông thường, việc xác định đồ thị dữ liệu có thẳng hàng hay không mang tính chủ quan.

Ví dụ 6

10 quan trắc về thời gian hoạt động(theo phút) của pin được sử dụng trong các laptop như sau : 176,191,214,229,205,192,201,190,183,185. Ta giả sử rằng thời gian hoạt động của pin tuân theo phân phối chuẩn. Sử dụng đồ thị xác suất để kiểm chứng giả sử này ?

Mô tả dữ liệu số

Độ đo trung tâm - Trung bình

Trung bình(mean) là đại lượng thường được sử dụng nhất để đo giá trị trung tâm của dữ liệu.

Định nghĩa 1

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

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

Độ đo trung tâm - Trung bình

Định nghĩa 2

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

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

Nhận xét 1

Trung bình : bị ảnh hưởng bởi các ngoại lai (outlier).

Độ đo trung tâm - Trung vị mẫu

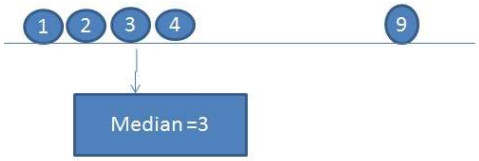
Định nghĩa 3

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 2

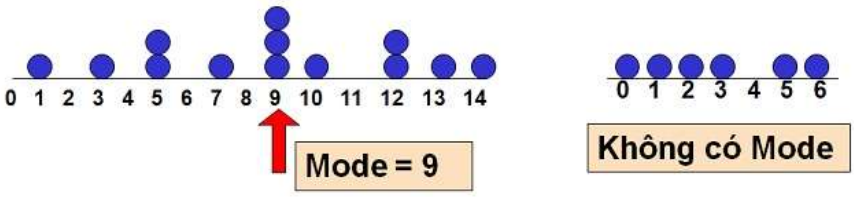
- *Trung vị không bị ảnh hưởng bởi các outlier.*
- *Vị trí của trung vị: sắp xếp dữ liệu tăng dần, gọi i là vị trí trung vị :* $i = \frac{n+1}{2}$
  - *Nếu* $i \in \mathbb{Z}$ *trung vị* $= x_i$
  - *Nếu* $i \notin \mathbb{Z}$ *trung vị* $= \frac{x_{[i]}+x_{[i]+1}}{2}$, *với [i] là phần nguyên của i.*

Độ đo trung tâm - Trung vị mẫu



Độ đo trung tâm - Mode

- Mode là một đại lượng để đo xu hướng trung tâm của dữ liệu,
- là giá trị thường xảy ra nhất,
- Không bị ảnh hưởng bởi các outlier
- Có thể sử dụng cho cả dữ liệu số và dữ liệu phân loại
- Có thể có nhiều mode hoặc không tồn tại mode

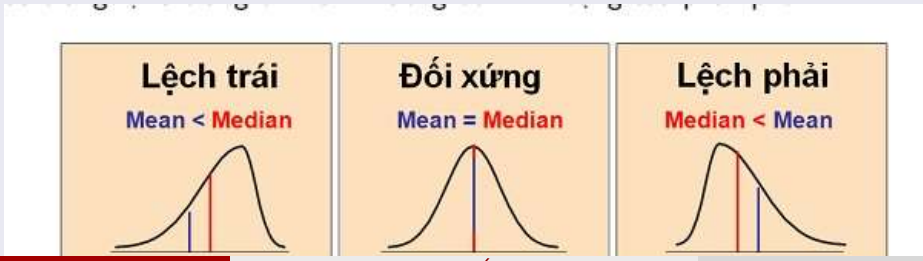


MÔ tả dữ liệu | Mô tả dữ liệu số
Dữ liệu thống kê | Ngày 25 tháng 11 năm 2016 | 30 / 37

Hình dung trung bình, trung vị, mode từ đồ thị tổ chức tần số

Ta có thể xác định vị trí tương đốii của trung bình, trung vị và mode từ đồ thị tổ chức tần số như sau:

- Nếu phân phối tần số là đối xứng, thì trung bình = trung vị =mode.
- Nếu phân phối tần số bị lệch (skewed) (tức là bất đối xứng, với một đuôi 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.



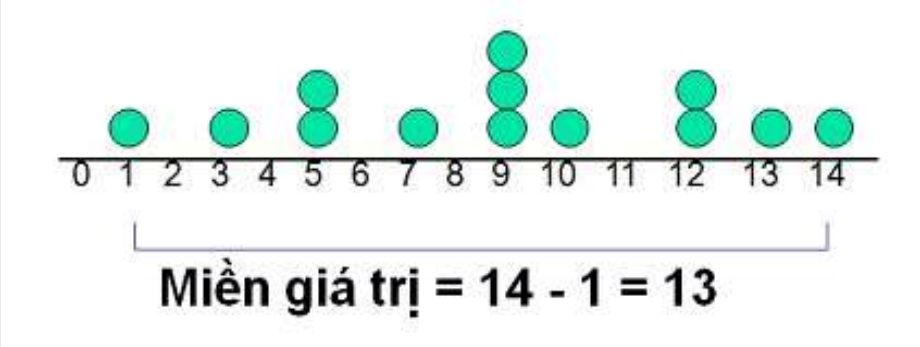
Sự biến thiên- Miền giá trị

Định nghĩa 4

- *Miền giá trị (range) là độ đo sự biến thiên đơn giản nhất.*
- *Là độ lệch giữa giá trị lớn nhất và bé nhất của dữ liệu quan trắc.*

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

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



Tứ phân vị, phân vị

Định nghĩa 5 (Tứ phân vị)

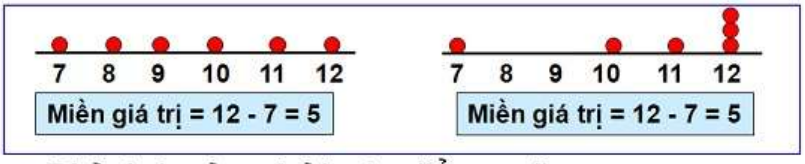
Nếu ta chia dữ liệu thành 4 phần bằng nhau. Các điểm chia này được gọi là các tứ phân vị (quartiles). Tứ phân vị đầu tiên, q_1 là giá trị xấp xỉ 25 % số quan sát nằm bên dưới nó và xấp xỉ 75 % số quan sát nằm trên nó. Tứ phân vị thứ hai, q_2, có xấp xỉ 50 % số quan sát nằm bên dưới nó và xấp xỉ 50 % số quan sát nằm trên nó, tứ phân vị thứ hai chính là trung vị. Tứ phân vị thứ ba, q_3, có xấp xỉ 75 % số quan sát nằm bên dưới nó.

Định nghĩa 6 (Phân vị)

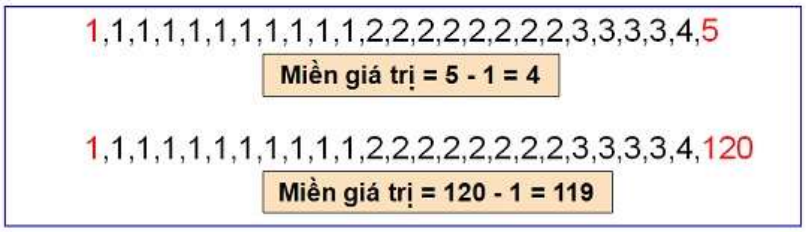
Phân vị mức $100 \times k$ là giá trị dữ liệu sao cho xấp xỉ $100 \times k\%$ số quan sát nằm dưới giá trị này và có xấp xỉ $100 \times (1 - k)$số quan sát nằm trên giá trị này.

Nhược điểm của miền tứ phân vị

- Bỏ qua phân bố của dữ liệu



- Bị ảnh hưởng bởi các điểm oulier



Miền tứ phân vị

Ta có thể loại bỏ các điểm ngoại lai (outlier) bằng cách sử dụng Miền phân vị.

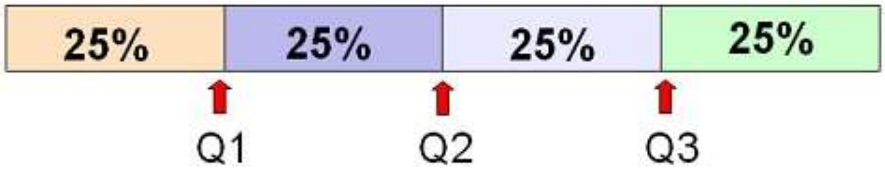
Định nghĩa 7

Miền phân vị IQR

$$IQR = Q_3 - Q_1$$

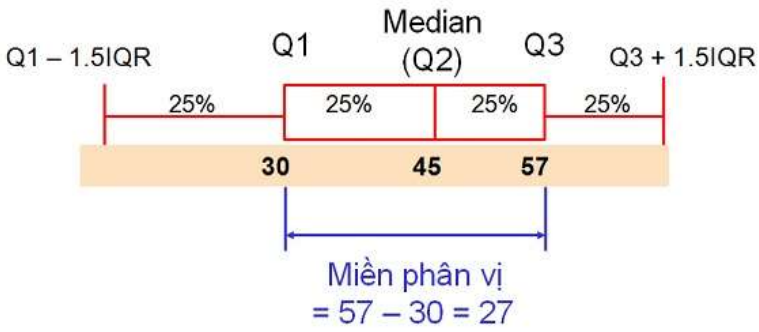
trong đó, Q_1 là mức phân vị thứ nhất (mức phân vị 25 %), Q_3 là mức phân vị thứ ba (mức phân vị 75 %).

Khi đó, oulier $< Q_1$, hoặc oulier $> Q_3$.



Đồ thị boxplot

■ Để biểu diễn miền phân vị và các điểm outlier ta dùng đồ thị boxplot,



Cách tìm tứ phân vị

Sắp xếp dữ liệu (khích thước n) theo thứ tự tăng dần. Gọi q_1, q_2, q_3 lần lượt là phân vị thứ nhất, hai, ba của dữ liệu và

$$k_1 = 0.25(n+1)$$
$$k_2 = 0.5(n+1)$$
$$k_3 = 0.75(n+1)$$

Khi đó,

$$q_i = \begin{cases} x_{k_i} & \text{nếu } k_i \text{ nguyên} \\ \frac{x_{[k_i]} + x_{[k_i]+1}}{2} & \text{ngược lại} \end{cases}$$

i=1,2,3

Sự biến thiên- Phương sai

Phương sai (Variance) là trung bình của bình phương độ lệch các giá trị so với trung bình. Phương sai phản ánh độ phân tán hay sự biến thiên của dữ liệu.

Định nghĩa 8 (Phương sai tổng thể)

Phương sai tổng thể

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

Với N là số phần tử của tổng thể.
Độ lệch chuẩn tổng thể là $\sigma = \sqrt{\sigma^2}$.

Sự biến thiên- Phương sai

Định nghĩa 9 (Phương sai mẫu)

Phương sai mẫu gồm n quan trắc là

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

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

Định lí Chebyshev

Định lý 1

Với một tổng thể bất kì có trung bình μ *và độ lệch chuẩn* σ, *và* $k > 1$, *phần trăm các giá trị quan trắc nằm trong khoảng*

$$[\mu - k\sigma, \mu + k\sigma]$$

bằng ít nhất $100\left[1 - \frac{1}{k^2}\right]\%$.

Ví dụ 7

ít nhất	*nằm trong*
$(1 - 1/1^2) = 0\%$	$(\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	$(\mu \pm 3\sigma)$

Quy tắc thực nghiệm

Quy tắc thực nghiệm (The Empirical Rule): nếu dữ liệu có phân phối chuẩn (hoặc tiệm cận chuẩn), thì khoảng

- $[\mu - 1\sigma, \mu + 1\sigma]$ chứa khoảng 68% giá trị dữ liệu của mẫu hoặc tổng thể.
- $[\mu - 2\sigma, \mu + 2\sigma]$ chứa khoảng 95% giá trị dữ liệu của mẫu hoặc tổng thể.
- $[\mu - 3\sigma, \mu - 3\sigma]$ chứa khoảng 99,7% giá trị dữ liệu của mẫu hoặc tổng thể.

Hệ số biến thiên

- Hệ số biến thiên được sử dụng để so sánh sự biến thiên của hai hay nhiều tập dữ liệu, có thể đo ở các đơn vị khác nhau.
- Đo mối liên hệ giữa sự biến thiên và trung bình.
- Đơn vị tính: %
- Công thức tính

$$CV = \frac{s}{\bar{x}} 100\%$$

Hệ số biến thiên

Ví dụ 8

- *Dữ liệu A có trung bình* $\bar{x}_A = 50$, *độ lệch chuẩn* $s_A = 5$

$$CV_A = \frac{s_A}{\bar{x}_A} 100\% = 10\%$$

- *Dữ liệu B có trung bình* $\bar{x}_B = 100$, *độ lệch chuẩn* $s_B = 5$

$$CV_B = \frac{s_B}{\bar{x}_B} 100\% = 5\%$$

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

Mẫu ngẫu nhiên

Định nghĩa 10

Các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ *là một mẫu ngẫu nhiên kích thước n nếu*

i X_i *là các biến ngẫu nhiên độc lập nhau.*

ii *Mọi* X_i *đều có cùng một phân phối xác suất.*

Thống kê

Định nghĩa 11

Một thống kê (statistic) là một hàm bất kì các quan sát trong một mẫu ngẫu nhiên.

- Trung bình mẫu: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Phương sai mẫu: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2$
- Độ lệch chuẩn mẫu: $S = \sqrt{S^2}$
- Giá trị nhỏ nhất của mẫu: $Y_1 = \min(X_1, X_2, \ldots, X_n)$
- Giá trị lớn nhất của mẫu: $Y_2 = \max(X_1, X_2, \ldots, X_n)$
- $R = Y_n - Y_1$

đều là các thống kê của mẫu ngẫu nhiên có kích thước n.

Phân phối mẫu

Bởi vì một thống kê là một biến ngẫu nhiên, nên nó có phân phối xác suất.

Định nghĩa 12

Phân phối xác suất của một thống kê được gọi là một phân phối mẫu.

Ví dụ 9

Phân phối xác suất của $\bar{X}$ được gọi là phân phối mẫu của trung bình

Nhận xét 3

Phân phối mẫu của một thống kê phụ thuộc vào phân phối của tổng thể, kích thước mẫu, và phương pháp chọn mẫu.

Phân phối mẫu của trung bình và phương sai

Định lý 2

Giả sử $(X_1, X_2, \ldots, X_n)$ *là một mẫu ngẫu nhiên được lấy từ một tổng thể có phân phối chuẩn với trung bình* μ *và phương sai* σ^2. *Khi đó,*

- $\bar{X}$ *và* S^2 *độc lập nhau.*
- $\bar{X} \sim N(\mu, \sigma^2/n)$.
- $\frac{(n-1)S^2}{\sigma} \sim \chi^2(n-1)$.
- $\frac{\bar{X}-\mu}{S\sqrt{n}} \sim T(n-1)$.

Phân phối mẫu của trung bình và phương sai

Trường hợp tổng thể có phân phối xác suất chưa biết

Trong trường hợp này, định lí giới hạn trung tâm khẳng định rằng $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$ $\frac{\bar{X}-\mu}{S/\sqrt{n}} \to N(0,1)$ Trong thực hành khi mẫu có kích thước đủ lớn $(n \geq 30)$, ta có các phân phối xấp xỉ chuẩn sau:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1);$$

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \approx N(0,1)$$

Phân phối mẫu của tỉ lệ

Giả sử cần khảo sát đặc trưng $\mathcal{A}$ của tổng thể, khảo sát n phần tử và đặt

$$X_i = \begin{cases} 1 & \text{nếu thỏa } \mathcal{A} \\ 0 & \text{nếu khác} \end{cases}$$

thu được mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$,với $X_i \sim B(1, p)$ trong đó p là tỉ lệ phần tử thỏa đặc trưng $\mathcal{A}$.

Đặt $X = \sum_{i=1}^{n} X_i$ là số phần tử thỏa đặc trưng $\mathcal{A}$ trong mẫu khảo sát, thì $X \sim B(n, p)$.

Tỉ lệ mẫu $\hat{p}$ là một ước lượng của tỉ lệ p được xác định bởi

$$\hat{p} = \frac{X}{n}$$

Phân phối mẫu của tỉ lệ

Kì vọng và phương sai của $\hat{p}$ là

$$\mathbb{E}(\hat{p}) = p, \ \mathbb{V}ar = \frac{p(1-p)}{n}$$

Theo định lí giới hạn trung tâm ta có

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0, 1).$$

Vì vậy trong thực hành, khi $np \geq 5, \ n(1-p) \geq 5$ thì $\hat{p} \approx N(p, \frac{p(1-p)}{n})$.

Qua bài giảng SV hiểu/ có thể áp dụng...

1. Một số khái niệm cơ bản: biến, dữ liệu, tổng thể, mẫu; chọn mẫu ngẫu nhiên, thống kê mô tả, thống kê suy luận
2. Mô tả dữ liệu
   - bằng đồ thị: đồ thị stem-leaf, đồ thị phân tán, đồ thị tổ chức tần số, đồ thị xác suất.
   - bằng các đặc trưng số của dữ liệu
     - Độ đo sự hướng tâm : trung bình, trung vị, mode.
     - Độ đo sự biến thiên: Miền giá trị, phân vị, tứ phân vị, khoảng tứ phân vị, định lí Chebyshev, hệ số biến thiên.
3. Phân phối mẫu: Mẫu ngẫu nhiên, thống kê, phân phối mẫu, một số phân phối mẫu cơ bản.