Phân tích số liệu thống kê

Đặng Hải Vân – Lê Phong – Nguyễn Đình Thúc
Khoa CNTT – ĐHKHTN
{dhvan,lphong,ndthuc}@fit.hcmus.edu.vn

Nội dung

- Phân tích dữ liệu mang tính khám phá (EDA)
- Lấy mẫu
  - Khái niệm
  - Lấy mẫu
  - Lấy mẫu xác suất
  - Xử lý mẫu
- Thống kê mô tả
  - Khái niệm
  - Các giá trị thống kê mô tả
  - Các kỹ thuật biểu diễn đồ thị
    - Histogram
    - Boxplot
    - Quantile-based plot
    - Scatter plot

Phân tích dữ liệu mang tính khám phá

- Phân tích dữ liệu mang tính khám phá (EDA – Exploratory Data Analysis) [John Tukey, 1977]
  - Dữ liệu nên được xem xét, khám phá trước khi đặt ra bất kỳ giả thuyết nào về mô hình xác suất, mối quan hệ giữa các biến,…
  - Hướng tiếp cận:

  - Đặc trưng: phụ thuộc nhiều vào các kỹ thuật biểu diễn đồ thị (graphical techniques)

Khái niệm

- Quần thể: là nhóm đối tượng cần tổng quát hóa
- Mẫu: là nhóm đối tượng thực sự được chọn trong nghiên cứu.
- Cơ cấu mẫu: là danh sách các phần tử của quần thể có thể.
- Thống kê (statistic): là hàm của các quan sát trong tập mẫu

  Ví dụ: trung bình mẫu, phương sai mẫu là các thống kê

Lấy mẫu

- Lấy mẫu: là một tiến trình chọn các mẫu cho mục đích tổng quát hóa.
  - Lấy mẫu xác suất (probability sampling)
    - Thủ tục chọn ngẫu nhiên . Xác suất các phần tử được chọn bằng nhau.
    - Mẫu chọn được gọi là mẫu ngẫu nhiên
  - Lấy mẫu không xác suất (non-probability sampling)

Ví dụ: UBND quận 5 thực hiện khảo sát lấy ý kiến của nhân dân quận 5 về tình hình trị an hiện tại của quận.

Cách khảo sát 1: tất cả hộ gia đình của quận đều có cơ hội được chọn và hỏi qua điện thoại. Xác suất 1 hộ gia đình được hỏi là xác định được. – Lấy mẫu xác suất

Cách khảo sát 2: Bảng câu hỏi được gửi đến các cư dân trong quận dựa vào 1 mailing list đã có sẵn. Ngoài ra các bảng câu hỏi được đặt ở các nơi công cộng. Theo cách này, không xác định được 1 cá nhân có thể trả lời bao nhiêu lần. Xác suất 1 cá nhân được hỏi là không xác định được. – Lấy mẫu không xác suất

Lấy mẫu xác suất

- Lấy mẫu ngẫu nhiên đơn giản (simple random sampling)
  - Chọn n từ cơ cấu mẫu N phần tử sao cho ${}_NC_n$ phần tử có cơ hội chọn ngang nhau
  - Kỹ thuật chọn
    - Bảng ngẫu nhiên với tỷ lệ mẫu: f = n/N

      Phát sinh số mầm s → mẫu: nhãn là s+i.n với i=0,1,…,1/f-1
    - Số ngẫu nhiên

      Phát sinh số ngẫu nhiên → mẫu: nhãn trùng với số ngẫu nhiên
- Lấy mẫu ngẫu nhiên phân tầng (stratified random sampling)
  - Nhóm thuần nhất (stratum, strata)
  - Lấy mẫu ngẫu nhiên đơn giản trong từng nhóm
  - Tỷ lệ mẫu cho từng nhóm $f_1, f_2, \ldots$
- Lấy mẫu ngẫu nhiên theo cụm (cluster sampling), lấy mẫu ngẫu nhiên một cách hệ thống (systematic random sampling) [Levy & Lemeshow, 1999]

Ví dụ lấy mẫu ngẫu nhiên đơn giản

Cần khảo sát các khách hàng của công ty, biết danh sách khách hàng gồm N=1000. Ta lấy mẫu gồm 100 khách hàng để thực hiện khảo sát (n=100)

a) Số mầm phát sinh s = 5. Xác định mẫu.

b) Dùng R để phát sinh mẫu theo phương pháp số ngẫu nhiên.

Trả lời:

a) f=100/1000 = 0,1. Mẫu là danh sách các phần tử ở các vị trí: 5,5+1.100, 5+2.100,…,5+9.100 hay 5, 105, 205,…,905

b)

```
> N<-1000
> n<-100
> sample(1:N,n,replace=FALSE)
```

Ví dụ lấy mẫu ngẫu nhiên phân tầng

- Cần khảo sát ý kiến của SV các khoa toán (gồm 200sv), cntt (gồm 500sv), lý (gồm 300sv), ta chọn một tập mẫu gồm n=100 sv.

Phân chia thành các nhóm thuần nhất: sv khoa toán (n_1=20), sv khoa cntt (n_2=50), sv khoa lý (n_3=30) (f_1=f_2=f_3=0,1)

Lấy mẫu ngẫu nhiên đơn giản cho từng nhóm.

- EDA
- Lấy mẫu
  - Khái niệm
  - Lấy mẫu
  - Lấy mẫu xác suất
  - Xử lý mẫu
- Thống kê mô tả
  - Khái niệm
  - Các giá trị thống kê mô tả
  - Các kỹ thuật biểu diễn đồ thị
    - Histogram
    - Boxplot
    - Quantile-based plot
    - Scatter plot

Bài tập lấy mẫu ngẫu nhiên đơn giản với R

sample(x, size, replace = FALSE, prob = NULL)
replace = TRUE: lấy mẫu có lặp lại
replace=FALSE: lấy mẫu không lặp lại

BT1: Dùng R chọn ngẫu nhiên 5 số từ 1 đến 40

BT2: Dùng R giả lập thí nghiệm tung đồng xu 10 lần

BT3: Dùng R giả lập thí nghiệm tung đồng xu 10 lần, biết khả năng tung mặt ngửa là 90%, mặt sấp là 10%

Trả lời:
BT2: sample(c("H","T"),10,replace=TRUE)
Lưu ý:
Dữ liệu vector: là một mảng
Khởi tạo vector:
1) Bằng cách nối kết: c(phần tử 1, phần tử 2,…). Vd: c("H","T"): tạo vector 2 phần tử
2) 1:10: tạo mảng từ 1 đến 10

Xử lý mẫu

- Khái niệm Giá trị bất thường (giá trị ngoại lệ): là giá trị có sự sai lệch quá rõ ràng so với các giá trị khác.
  - Phát hiện mẫu bất thường
  - Xử lý mẫu bất thường

Khái niệm thống kê mô tả

- Khái niệm Thống kê mô tả: là phương pháp thống kê toán được dùng để mô tả các đặc trưng cơ bản của dữ liệu, cung cấp tóm tắt cô đọng về mẫu và các thước đo.

Các thống kê mô tả

Biến ngẫu nhiên X, tập mẫu gồm n phần tử $\{x_i\}$, i=1,…,n

- Moment thứ r của mẫu: $\mu_r' = E[X^r] \Rightarrow \mu_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r$

Moment trung tâm: $\mu_r = E[(X-\mu)^r] \Rightarrow \mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^r$

- Trung bình mẫu (sample mean, sample expected value): mô tả khuynh hướng của tâm dữ liệu $\mu = E[X] = \sum x_i f(x_i) \text{ hay } \int x f(x)dx \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

- Phương sai mẫu (sample variance)

$$\sigma^2 = \text{var}(X) = E[(X-\mu)^2] = \sum (x_i-\mu)^2 f(x_i) \text{ hay } \int (x-\mu)^2 f(x)dx \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \text{ hay } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$$

Các thống kê mô tả (tt)

- Trị SKEW: diễn tả tính bất đối xứng của phân phối dữ liệu quanh trị trung bình: skew<0 (lệch trái), skew>0 (lệch phải), skew=0 (đối xứng)

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \Rightarrow \gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)^{3/2}}$$

- Tri KURT: diễn tả độ phẳng của đỉnh phân phối dữ liệu: kurt<3 (bằng), kurt>3 (nhọn), kurt=3 (vừa phải, hình chuông)

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)^2}$$

Các thống kê mô tả (tt)

- Trung vị (median): là điểm nằm chính giữa dãy dữ liệu. $median = \begin{cases} x_{(n+1)/2} & n \bmod 2 = 1 \\ (x_{n/2} + x_{n/2+1})/2 & n \bmod 2 = 0 \end{cases}$
- Yếu vị (mode): là giá trị có tần số xuất hiện cao nhất trong tập dữ liệu.
- Độ phân tán: biểu diễn sự phân tán các giá trị quanh tâm dữ liệu
  - Khoảng quan sát (range): *range = Max – Min*
  - Độ lệch chuẩn
- Phân vị (quantile): phân vị q_p là giá trị q nhỏ nhất sao cho phân phối tích lũy của nó lớn hơn hoặc bằng p, với 0<p<1

$$F(q_p) = P[X \le q_p] = p \text{ hay } q_p = F^{-1}(p)$$

Ví dụ: $q_{0,25}$,$q_{0,5}$,$q_{0,75}$: các phần tư vị (quartile)

Ví dụ tính giá trị thống kê

- Cho tập mẫu X={1,3,2,4,6,2,2,5,6}
- Tính
  - Trung bình mẫu $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1+3+2+4+6+2+2+5+6}{9} = \frac{31}{9} = 3{,}44$
  - Phương sai mẫu

$$\sigma^2 = \frac{1}{9} \times \frac{(1-3{,}44)^2 + (3-3{,}44)^2 + (2-3{,}44)^2 + (4-3{,}44)^2 + (6-3{,}44)^2 + (2-3{,}44)^2 + (2-3{,}44)^2 + (5-3{,}44)^2 + (6-3{,}44)^2}{9} = 3{,}14$$

$$S^2 = \frac{1}{9-1} \times \frac{(1-3{,}44)^2 + (3-3{,}44)^2 + (2-3{,}44)^2 + (4-3{,}44)^2 + (6-3{,}44)^2 + (2-3{,}44)^2 + (2-3{,}44)^2 + (5-3{,}44)^2 + (6-3{,}44)^2}{9} = 3{,}53$$

  - Trị skew

$$\mu_3 = \frac{1}{9} \times \frac{(1-3{,}44)^3 + (3-3{,}44)^3 + (2-3{,}44)^3 + (4-3{,}44)^3 + (6-3{,}44)^3 + (2-3{,}44)^3 + (2-3{,}44)^3 + (5-3{,}44)^3 + (6-3{,}44)^3}{9} = 1{,}551$$

$$\mu_2 = \sigma^2 = 3{,}14 \Rightarrow skew = \frac{\mu_3}{\mu_2^{3/2}} = \frac{1{,}551}{3{,}14^{3/2}} = 0{,}279$$

  - Trị kurt

$$\mu_4 = \frac{1}{9} \times \frac{(1-3{,}44)^4 + (3-3{,}44)^4 + (2-3{,}44)^4 + (4-3{,}44)^4 + (6-3{,}44)^4 + (2-3{,}44)^4 + (2-3{,}44)^4 + (5-3{,}44)^4 + (6-3{,}44)^4}{9} = 1{,}551$$

$$\mu_2 = \sigma^2 = 3{,}14 \Rightarrow kurt = \frac{\mu_4}{\mu_2^2} = \frac{15{,}59}{3{,}14^2} = 1{,}58$$

Ví dụ tính giá trị thống kê mẫu với R

```
> x<-c(1,3,2,4,6,2,2,5,6)
> library(moments)
> mean(x)
[1] 3.444444
> var(x)
[1] 3.527778
> quantile(x,0.25)
25%
  2
> quantile(x)
  0%  25%  50%  75% 100%
   1    2    3    5    6
> kurtosis(x)
[1] 1.582584
> skewness(x)
[1] 0.2717328
```

Các kỹ thuật biểu diễn bằng đồ thị

- Dữ liệu một chiều (univariate data)
  - Histogram
  - Boxplot
  - Quantile-based plot
  - Stem and leaf
- Dữ liệu hai hoặc ba chiều
  - Scatter plot
  - Surface plot
  - Contour plot
  - Bivariate histogram
- Dữ liệu nhiều chiều: scatter plot matrix,…

Khái niệm histogram theo tần số

- Khái niệm Histogram theo tần số (frequency histogram):
  - Trục ngang: miền dữ liệu được chia thành các bin (khoảng giá trị). Các giá trị thuộc bin nào thì sẽ được đếm cho bin đó. Cách phân chia các bin: tùy ý theo người dùng hoặc theo một hệ thống luật [Scott 1992]
  - Trục dọc: tần số của từng bin (số lượng dữ liệu thuộc từng bin)
  - $y(x) = v_k$ với x thuộc B_k

  với y(x): giá trị trên trục dọc ứng với x ; v_k: số lượng dữ liệu thuộc bin thứ k; B_k: bin thứ k

Khái niệm các histogram biến thể

- Histogram tần số tương đối (Relative frequency histogram):
  - $y(x) = v_k / n$ với x thuộc B_k

  với n là tổng số dữ liệu

- Histogram theo mật độ (Density histogram):

  $y(x) = v_k / (nh)$ với x thuộc B_k

  với h là độ rộng của bin

  Đặc điểm: tổng diện tích các cột bằng 1.

  Ví dụ: Vẽ histogram bằng ngôn ngữ R:

Xây dựng histogram

- Cho tập dữ liệu X = {1,1,1,2,2,3,4,5,7}

Giả sử cần xây dựng histogram với 4 bin: {1,2} (bin 1), {3,4} (bin 2), {5,6} bin 3, {7,8} bin 4. Hãy xây dựng histogram tần số và histogram theo mật độ.

Sử dụng R để xây dựng histogram. So khớp kết quả.

Xây dựng histogram với R

- Cho tập dữ liệu: x={1,1,1,1,3,4,6,4,4,6,7}
- Vẽ histogram tần số

hist(x,so_bin=3,freq=TRUE)

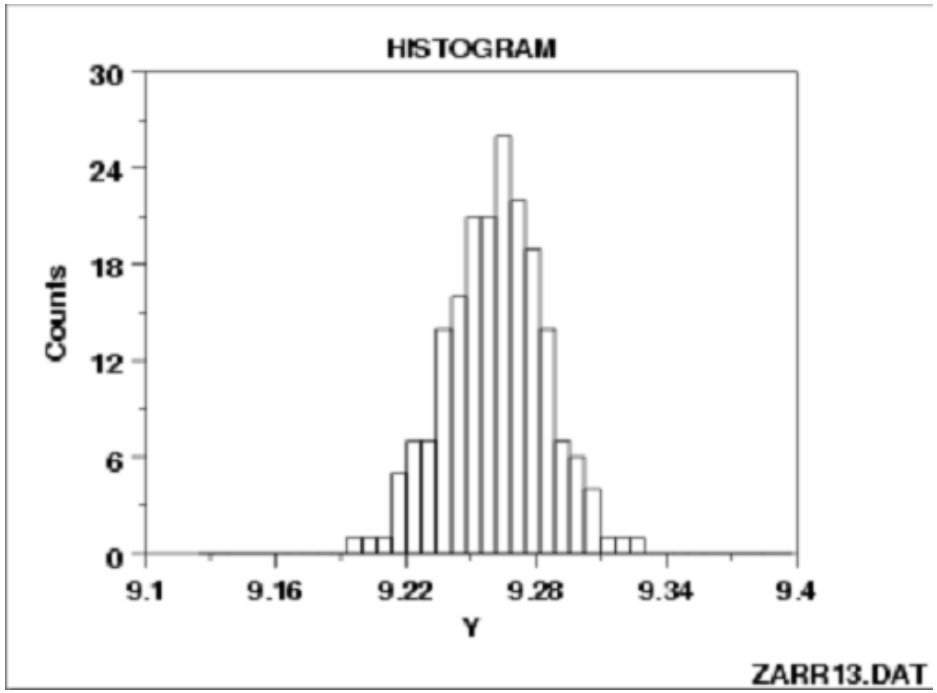
- Vẽ histogram mật độ

hist(x,so_bin=3,freq=FALSE)

Đặc trưng của histogram

- Suy ra các đặc điểm phân bố của dữ liệu:
  - Vị trí tâm của dữ liệu (center)
  - Độ phân tán (spread)
  - Độ lệch (skewness)
  - Giá trị ngoại lệ (outlier)
  - Yếu vị (mode)
- Đề xuất mô hình phân phối xác suất phù hợp
- Phát hiện các bất thường

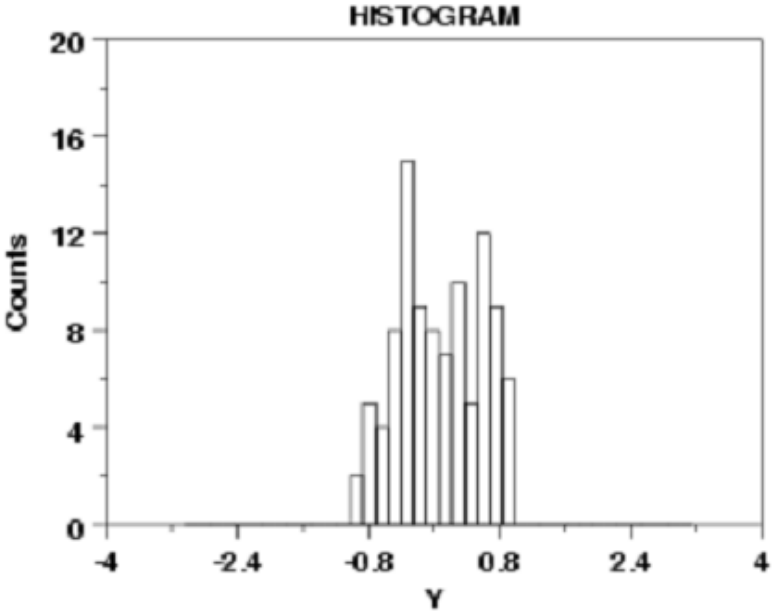
Ví dụ - Histogram đối xứng, phần đuôi vừa phải, dạng chuẩn



- Vị trí tâm của dữ liệu (center): ở giữa
- Độ phân tán (spread): tập trung ở giữa, giảm dần ở hai bên, phần đuôi vừa phải
- Độ lệch (skewness): đối xứng
- Giá trị ngoại lệ (outlier): không có
- Yếu vị (mode): 1 yếu vị

-> Kiểm tra phân phối chuẩn

Ví dụ - Histogram đối xứng, phần đuôi ngắn

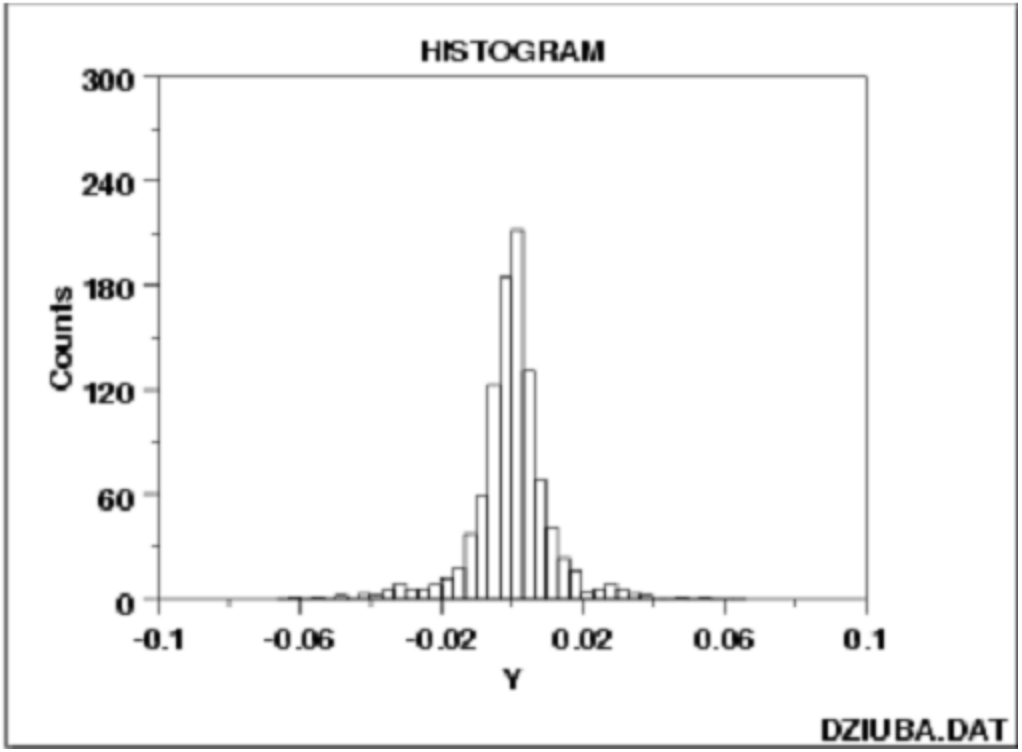


- Vị trí tâm của dữ liệu (center): chưa xác định
- Độ phân tán (spread): đều, giảm dần hai bên, phần đuôi ngắn
- Độ lệch (skewness): đối xứng
- Giá trị ngoại lệ (outlier): không có
- Yếu vị (mode): >1 yếu vị

-> Kiểm tra phân phối đều

- ✓ EDA
- ✓ Lấy mẫu
  - ✓ Khái niệm
  - ✓ Lấy mẫu
  - ✓ Lấy mẫu xác suất
  - ✓ Xử lý mẫu
- ✓ Thống kê mô tả
  - ✓ Khái niệm
  - ✓ Các giá trị thống kê mô tả
  - ✓ Các kỹ thuật biểu diễn đồ thị
    - ✓ Histogram
    - ✓ Boxplot
    - ✓ Quantile-based plot
    - ✓ Scatter plot

Ví dụ - Histogram đối xứng, phần đuôi dài, không có dạng chuẩn

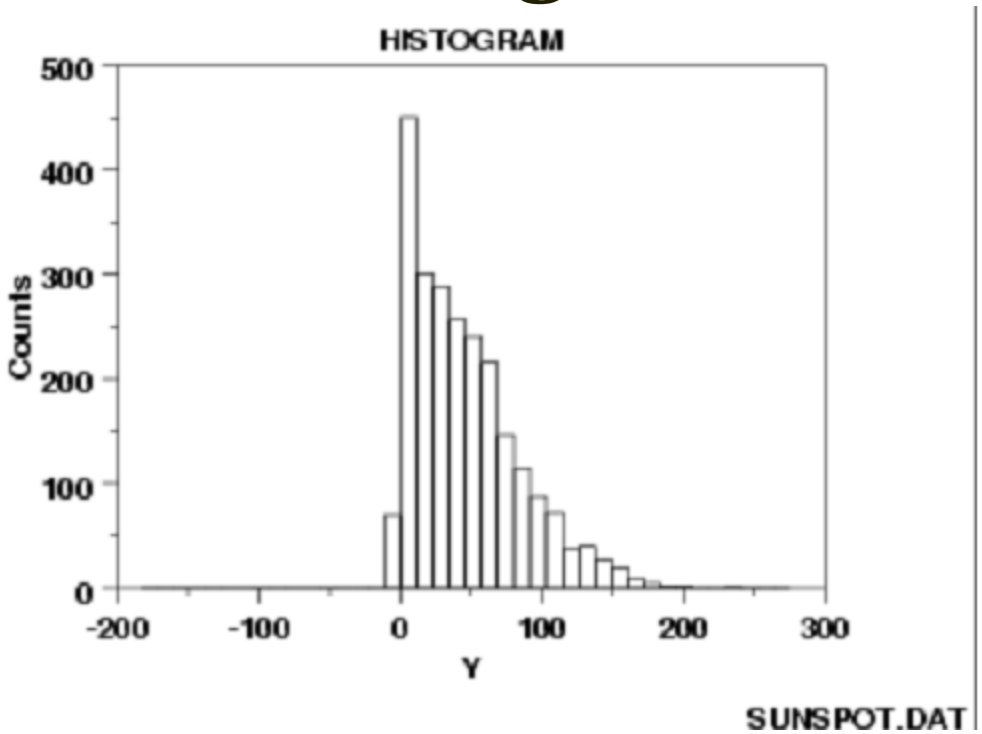


- Vị trí tâm của dữ liệu (center): ở giữa
- Độ phân tán (spread): tập trung ở giữa, giảm dần hai bên, phần đuôi dài
- Độ lệch (skewness): đối xứng
- Giá trị ngoại lệ (outlier): không có
- Yếu vị (mode): 1 yếu vị

-> Kiểm tra phân phối Cauchy (chưa đi chi tiết vào phân phối Cauchy)

- ✓ EDA
- ✓ Lấy mẫu
  - ✓ Khái niệm
  - ✓ Lấy mẫu
  - ✓ Lấy mẫu xác suất
  - ✓ Xử lý mẫu
- ✓ Thống kê mô tả
  - ✓ Khái niệm
  - ✓ Các giá trị thống kê mô tả
  - ✓ Các kỹ thuật biểu diễn đồ thị
    - ✓ Histogram
    - ✓ Boxplot
    - ✓ Quantile-based plot
    - ✓ Scatter plot

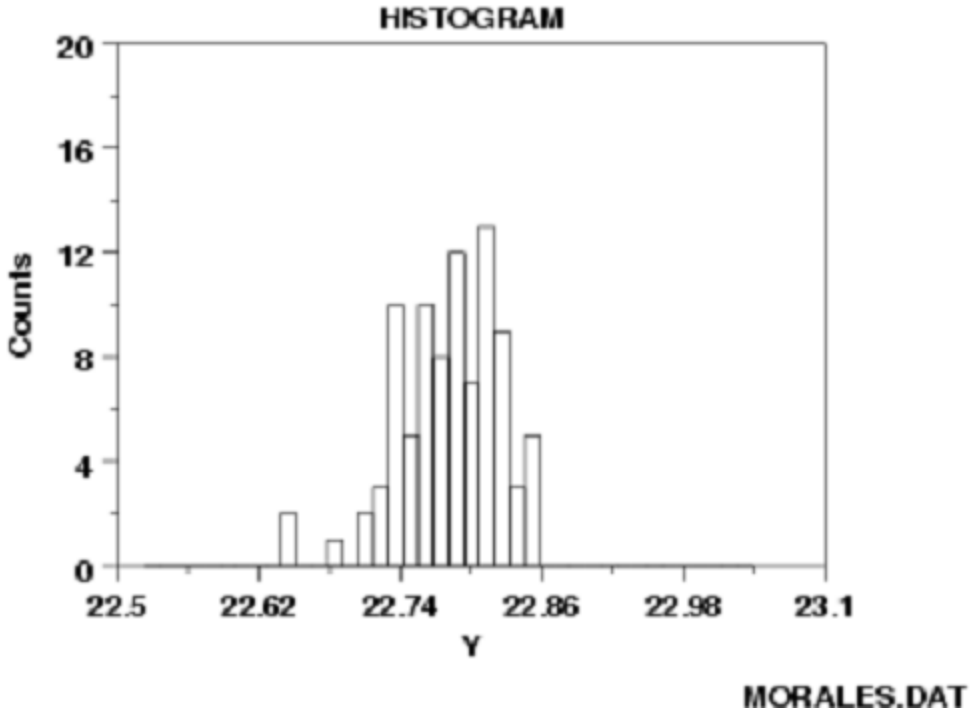
Ví dụ - Histogram lệch phải



- Vị trí tâm của dữ liệu (center): không có
- Độ phân tán (spread): tập trung bên trái, giảm dần sang phải, phần đuôi phải dài
- Độ lệch (skewness): lệch phải
- Giá trị ngoại lệ (outlier): không có
- Yếu vị (mode): 1 yếu vị
- -> Tính mean, median, mode để đặc trưng cho vị trí dữ liệu
- -> Kiểm tra các họ phân phối: Chi-square, lognormal, gamma…

- ✓ EDA
- ✓ Lấy mẫu
  - ✓ Khái niệm
  - ✓ Lấy mẫu
  - ✓ Lấy mẫu xác suất
  - ✓ Xử lý mẫu
- ✓ Thống kê mô tả
  - ✓ Khái niệm
  - ✓ Các giá trị thống kê mô tả
  - ✓ Các kỹ thuật biểu diễn đồ thị
    - ✓ Histogram
    - ✓ Boxplot
    - ✓ Quantile-based plot
    - ✓ Scatter plot

Ví dụ - Histogram lệch trái



- Tương tự lệch phải

Bài tập

Sử dụng R để:

- Phát sinh tập mẫu 1000 phần tử theo phân phối chuẩn chính tắc.
- Xây dựng histogram theo tần số, histogram theo mật độ.
- Đối với histogram theo mật độ, vẽ thêm đường cong hàm mật độ xác suất lý thuyết của phân phối chuẩn. Nhận xét đường cong có khớp với histogram không.

Boxplot (Box and whisker)

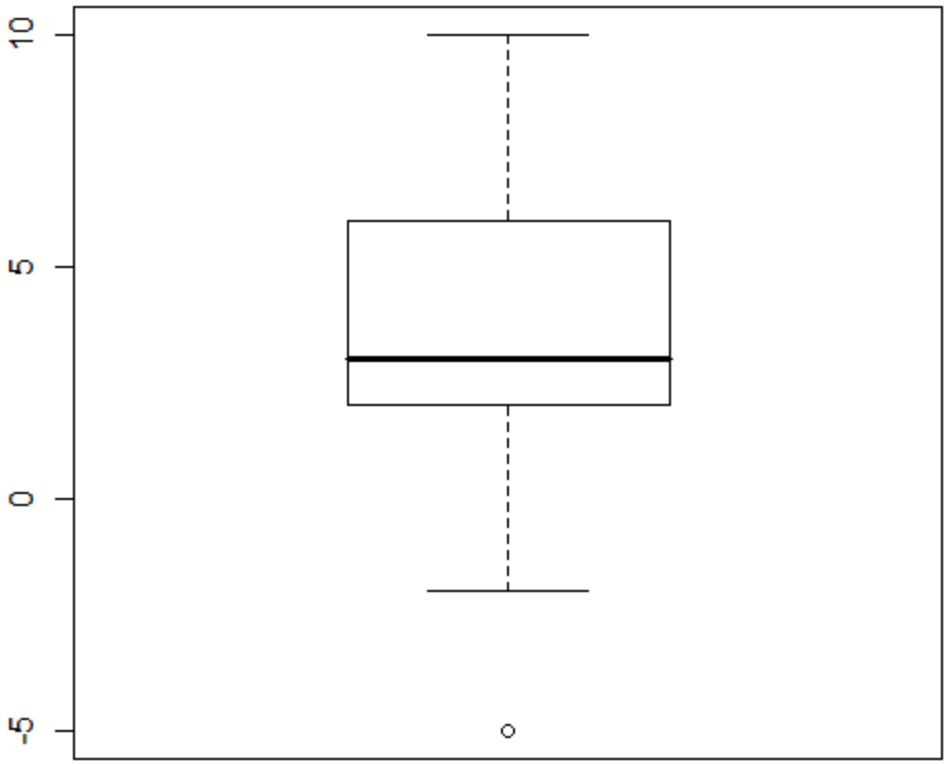
- Khoảng cách giữa hai phần tư vị (IQR, interquartile range): IRQ = $q_{0.75}$ - $q_{0.25}$
- Giới hạn dưới (lower limit): LL = $q_{0.25}$ - 1.5×IQR
- Giới hạn trên (upper limit): UL = $q_{0.25}$ + 1.5×IQR
- Ngoại lệ: các giá trị cách biệt phần còn lại của dữ liệu
  - Có thể do lấy mẫu sai sót
  - Có thể là các điểm cực trị của phân bố
  - Tóm lại: cần xem xét kỹ
- Các giá trị kề (adjacent values): các giá trị cực trị trong tập mẫu nằm trong giới hạn LL và UL. Nếu không có các giá trị ngoại lệ có khả năng, đây là min, max của tập dữ liệu.

Khái niệm Boxplot

- Ý tưởng chính: các giá trị nằm ngoài UL, LL có khả năng là ngoại lệ.
- Cấu tạo: Tạo thành bởi 5 giá trị: 3 giá trị quartile mẫu $q_{0.25}$, $q_{0.5}$, $q_{0.75}$, min, max
- Nhiều phiên bản (phụ thuộc phần mềm)
- Có thể vẽ các boxplot của các mẫu khác nhau để so sánh phân phối xác suất của các mẫu.

Ví dụ boxplot

```
> x<-c(1,3,2,4,6,2,2,5,6,6,7,8,2,5,2,-5,10,-2,-1)
> quantile(x)
  0%  25%  50%  75% 100%
  -5    2    3    6   10
> boxplot(x)
```



Khái niệm Boxplot

- > x<-c(1,3,2,4,6,2,2,5,6,6,7,8,2,5)
- > quantile(x)
- 0% 25% 50% 75% 100%
- 1.0 2.0 4.5 6.0 8.0
- > boxplot(x)
- > sort(x)

[1] 1 2 2 2 2 3 4 5 5 6 6 6 7 8

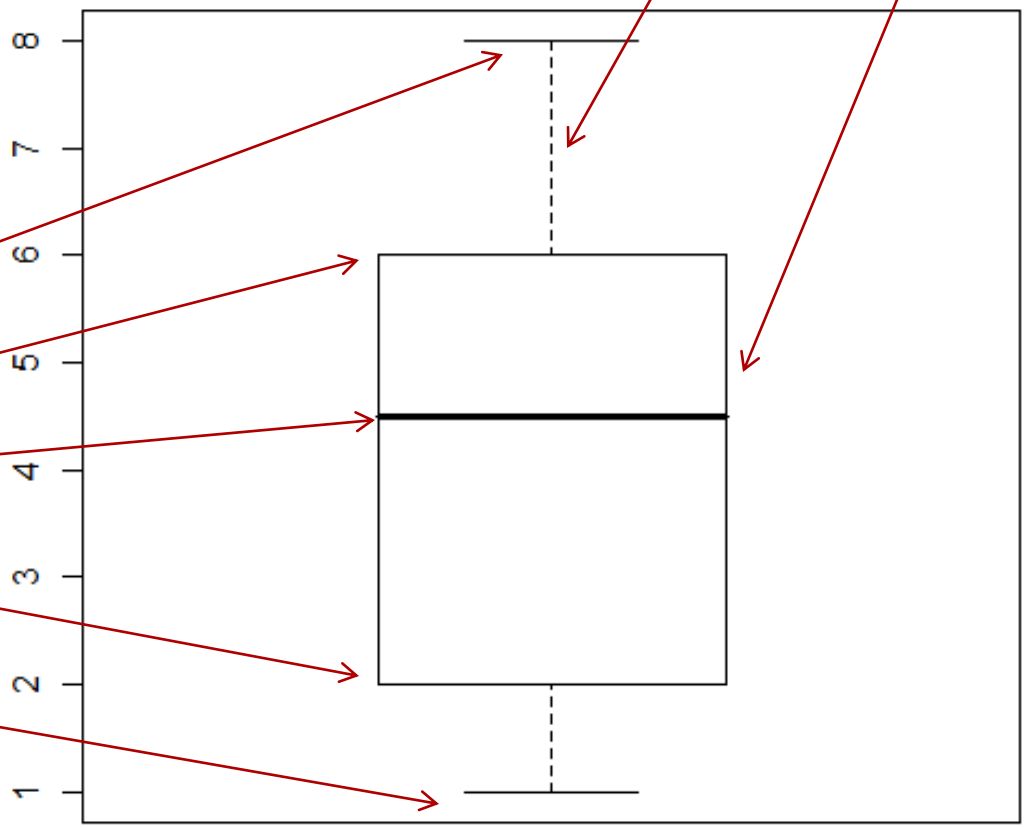
Giá trị kề

$q_{0.75}$

$q_{0.5}$

$q_{0.25}$

Giá trị kề



Đặc trưng của Boxplot

- Độ dài 2 whisker và hai phần của box
  - Nếu xấp xỉ bằng nhau: dữ liệu phân bố đối xứng => Phân phối chuẩn hoặc đều
  - Nếu lệch về một bên: dữ liệu lệch
- IQR
  - Nếu IQR nhỏ: Dữ liệu tập trung quanh trung vị
  - Nếu IQR lớn: dữ liệu phân tán rộng

Các đồ thị dựa trên quantile

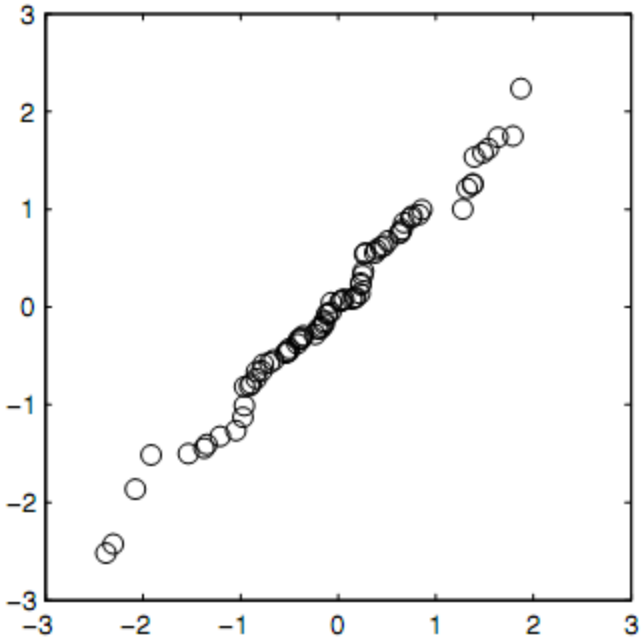
- Q-q plot
- Quantile plot (Probability plot)

Khái niệm q-q plot

- Xác định 2 tập dữ liệu có cùng phân phối xác suất không
- Ý chính: Vẽ đồ thị các phân vị ước lượng của tập dữ liệu 1 và các phân vị ước lượng của tập dữ liệu 2
- Phân vị ước lượng của tập dữ liệu: lấy tập giá trị sắp xếp rồi của tập mẫu
- Thuận lợi:
  - Kích thước 2 tập mẫu không cần bằng nhau
  - So sánh được nhiều khía cạnh của phân bố: vị trí, sự phân tán, tính đối xứng, ngoại lệ

Ví dụ q-q plot

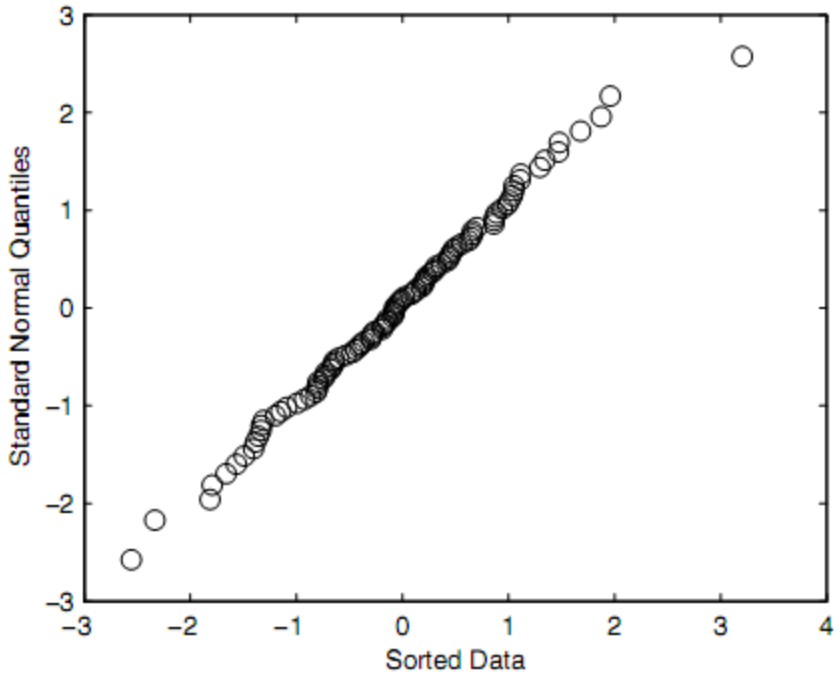
- Vẽ q-q plot:
  - Tập dữ liệu 1: sắp thứ tự (n phân vị): $x_1, x_2, \ldots, x_n$
  - Tập dữ liệu 2: sắp thứ tự (m phân vị): $y_1, y_2, \ldots, y_m$
  - Nếu m=n: Vẽ $\{x_i, y_i\}$
  - Nếu m<n: $\{(i-0.5)/m , y_i\}$ với i=1,…,m
  - Nếu 2 tập mẫu thuộc 2 quần thể có cùng phân phối, các điểm của đồ thị xấp xỉ đường thẳng.



Khái niệm quantile plot (Probability plot)

- Vẽ đồ thị biểu diễn các quantile lý thuyết với các quantile của tập mẫu.
- Vẽ đồ thị:

$\{x_i, F^{-1}((i-0.5)/n)\}$ với i=1,…,n; F là hàm cdf



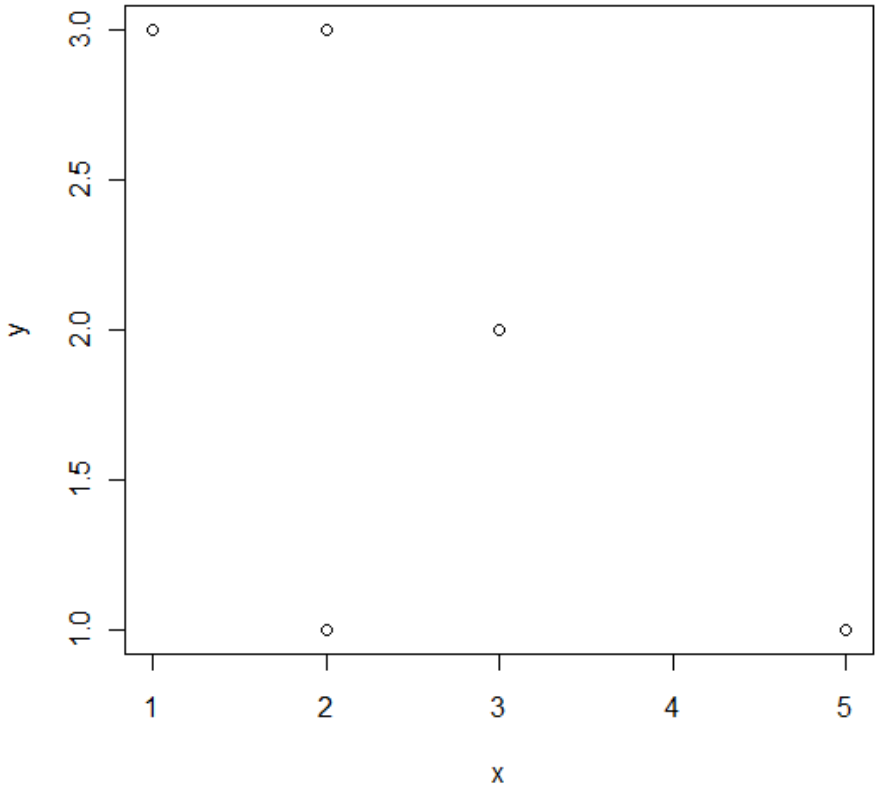
Scatter plot

- Thể hiện:
  - Dữ liệu phân bố theo 2 chiều như thế nào
  - Hai biến liên hệ như thế nào: tuyến tính, phi tuyến tính
  - Cách vẽ:
    - Giả sử có 2 tập mẫu $X=\{x_1,x_2,\ldots\}$, $Y=\{y_1,y_2,\ldots\}$
    - Vẽ các cặp điểm (x_1,y_1), (x_2,y_2), …

- ✓ EDA
- ✓ Lấy mẫu
  - ✓ Khái niệm
  - ✓ Lấy mẫu
  - ✓ Lấy mẫu xác suất
  - ✓ Xử lý mẫu
- ✓ Thống kê mô tả
  - ✓ Khái niệm
  - ✓ Các giá trị thống kê mô tả
  - ✓ Các kỹ thuật biểu diễn đồ thị
    - ✓ Histogram
    - ✓ Boxplot
    - ✓ Quantile-based plot
    - ✓ Scatter plot

Ví dụ scatter plot

```
> x<-c(2,1,2,3,5)
> y<-c(3,3,1,2,1)
> plot(x,y)
```



- ✓ Lấy mẫu
  - ✓ Khái niệm
  - ✓ Lấy mẫu
  - ✓ Lấy mẫu xác suất
  - ✓ Xử lý mẫu
- ✓ Thống kê mô tả
  - ✓ Khái niệm
  - ✓ Thống kê mô tả bằng số
  - ✓ Thống kê mô tả bằng hình
    - ✓ Histogram
    - ✓ Boxplot

Tài liệu trích dẫn

- Giáo trình "Thống kê máy tính và ứng dụng" (Khoa CNTT, NXB KHKT, 2010)
- [4] NIST/SEMATECH e-Handbook of Statistical Methods (http://www.itl.nist.gov/div898/handbook/
- Levy, Paul S. and Stanley Lemeshiow. 1999. *Sampling of Populations: Methods and Applications,* New Yorik: John Wiley & Sons.
- Martinez & Martinez . *Computational Statistics Handbook with MATLAB.*
- Montgomery && Runger. *Applied Statistics And Probability For Engineers.*