THỐNG KÊ MÁY TÍNH

(Computational Statistics)

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giảng viên: TS.Nguyễn Khắc Cường

CHƯƠNG 3

TÓM TẮT DỮ LIỆU

3.1.Đo lường mức độ tập trung của dữ liệu

- Trung bình cộng (Arithmetic mean)
  - Công thức:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

    - x_i: giá trị quan sát thứ i của mẫu; n: kích thước mẫu.
  - Hàm trong R: mean(dữ liệu)
  - VD:
    - Doanh số của các xí nghiệp

Xí nghiệp	A	B	C	D	E	F
Doanh số	25	17	34	26	43	35

    - Doanh số trung bình:

```
> ds<-c(25,17,34,26,43,35)
> mean(ds)
[1] 30
```

  - Nhận xét: Các phần tử có giá trị tương đồng và ít tương đồng → mean ?

3.1.Đo lường mức độ tập trung của dữ liệu

- Trung bình cộng có trọng số (Weighted mean)
  - Công thức:

$$\overline{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

    - x_i: giá trị quan sát thứ i của mẫu; n: kích thước mẫu
    - w_i: trọng số/tần số xuất hiện của quan sát thứ i
  - VD:

Tổ	A	B	C	D	E
Số lượng thành viên	10	15	14	10	16
Sản phẩm của mỗi thành viên trong tổ	30	20	25	20	25

    - Số sản phẩm trung bình mỗi thành viên làm được

$$\overline{x}_w = \frac{(10*30+15*20+14*25+10*20+16*25}{(10+15+14+10+16)} = 23.84615$$

3.1.Đo lường mức độ tập trung của dữ liệu

- Trung bình cộng có trọng số (Weighted mean)
  - VD:

Tổ	A	B	C	D	E
Số lượng thành viên	10	15	14	10	16
Sản phẩm của mỗi thành viên trong tổ	30	20	25	20	25

  - Số sản phẩm trung bình mỗi thành viên làm được

$$\overline{x}_w = \frac{(10*30+15*20+14*25+10*20+16*25}{(10+15+14+10+16)} = 23.84615$$

  - Tính bằng R

```
> sl<-c(10,15,14,10,16)
> sp<-c(30,20,25,20,25)
> ts<-sl*sp
> ts
[1] 300 300 350 200 400
> x<-sum(ts)/sum(sl)
> x
[1] 23.84615
```

3.1.Đo lường mức độ tập trung của dữ liệu

- Trung bình hình học (Geometric Mean)
  - Công thức: $\bar{x} = (x_1 \times x_2 \times .. \times x_n)^{1/n} = (\prod_{i=1}^{n} x_i)^{1/n}$ (trung bình nhân)
  - Công thức tương đương

$$(\prod_{i=1}^{n} x_i)^{1/n} = \exp(\frac{1}{n}\sum_{i=1}^{n} \ln(x_i))$$ (TB hình học = TB logarit)

  - VD:

```
> data <- c(1, 15, 12, 5, 18, 11, 12, 15, 18, 25)
> exp(mean(log(data)))
[1] 10.37383
```

  - VD: tính TB hình học cho 3 biến

```
> data<- data.frame(x=c(10, 13, 14, 26, 38, 28, 29),
+                   y=c(15, 8, 18, 17, 1, 1, 6),
+                   z=c(12, 10, 18, 28, 29, 29, 12))
> apply(data[ , c('x', 'y', 'z')], 2, function(x) exp(mean(log(x))))
        x         y         z
20.379699  5.798203 17.992195
```

3.1.Đo lường mức độ tập trung của dữ liệu

- Trung bình cộng điều hòa (Harmonic mean)
  - Công thức: $\bar{x} = H = n / \sum_{i=1}^{n} \frac{1}{x_i}$
  - Tác dụng:
    - Thường dùng để đo xu hướng tập trung về trung tâm đối với dữ liệu chứa các giá trị đại diện cho tốc độ thay đổi
  - VD:
    - Khoảng cách giữa A và B là 120 km
    - Vòng đi: đi trong 3 giờ
    - Vòng về: đi trong 2 giờ

    → tổng khoảng cách đi và về: 240 km, tổng thời gian: 5 giờ

    → Tốc độ trung bình = 240 km / 5 giờ = 48 km/h

3.1.Đo lường mức độ tập trung của dữ liệu

- Trung bình cộng điều hòa (Harmonic mean)
  - VD: tính bằng Harmonic mean R

```
> allSpeed= c(40,60)
> N=length(allSpeed)
> inverseOfAllSpeed=allSpeed^(-1)
> sumOfInverse=sum(inverseOfAllSpeed)
> HM=N/sumOfInverse
> print(HM)
[1] 48
```

→Trong khi TB cộng = (40 km/h + 60 km/h) / 2 = 50 km/h

→Nhận xét:

Harmonic mean chính xác hơn trong loại dữ liệu này

  - VD

```
> # create a data vector
> x <- c(30, 35, 45)
> # calculate harmonic mean
> HM <- 1/mean(1/x)
> HM
[1] 35.66038
```

3.1.Đo lường mức độ tập trung của dữ liệu

- Median (trung vị)
  - Median là giá trị đứng giữa của dãy đã sắp xếp tăng dần
  - Công thức:
    - Sắp xếp phần tử tăng dần
    - Tìm median M_e

$$M_e = \begin{cases} (x_{[n/2]} + x_{[(n+2)/2]})/2 & \text{nếu n chẵn} \\ x_{[n+1]/2} & \text{nếu n lẻ} \end{cases}$$

  - Tính median trong R

```
> x<-c(10,18,12,30,20)
> median(x)
[1] 18
```

  - Nhận xét:
    - Median không phụ thuộc vào giá trị biên

3.1.Đo lường mức độ tập trung của dữ liệu

- Mode
  - Giới thiệu
    - Mode là giá trị xuất hiện nhiều lần nhất
    - Có thể có nhiều mode trong một tập dữ liệu
  - VD:
    - Dãy các giá trị quan sát: 7 15 18 22 25 37
      → Không có mode
    - Dãy các giá trị quan sát: 7 15 18 25 25 37
      → mode là 25
    - Dãy các giá trị quan sát: 7 7 15 18 25 25
      → mode là 7 và 25

3.1.Đo lường mức độ tập trung của dữ liệu

- Mode
  - Tìm mode trong R

```
> # Create the function.
> getmode <- function(v) {
+    uniqv <- unique(v)
+    uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
>
> # Create the vector with numbers.
> v <- c(2,1,2,3,1,2,3,4,1,5,5,3,2,3)
>
> # Calculate the mode using the user function.
> result <- getmode(v)
> print(result)
[1] 2
```

3.1.Đo lường mức độ tập trung của dữ liệu

- Midrange
  - Giới thiệu:
    - Midrange là giá trị trung bình cộng của giá trị lớn nhất và giá trị nhỏ nhất của dữ liệu
  - VD

```
> x<-c(27,17,34,26,43,35)
> range(x)
[1] 17 43
> mean(range(x))
[1] 30
```

3.1.Đo lường mức độ tập trung của dữ liệu

- Quartiles (Tứ phân vị)
  - Giới thiệu:
    - Q1 (tứ phân vị thứ nhất): Là giá trị trị sao cho có
      - 25% số quan sát nhỏ hơn Q1
      - và 75% số quan sát lớn hơn Q1
    - Q2 (tứ phân vị thứ hai): Là số trung vị, có
      - 50% số quan sát nhỏ hơn Q2
      - và 50% số quan sát lớn hơn Q2
    - Q3 (tứ phân vị thứ ba): Là giá trị sao cho có
      - 75% số quan sát nhỏ hơn Q3
      - và 25% số quan sát lớn hơn Q3
  - Tìm Quartiles trong R

```
> x<-c(6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   25.50   40.00   33.18   42.50   49.00
```

3.2.Đo lường mức độ phân tán của dữ liệu

- Range (Khoảng biến thiên)
  - Giới thiệu:
    - Là hiệu số giữa giá trị lớn nhất và nhỏ nhất trong dữ liệu quan sát

$$R = a_{\max} - a_{\min}$$

  - Tác dụng:
    - Cho biết độ trải của dữ liệu
    - Không cho biết mức độ phân bố của dữ liệu
  - VD:

```
> x<-c(12,15,23,45,67,89)
> range(x)
[1] 12 89
> min(x)
[1] 12
> max(x)
[1] 89
```

3.2.Đo lường mức độ phân tán của dữ liệu

- Interquartile Range (Độ trải trong)
  - Ký hiệu: IQR
  - Cách tính: $IQR = Q_3 - Q_1$
    - Là hiệu số giữa tứ phân vị thứ 3 (Q3) và tứ phân vị thứ nhất (Q1)
- Variance (Phương sai)
  - Tác dụng :
    - Phương sai dùng để đánh giá mức độ biến thiên của các giá trị quan sát quanh giá trị trung bình
  - Công thức:

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

  - VD:

```
> x<-c(25, 17, 34, 26, 43, 35)
> var(x)
[1] 84
```

3.2.Đo lường mức độ phân tán của dữ liệu

- Standard Deviation (Độ lệch chuẩn)
  - Công thức: $sd = \sqrt{s^2}$
    - Độ lệch chuẩn là căn bận hai của phương sai
  - Ý nghĩa:
    - Phương sai : đánh giá ở miền bình phương của đơn vị đo dữ liệu
    - Độ lệch chuẩn : đánh giá ở miền đơn vị đo dữ liệu
    - Đa số các giá trị quan sát được nằm trong phạm vi $(\bar{x}-\sigma, \bar{x}+\sigma)$
      σ (=sd)
  - VD

```
> x<-c(25, 17, 34, 26, 43, 35)
> var(x)
[1] 84
> sd(x)
[1] 9.165151
> mean(x)
[1] 30
```

3.2.Đo lường mức độ phân tán của dữ liệu

- Sampling Distribution (Phân phối mẫu)
  - Giới thiệu:
    - Thực hiện lấy mẫu lặp lại N lần → thu được N bộ mẫu
    - Tính giá trị cần khảo sát (trung bình, median, …) đối với N bộ mẫu đó → thu được một phân phối mẫu đối với giá trị cần khảo sát
- Standard error (Sai số chuẩn)
  - Giới thiệu:
    - Thực hiện lấy mẫu (từ quần thể) N lần, mỗi lần lấy n phần tử
    - Sai số chuẩn chính là độ lệch chuẩn của tập hợp mẫu sau khi chọn mẫu N lần
  - Công thức: $SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
    - σ : là độ lệch chuẩn của quần thể; N: số lần lấy mẫu;
    - n: số lượng phần tử trong mẫu

3.2.Đo lường mức độ phân tán của dữ liệu

- Standard error (Sai số chuẩn)
  - Sử dụng sai số chuẩn:
    - Trong thực tế: tính được giá trị trung bình của quần thể là rất khó ?
    - Do đó, thường giá trị trung bình được tính từ mẫu
    - Mà sai số chuẩn thể hiện sự sai khác giữa giá trị trung bình của mẫu và giá trị trung bình của quần thể

    → Có thể ước lượng được giá trị trung bình của quần thể nhờ
      - Giá trị trung bình của mẫu
      - Sai số chuẩn

3.2.Đo lường mức độ phân tán của dữ liệu

- Coefficient of Variance (Hệ số biến thiên CV)
  - Công thức: $CV = \frac{s}{\bar{x}} * 100\%$
    - s : độ lệch chuẩn; $\bar{x}$: giá trị trung bình;
  - Ý nghĩa:
    - CV thể hiện độ phân tán trên một đơn vị trung bình
  - Sử dụng:
    - CV được sử dụng để so sánh độ phân tán của 2 tập dữ liệu (kể cả các tập dữ liệu đo lượng ở 2 đơn vị khác nhau)
  - VD:

```
> CV <- function(x)
+ {   sd<-sd(x)
+     xn<-mean(x)
+     CoV<-sd/xn*100
+     c(CV=CoV)
+ }
> x<-c(12,34,45,56,67,89)
> CV(x)
       CV
52.90873
```

3.3.Ứng dụng của thống kê mô tả

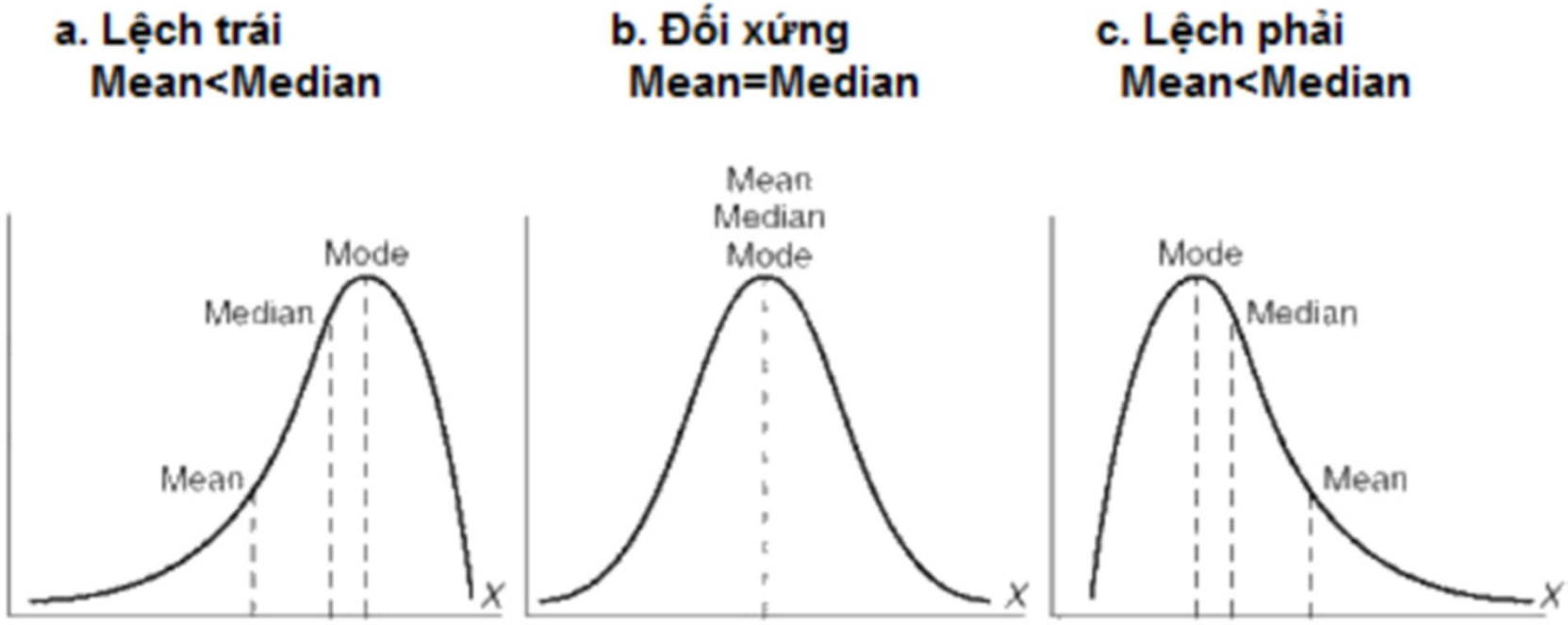
- Mối quan hệ thực nghiệm giữa average, median, mode
  - 3 giá trị này là các độ đo mức độ hướng tâm của dữ liệu
  - Không có độ đo duy nhất nào thể hiện chính xác mức độ hướng tâm của dữ liệu
  - Việc chọn độ đo nào phụ thuộc vào phân bố của dữ liệu
- Chú ý khi chọn độ đo:
  - Average:
    - Ưu:
      - Được tính từ dữ liệu số cụ thể của dữ liệu
      - Sử dụng tất cả các giá trị của các quạn sát
      - Có giá trị đơn nhất
    - Nhược: bị ảnh hưởng bởi giá trị biên (cực đoan)

3.3.Ứng dụng của thống kê mô tả

- Chú ý khi chọn độ đo:
  - Median:
    - Ưu:
      - Khắc phục được nhược điểm của average → không bị ảnh hưởng bởi giá trị biên (cực đoan)
    - Nhược:
      - Không tốt khi số lượng quan sát nhỏ, vì
        - chỉ là trung bình vị trí
        - Không phải là trung bình giá trị của dữ liệu
  - Mode:
    - Chỉ dùng tốt khi:
      - Dữ liệu có phân bố rõ ràng, cân đối về tính hướng tâm
      - Số lượng quan sát đủ lớn

3.3.Ứng dụng của thống kê mô tả

- Chú ý khi chọn độ đo:
  - Nắm rõ vị trí của các độ đo khi hình dạng phân phối của dữ liệu thay đổi



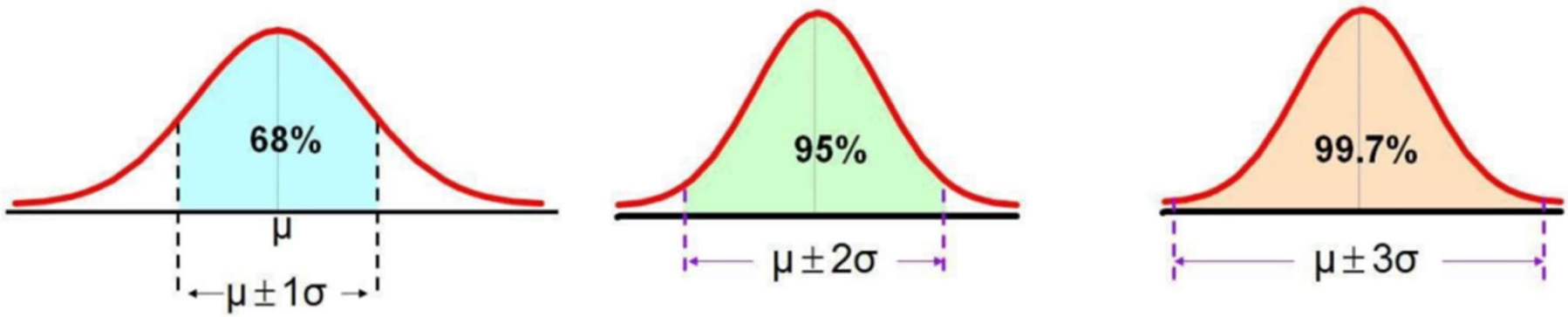
3.3.Ứng dụng của thống kê mô tả

- Định lý Chebychev
  - Với 1 quần thể bất kỳ có trung bình μ, độ lệch chuẩn σ, k là giá trị bất kỳ lớn hơn 1.
  - Tối thiểu $(1-1/k^2)x100\%$ các giá trị quan sát nằm trong khoảng (μ -k σ, μ +k σ)

k	Số phần trăm giá trị các quan sát	Thuộc khoảng
1	68%	$(\mu-\sigma,\mu+\sigma)$
2	95%	$(\mu-2\sigma,\mu+2\sigma)$
3	99%	$(\mu-3\sigma,\mu+3\sigma)$

3.3.Ứng dụng của thống kê mô tả

- Định lý Chebychev
  - Cụ thể:
    - μ + 1σ : chứa khoảng 68% giá trị dữ liệu của mẫu hoặc quần thể
    - μ + 2σ : chứa khoảng 95% giá trị dữ liệu của mẫu hoặc quần thể
    - μ + 3σ : chứa khoảng 99.7% giá trị dữ liệu của mẫu hoặc quần thể



Q / A