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 2

THỐNG KÊ HỌC

2.4.Trình bày dữ liệu

- Dùng bảng tần số
  - Dạng cơ bản: có 3 cột
    - Cột 1: đặc điểm / khoảng giá trị quan sát
    - Cột 2: tần số (số lần xuất hiện)
    - Cột 3: tần suất (tỉ lệ %)
    - VD:

Độ tuổi	Tần số (SV)	Tần suất
19-24	9	30.00
24-29	10	33.33
29-34	8	26.67
34 trở lên	3	10.00
Tổng	30	100.00

  - Dạng mở rộng: có thể thêm cột
    - Mô tả tính chất của dữ liệu
    - Tần suất tích lũy
    - . . .

2.4.Trình bày dữ liệu

- Dùng bảng phân tổ thống kê
  - Giới thiệu:
    - Tìm trong tập dữ liệu → các phần tử có tính chất giống hoặc tương tự nhau (xét theo tiêu chí nào đó ???) → gom lại thành một tổ
  - Yêu cầu:
    - Các phần tử khác tổ phải có tính chất khác nhau rõ rệt
    - Một phần tử chỉ thuộc một tổ → để các tổ không trùng nhau
  - Cách chia tổ:
    - Không có cách chia cụ thể, chủ yếu chia theo kinh nghiệm

2.4.Trình bày dữ liệu

- Dùng bảng phân tổ thống kê
  - Một số cách chia tham khảo:

Số tổ

$k = [(2*n)^{1/3}]$

$k = [1+3.3*\log_{10}(n)]$

Khoảng cách giữa các dữ liệu trong tổ (khoảng cách tổ)

$$h = \frac{X_{max} - X_{min}}{k}$$

Trong đó

k : số tổ; n : số lượng quan sát (phần tử) của dữ liệu;

h : khoảng cách tổ; X_{max}, X_{min} : giá trị lớn nhất, nhỏ nhất của dữ liệu

  - Ví dụ: cần chia tổ cho bộ dữ liệu sau

28	23	30	24	19	21	39	22	22	31	37
33	20	30	35	21	26	27	25	29	27	21
25	28	26	29	29	22	32	27			

2.4.Trình bày dữ liệu

- Dùng bảng phân tổ thống kê
  - Ví dụ: cần chia tổ cho bộ dữ liệu sau

28	23	30	24	19	21	39	22	22	31	37
33	20	30	35	21	26	27	25	29	27	21
25	28	26	29	29	22	32	27			

  - Tính số tổ cần chia: $k = [(2*n)^{1/3}] = [(2*30)^{1/3}] = [3.9] = 4$
  - Tính khoảng cách dữ liệu trong tổ: $h = \frac{X_{max} - X_{min}}{k} = \frac{39-19}{4} = 5$
  - Vậy số tổ được chia là:

Tổ 1: (19,24) tuổi

Tổ 2: (24,29) tuổi

Tổ 3: (29,34) tuổi

Tổ 4: (34,39) tuổi

Dữ liệu được chia thành 4 tổ

Mỗi tổ chứa các khoảng 5 tuổi

2.4.Trình bày dữ liệu

- Dùng biểu đồ phân phối tần số (Histogram)
  - Tác dụng:
    - Biểu diễn tần số xuất hiện các giá trị của một biến
  - Cú pháp: hist(v,main,xlab,ylab,xlim,ylim,col,border)
    - v : vector dữ liệu
    - main : tên biểu đồ
    - xlab, ylab: tên các trục
    - xlim, ylim : thiết lập tỉ lệ xích và khoảng giá trị trên các trục
    - col : màu của thanh biểu đồ
    - border : màu của được viền

2.4.Trình bày dữ liệu

- Dùng biểu đồ phân phối tần số (Histogram)
  - VD:
    - hist(distance)
    - hist(distance, main = "Frequency histogram")
    - hist(distance, main = "Frequency histogram", xlab="x axis", ylab="y axis")
    - hist(distance, main = "Frequency histogram", xlab="x axis", ylab="y axis", xlim=c(0,500), ylim=c(0,10))
    - hist(distance, main = "Frequency histogram", xlab="x axis", ylab="y axis", xlim=c(0,500), ylim=c(0,10), col="red")
    - hist(distance, main = "Frequency histogram", xlab="x axis", ylab="y axis", xlim=c(0,500), ylim=c(0,10), col="red", border="lightblue")
    -

2.4.Trình bày dữ liệu

- Dùng biểu đồ thanh (Bar chart)
  - Tác dụng:
    - Dùng biểu diễn các loại dữ liệu
      - Phân loại (categorical data)
      - Rời rạc (discrete data)
    - Mỗi thanh ứng với mỗi mỗi yếu tố (factor)
    - Chiều cao mỗi thanh là số lượng các quan sát ứng với mỗi yếu tố
  - Cú pháp:
    - barplot(h,xlab,ylab,main,names.arg,col)
  - VD 1:
    - data<-c(20,15, 17, 30, 35)
    - month<-c("Mar", "Apr", "May", "Jun", "Jul")
    - color<-c("yellow", "red", "blue", "pink", "green")
    - barplot(data, xlab="Month", ylab="Revenues", col=color, names.arg=month, main="Revenues Chart")

2.4.Trình bày dữ liệu

- Dùng biểu đồ thanh (Bar chart)
  - VD 2:
    - data<-c(20,15, 17, 30, 35)
    - month<-c("Mar", "Apr", "May", "Jun", "Jul")
    - color<-c("yellow", "red", "blue", "pink", "green")
    - barplot(data, xlab="Month", ylab="Revenues", col=color, names.arg=month, main="Revenues Chart")
  - VD 3:
    - data <- as.matrix(data.frame(A = c(0.2, 0.4),
      B = c(0.3, 0.1),
      C = c(0.7, 0.1),
      D = c(0.1, 0.2),
      E = c(0.3, 0.3)))
    - rownames(data) <- c("Group 1", "Group 2")
    - data
    - barplot(data, col = c("red", "blue"))
    - legend("topright", legend = c("Group 1", "Group 2"), fill = c("red", "blue"))

2.4.Trình bày dữ liệu

- Dùng biểu đồ thanh (Bar chart)
  - VD 4:
    - colors = c("green", "orange", "brown")
    - months <- c("Mar", "Apr", "May", "Jun", "Jul")
    - regions <- c("East", "West", "North")
    - Values <- matrix(c(2, 9, 3, 11, 9, 4, 8, 7, 3, 12, 5, 2, 8, 10, 11),
      nrow = 3, ncol = 5, byrow = TRUE)
    - barplot(Values, main = "Total Revenue", names.arg = months,
      xlab = "Month", ylab = "Revenue",
      col = colors, beside = TRUE)
    - legend("topleft", regions, cex = 0.7, fill = colors)

2.4.Trình bày dữ liệu

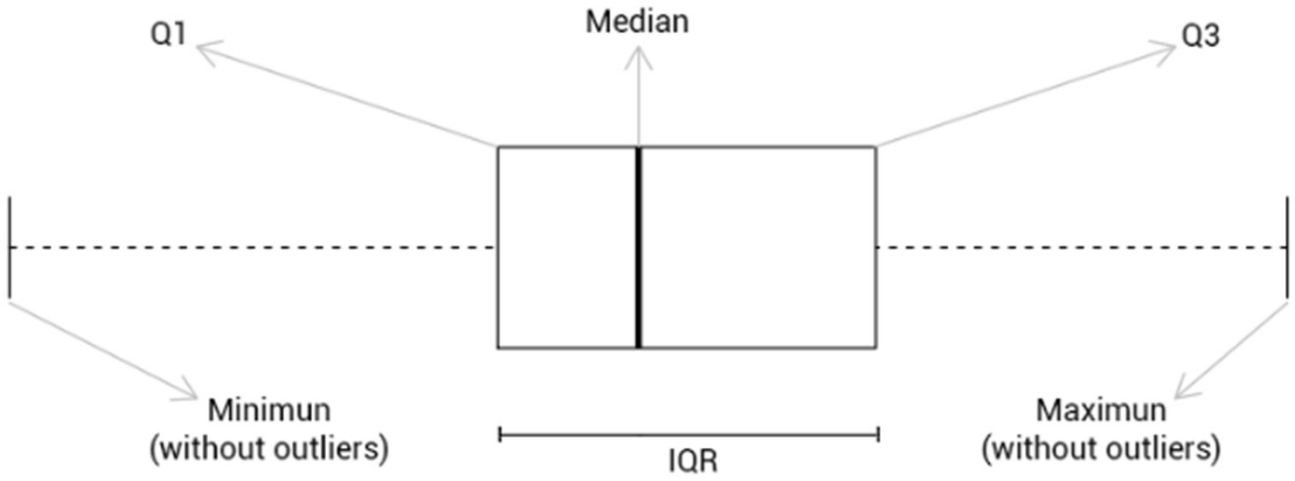
- Dùng biểu đồ tròn (Piecharts)
  - Tác dụng:
    - Biểu diễn dữ liệu rời rạc (Discrete data)
    - Thể hiện kết cấu của một tổng thể
  - Cú pháp:
    - pie(x,lables, radius, main, col, clockwise)
      - x : dữ liệu
      - labels : tên các mảnh
      - radius : bán kính hình tròn (-1,1)
      - clockwise :
        - TRUE : chia theo chiều kim đồng hồ
        - FALSE : chia theo ngược chiều kim đồng hồ

2.4.Trình bày dữ liệu

- Dùng biểu đồ tròn (Piecharts)
  - VD 1:
    - slices <- c(10, 12,4, 16, 8)
    - lbls <- c("US", "UK", "Australia", "Germany", "France")
    - pie(slices, labels = lbls, main="Pie Chart of Countries")
  - VD 2:
    - # Create data for the graph.
    - geeks <- c(23, 56, 20, 63)
    - labels <- c("Mumbai", "Pune", "Chennai", "Bangalore")
    - piepercent<- round(100 * geeks / sum(geeks), 1)
    - pie(geeks, labels = piepercent,
      main = "City pie chart", col = rainbow(length(geeks)))
    - legend("topright", c("Mumbai", "Pune", "Chennai", "Bangalore"),
      cex = 0.5, fill = rainbow(length(geeks)))

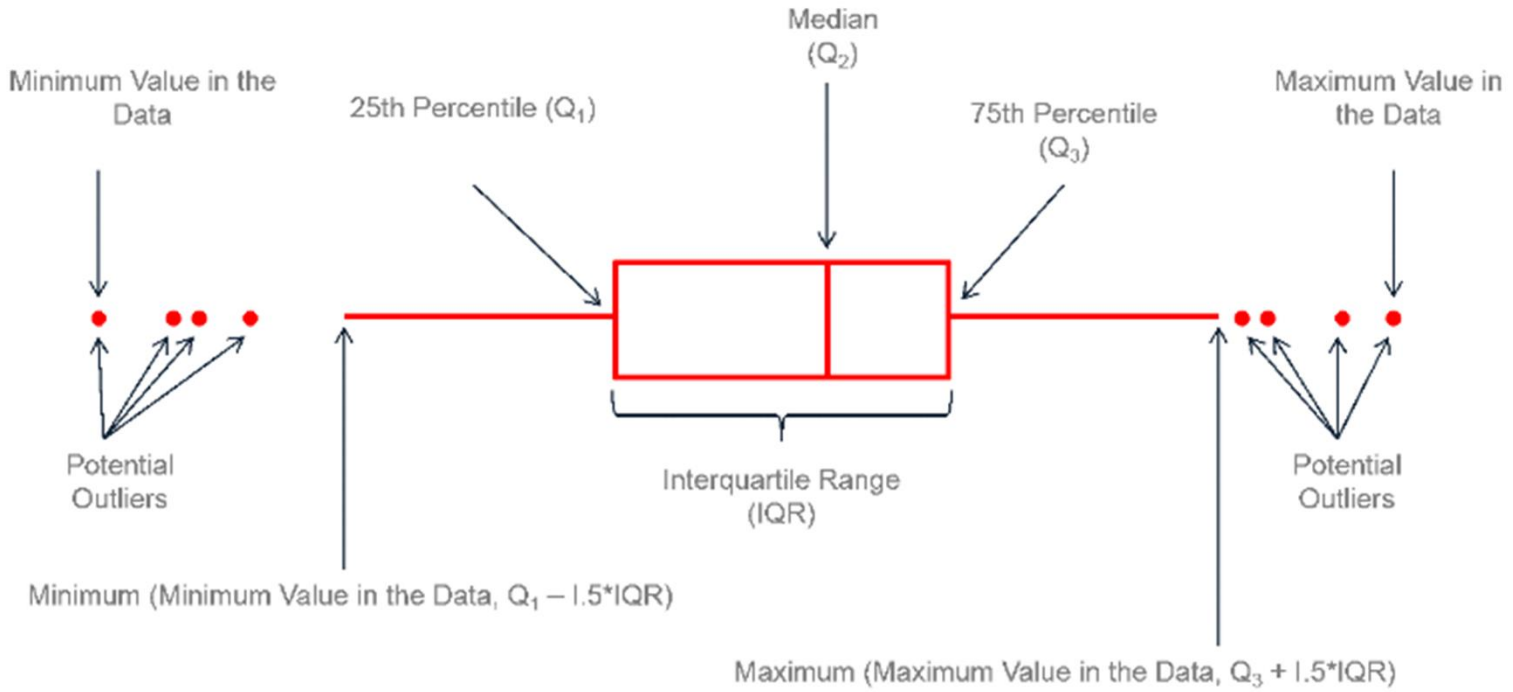
2.4.Trình bày dữ liệu

- Dùng biểu đồ hộp (Boxplots)
  - Tác dụng:
    - Thể hiện đồng thời các giá trị max, min, percentile
    - Thể hiện mối liên hệ giữa các đại lượng thống kê trên



2.4.Trình bày dữ liệu

- Dùng biểu đồ hộp (Boxplots)
  - Tác dụng:
    - Thể hiện đồng thời các giá trị max, min, percentile
    - Thể hiện mối liên hệ giữa các đại lượng thống kê trên



2.4.Trình bày dữ liệu

- Dùng biểu đồ hộp (Boxplots)
  - Cú pháp:
    - boxplot(x, data, notch, varwidth, names, main)
      - x : vector hay công thức dữ liệu
      - data : dữ liệu
      - notch : TRUE / FALSE
      - varwidth: TRUE / FALSE (chỉnh biểu đồ cân xứng với dữ liệu)
      - names : nhãn các nhóm
  - VD 1:
    - x <- c(8, 5, 14, -9, 19, 12, 3, 9, 7, 4,
      4, 6, 8, 12, -8, 2, 0, -1, 5, 3)
    - boxplot(x)
    - boxplot(x, horizontal = TRUE)

2.4.Trình bày dữ liệu

- Dùng biểu đồ đường (Line graph)
  - Tác dụng:
    - Nối các điểm dữ liệu lại trên một đường
    - Thể hiện sự thay đổi của dữ liệu
  - Cú pháp:
    - plot(v, type, col, xlab, ylab, lty)
      - v : vector dữ liệu
      - type:
        - type=”h”: vẽ các đường thẳng đứng
        - type=”o”: vẽ đường ngang qua các điểm
        - type=”b”: vẽ đường không ngang qua các điểm
        - type=”s”: vẽ đường theo kiểu bậc cấp
      - col : màu

2.4.Trình bày dữ liệu

- Dùng biểu đồ đường (Line graph)
  - Cú pháp:
    - plot(v, type, col, xlab, ylab, lty)
      - lty: kiểu đường

2.4.Trình bày dữ liệu

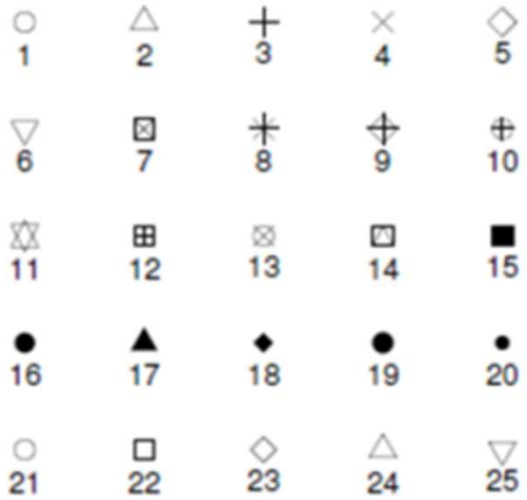
- Dùng biểu đồ đường (Line graph)
  - VD 1:
    - x <- 1:10
    - y1 <- x*x
    - y2 <- 2*y1
    - plot(x, y1, type = "S")
    - plot(x, y1, type = "b", pch = 19,
      col = "red", xlab = "x", ylab = "y")
    - plot(x, y1, type = "b", frame = FALSE, pch = 19,
      col = "red", xlab = "x", ylab = "y")
    - lines(x, y2, pch = 18, col = "blue", type = "b", lty = 2)
    - legend("topleft", legend=c("Line 1", "Line 2"),
      col=c("red", "blue"), lty = 1:2, cex=0.8)

2.4.Trình bày dữ liệu

- Dùng biểu đồ đường (Line graph)
  - VD 2:
    - t=0:10
    - z= exp(-t/2)
    - plot(t,z)
    - plot(t,z, type="l", col="green", lwd=5, xlab="time", ylab="concentration", main="Exponential decay")

2.4.Trình bày dữ liệu

- Dùng biểu đồ điểm (Dot chart)
  - Tác dụng:
    - Tương tự biểu đồ thanh
    - Thông tin được thu gọn lại thành điểm
  - Cú pháp:
    - dotchart(v, labels, col, xlab, ylab,pch)
      - pch : kiểu điểm cần vẽ



2.4.Trình bày dữ liệu

- Dùng biểu đồ điểm (Dot chart)
  - VD:
    - set.seed(1)
    - month <- month.name
    - expected <- c(15, 16, 20, 31, 11, 6,
      17, 22, 32, 12, 19, 20)
    - sold <- c(8, 18, 12, 10, 41, 2,
      19, 26, 14, 16, 9, 13)
    - quarter <- c(rep(1, 3), rep(2, 3), rep(3, 3), rep(4, 3))
    - data <- data.frame(month, expected, sold, quarter)
    - data
    - dotchart(data$sold, labels = data$month, pch = 21, bg = "green", pt.cex = 1.5)

2.4.Trình bày dữ liệu

- Dùng biểu đồ tán xạ (Scatter plot)
  - Tác dụng:
    - Dùng các điểm để thể hiện mối liên hệ giữa 2 biến
  - Cú pháp:
    - plot(x, y, main, xlab, ylab, xlim, ylim, pch)
  - VD:
    - x <- mtcars$wt
    - x
    - y <- mtcars$mpg
    - y
    - plot(x, y, main = "Main title",
      xlab = "X axis title", ylab = "Y axis title",
      pch = 19, frame = FALSE)
    - plot(x, y, main = "Main title",
      xlab = "X axis title", ylab = "Y axis title",
      pch = 19, frame = FALSE)
      abline(lm(y ~ x, data = mtcars), col = "blue")

Q / A