

# BÀI TẬP THỐNG KÊ ỨNG DỤNG

## Buổi 2

### YÊU CẦU BUỔI HỌC

#### 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

*Ví dụ: Giả sử số tiền (đơn vị nghìn VND) dùng cho chi tiêu thực phẩm trong một tuần là: 120, 150, 125, 100, 180, 140, 200. Khi đó số tiền trung bình một ngày trong tuần dành cho chi tiêu thực phẩm là:*

?

(Cho số liệu mẫu: 120, 150, 125, 100, 180, 140, 200. Hoặc 120 150 125 100 180 140 200  
Copy số liệu mẫu vào R để thực hành).

```
> x=c(120, 150, 125, 100, 180, 140, 200)
HỌC
> x=scan()
1: 120 150 125 100 180 140 200
8:
Read 7 items

> mean(x)
[1] 145
```

## Ví dụ

Điểm thi kết thúc học kì của một sinh viên được cho trong bảng dưới đây:

Môn học	Điểm	Số đ.v tín chỉ
TK Ứng dụng	7.0	3
Học máy	5.0	2
Trí tuệ nhân tạo	8.0	2
Hệ điều hành	4.0	2
Mạng máy tính	6.0	2

Khi đó điểm trung bình các môn học của sinh viên trên trong kì này sẽ là:

?

Số liệu mẫu:

7.0 3

5.0 2

8.0 2

4.0 2

6.0 2

dùng lệnh rep( ) ghép Diem và Tanso

```
> Diem=c(7,5,8,4,6)
```

```
> TanSo=c(3,2,2,2,2)
```

```
> DuLieu=rep(Diem,TanSo)
```

```
> DuLieu
```

```
[1] 7 7 7 5 5 8 8 4 4 6 6
```

```
> mean(DuLieu)
```

```
[1] 6.090909
```

Hoặc gộp lệnh

```
> mean(rep(Diem,TanSo))  
[1] 6.090909
```

## Ví dụ 2

Giả sử điểm thi cuối kỳ môn TKUD được cho bởi bảng sau đây, tính điểm trung bình môn TKUD của K58.

Khoảng điểm	Điểm đại diện ( $x_i^*$ )	Tần số ( $f_i$ )
[0.0, 1.5]	0.75	13
(1.5, 3.0]	2.25	27
(3.0, 4.5]	3.75	41
(4.5, 6.0]	5.25	31
(6.0, 7.5]	6.75	18
(7.5, 9.0]	8.25	13
(9.5, 10]	9.75	3

(Làm như VD trên)

## Ví dụ 1

Tính trung vị của tập dữ liệu về tuổi của 9 người như sau:

5, 11, 9, 12, 10, 20, 15, 30, 25.

Số liệu mẫu: 5, 11, 9, 12, 10, 20, 15, 30, 25

```
> x=c(5, 11, 9, 12, 10, 20, 15, 30, 25)  
> median(x)  
[1] 12
```

## Ví dụ 2

Tính trung vị của tập dữ liệu về tuổi của 8 người như sau:

5, 11, 9, 12, 10, 20, 15, 30.

Số liệu mẫu: 5, 11, 9, 12, 10, 20, 15, 30

```
> y=c(5, 11, 9, 12, 10, 20, 15, 30)
> median(y)
[1] 11.5
```

## Ví dụ

*Mode của tập dữ liệu:*

- 0, 1, 3, 1, 5, 2, 6, 2, 9, 2 là 2.
- 2, 3, 2, 5, 7, 8, 7, 15 là 2 và 7.
- 0, 1, 2, 3, 4, 5, 6 là tất cả các phần tử của tập.

Số liệu mẫu: 0, 1, 3, 1, 5, 2, 6, 2, 9, 2

```
> x=c(0, 1, 3, 1, 5, 2, 6, 2, 9, 2)
> table(x)
x
0 1 2 3 5 6 9
1 2 3 1 1 1 1
> which(table(x)==max(table(x)))
2
3
```

Mẫu thứ nhất cho mode=2, xuất hiện ở vị trí thứ 3.

Tương tự 2 mẫu còn lại.

## Ví dụ

Tìm phân vị thứ 25, thứ 50, thứ 60 và thứ 75 của tập dữ liệu sau:

9 9 10 11 13 13 13 15 16 20 20 24

Số liệu mẫu: 9 9 10 11 13 13 13 15 16 20 20 24

```
> x=scan()
1: 9 9 10 11 13 13 13 15 16 20 20 24
```

```
13:
Read 12 items
> quantile(x,probs = c(0.25,0.50,0.6,0.75))
 25%   50%   60%   75%
10.75 13.00 14.20 17.00
```

### Example

Tứ phân vị của tập dữ liệu: 10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70 tính trên phần mềm R lần lượt là 12.25, 22.50, 36.25.

Số liệu mẫu: 10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70

```
> x=c(10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70)
> quantile(x,probs = c(0.25,0.50,0.75))
 25%   50%   75%
12.25 22.50 36.25
```

### Định nghĩa

*Độ trải giữa còn được gọi là khoảng tứ phân vị được tính bằng hiệu giữa phân vị thứ ba và phân vị thứ nhất.*

*Công thức*

$$R_Q = Q_3 - Q_1$$

- **Ví dụ** Xét tập dữ liệu:

10 15 20 25 30 40 80 90

Khi đó  $R_Q = Q_3 - Q_1 = 70 - 16.26 = 53.75$ .

(làm tương tự VD trên)

### Example

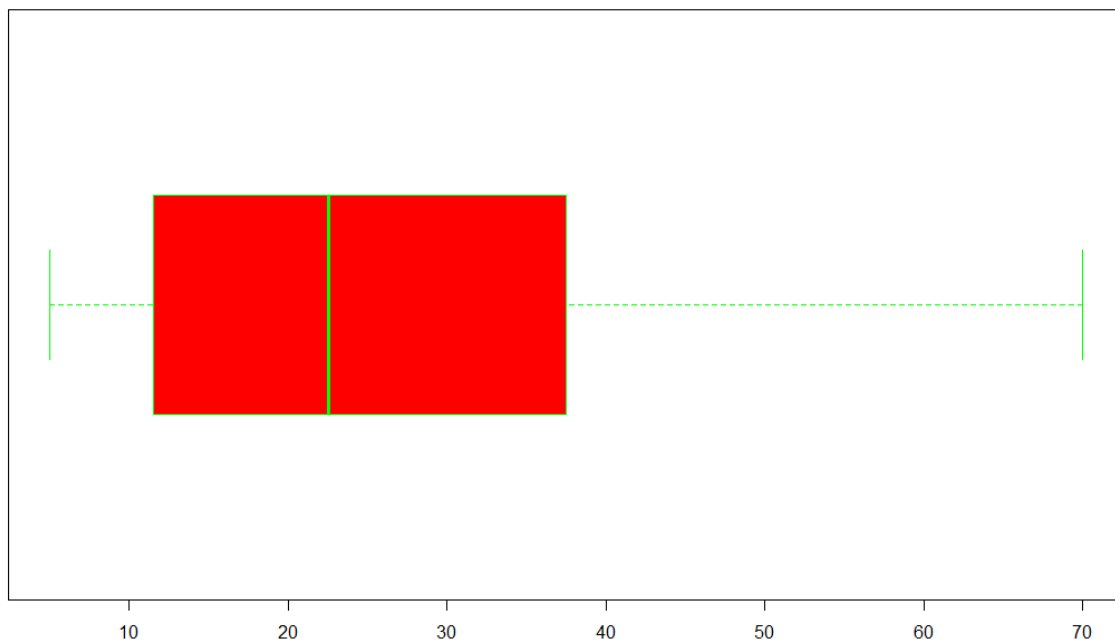
Tập dữ liệu

10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70

có giá trị nhỏ nhất là 5, lớn nhất là 70 và tứ phân vị là: 12.25 22.50 36.25.

Hãy vẽ biểu đồ hộp và râu?

```
> x=c(10, 8, 5, 13, 15, 25, 35, 20, 25, 40, 60, 70)
> boxplot(x,border="green",col="red",horizontal=TRUE)
```



### Example

Tính phương sai và độ lệch chuẩn mẫu của tập dữ liệu

1, 3, 3, 5, 8.

Số liệu mẫu: 1, 3, 3, 5, 8

```
> x=c(1, 3, 3, 5, 8)
> var(x)
[1] 7
> sd(x)
[1] 2.645751
```

### Example

Cho tập dữ liệu

10, 15, 32, 18, 25, 65, 30, 38.

Ta coi nó như là một mẫu thì phương sai là 297.2679 và độ lệch chuẩn là 17.24146.

(làm tương tự VD trên)

Tìm các số đo trung tâm của mỗi cột trong tập dữ liệu sau:

Thứ tự	Xếp loại	Giới tính	Lương
1	Kha	Nu	14
2	Kha	Nu	10
3	Kha	Nam	14
4	TrungBinh	Nu	12
5	TrungBinh	Nu	14
6	Kha	Nam	5
7	Gioi	Nam	11
8	Kha	Nu	6
9	Gioi	Nam	12
11	TrungBinh	Nu	10

- 1 So sánh lương trung bình của nhóm nam và nhóm nữ.
- 2 Trong các nhóm xếp loại tốt nghiệp khá, giỏi, trung bình, nhóm nào có lương trung bình lớn nhất?

```
> DL=edit(data.frame())
#Nhập bảng DL như trong đề bài, đóng cửa sổ nhập liệu.
> DL
  TT  XepLoai GioiTinh Luong
1  1      Kha      Nu    14
2  2      Kha      Nu    10
3  3      Kha      Nam    14
4  4 TrungBinh      Nu    12
5  5 TrungBinh      Nu    14
6  6      Kha      Nam     5
7  7      Gioi      Nam    11
8  8      Kha      Nu     6
9  9      Gioi      Nam    12
10 10 TrungBinh      Nu    10
```

1) So sánh lương TB của 2 nhóm Nam, Nữ

```
> attach(DL)
> tapply(Luong,list(GioiTinh),mean)
  Nam  Nu
10.5 11.0
```

Nên lương TB của nhóm Nam nhỏ hơn lương TB của nhóm Nữ

2)

```
> tapply(Luong,list(XepLoai),mean)
  Gioi      Kha TrungBinh
11.0    7.5    12.0
```

11.5

9.8

12.0

Vậy nhóm Trung bình có lương TB lớn nhất là 12.

Tính phương sai và độ lệch chuẩn của lương của mỗi nhóm nam và nữ trong tập dữ liệu sau, lương của nhóm nào đồng đều hơn?

Thứ tự	Xếp loại	Giới tính	Lương
1	Kha	Nu	14
2	Kha	Nu	10
3	Kha	Nam	14
4	TrungBinh	Nu	12
5	TrungBinh	Nu	14
6	Kha	Nam	5
7	Gioi	Nu	11
8	Kha	Nu	6
9	Kha	Nu	13
10	Gioi	Nam	12
11	TrungBinh	Nu	10
12	Kha	Nam	6

**HD:** Lập bảng DL như ví dụ trên, ở đây ta lấy tạm bảng DL đã có ở VD trên để thực hành. Với bảng dữ liệu trong ví dụ này cho kết quả gần tương tự.

```
> tapply(Luong,list(GioiTinh),var)
      Nam      Nu
15.0    9.2
> tapply(Luong,list(GioiTinh),sd)
      Nam      Nu
3.872983 3.033150
```

Lương nhóm Nam có phương sai là 15, độ lệch chuẩn là 3.872983

Lương nhóm Nữ có phương sai là 9.2, độ lệch chuẩn là 3.033150

Vậy lương nhóm Nữ đồng đều hơn trong bảng ở VD1.

**Example(làm việc với file excel)** Cho file dữ liệu điểm thi KT1.csv; tính điểm trung bình, phương sai và độ lệch chuẩn mẫu của điểm TKHP; tính cỡ mẫu n.

**HD:** #Kiểm tra thư mục R đang làm việc



```
> getwd()  
[1] "C:/Users/Phuong/Documents"
```

#Như vậy R đang làm việc với thư mục "C:/Users/Phuong/Documents"

#Copy file KT1.csv vào thư mục có đường dẫn trên (mặc định là Documents)

```
> DL=read.csv("KT1.csv")  
> attach(DL)  
> mean(TKHP)  
[1] 6.15035  
> var(TKHP)  
[1] 5.295898  
> sd(TKHP)  
[1] 2.301282  
> length(TKHP)  
[1] 143
```

Nên có n=143 sinh viên.

**Chú ý:** Nếu file excel gốc .xls, ta mở file excel .xls, chọn “Save As”, “Save as type” chọn lưu file dưới dạng đuôi .CSV(comma delimited).