Bài tập thực hành

XÁC SUẤT THỐNG KÊ

Bộ môn Toán - ĐẠI HỌC THĂNG LONG

Năm học 2008 - 2009

Mục lục

Trang

Bài 2

TÓM TẮT DỮ LIỆU

2.1 Tóm tắt dữ liệu bằng bảng tần số

Bảng tần số của tập dữ liệu

Trong phần này chúng ta sẽ xem xét một số hàm trong R hỗ trợ tóm tắt dữ liệu bằng bảng:

Tính tần số của tập dữ liệu

Để tính tần số của tập dữ liệu, ta dùng hàm:
`table(x, exclude = c(NA, NaN))`
trong đó

- `x` là véc tơ dữ liệu hoặc bảng dữ liệu cần tính tần số của các phần tử;

- `exclude` là tham số chỉ những phần tử không tham gia vào quá trình tính tần số, mặc định `exclude = c(NA, NaN)`, tức là không tính tần số những dữ liệu trống `NA(Not Available)` và những dữ liệu không phải số `NaN(Not a Number)`.

Giả sử ta có số liệu về điểm thi một môn học và đánh giá về thời gian tự học cho môn đấy được cho trong bảng dưới đây:

SinhVien	GioiTinh	Diem	DanhGiaTGTuHoc
1	Nam	7.0	Nhieu
2	Nam	6.0	BT
3	Nu	6.5	It
4	Nam	7.5	BT
5	Nu	6.5	BT
6	Nu	4.5	It
7	Nam	5.0	Nhieu
8	Nam	2.5	BT
9	Nam	5.0	It
10	Nu	6.0	BT
11	Nu	8.0	Nhieu
12	Nu	6.5	It
13	Nam	4.5	Nhieu
14	Nu	5.0	BT
15	Nam	6.0	It

Để nhập dữ liệu trên vào trong R ta thực hiện lệnh:

```
> GioiTinh = c("Nam", "Nam", "Nu", "Nam", "Nu", "Nu",
"Nam", "Nam", "Nam", "Nu", "Nu", "Nu", "Nam", "Nu",
"Nam")
> Diem = c(7.0, 6.0, 6.5, 7.5, 6.5, 4.5, 5.0, 2.5,
5.0, 6.0, 8.0, 6.5, 4.5, 5.0, 6.0)
> DanhGiaTGTuHoc = c("Nhieu", "BT", "It", "BT", "BT",
"It", "Nhieu", "BT", "It", "BT", "Nhieu", "It", "Nhieu",
"BT", "It")
> DuLieu = data.frame(GioiTinh, Diem, DanhGiaTGTuHoc)
```

Để tính số sinh viên nam và nữ trong bảng dữ liệu, tần số điểm hay tần số chéo giữa các cột dữ liệu, ta thực hiện lệnh các lệnh sau:

```
> table(GioiTinh)                    # tính số sinh viên nam và nữ
GioiTinh
 Nam  Nu
   8   7
> table(Diem)                        # tính tần số điểm
Diem
 2.5  4.5  5  6  6.5  7  7.5  8
   1    2  3  3    3  1    1  1
```

```
> table(DanhGiaTGTuHoc, exclude = "It")
```

```
# tính tần số sinh viên theo thời gian tự học trừ sinh viên học ít
```

```
DanhGiaTGTuHoc
   BT  Nhieu
    6      4
```

```
> table(Diem, GioiTinh)
```

```
# tính tần số chéo giữa điểm và giới tính
```

```
      GioiTinh
 Diem  Nam  Nu
  2.5    1   0
  4.5    1   1
    5    2   1
    6    2   1
  6.5    0   3
    7    1   0
  7.5    1   0
    8    0   1
```

```
> table(DanhGiaTGTuHoc, GioiTinh)
```

```
# tính tần số chéo giữa giới tính và thời gian tự học
```

```
                GioiTinh
 DanhGiaTGTuHoc  Nam  Nu
    BinhThuong     3   3
    It             2   3
    Nhieu          3   1
```

Kết quả trong R cho ta các thông tin:

- Bảng dữ liệu có 8 sinh viên nam và 7 sinh viên nữ;
- Tần số xuất hiện tương ứng của tám điểm khác nhau, chẳng hạn trong 15 sinh viên có hai sinh viên được điểm 4.5 hay có ba sinh viên được điểm 6.5;
- Tần số chéo giữa các cột, chẳng hạn trong ba sinh viên được điểm 6 có hai sinh viên nam và một sinh viên nữ hay có ba sinh viên nam và ba sinh viên nữ tự đánh giá thời gian học môn học này là bình thường.

Tính tần suất của tập dữ liệu

Để tính tần suất của tập dữ liệu, ta dùng hàm:

```
prop.table(x, margin = NULL)
```

trong đó

- `x` là véc tơ dữ liệu hoặc bảng dữ liệu cần tính tần suất của các phần tử;
- `margin` là tham số chỉ cách tính tần suất trong bảng dữ liệu hai chiều. Nếu `margin=1` thì tính tần suất của các phần tử trên mỗi hàng, nếu `margin=2` thì tính tần suất các phần tử trên mỗi cột. Mặc định `margin=NULL` tức là tính tần số cho mọi phần trong bảng dữ liệu.

Để tính tỉ lệ sinh viên nam và nữ trong bảng dữ liệu, tần suất điểm hay tần suất chéo giữa các cột dữ liệu, ta thực hiện các lệnh sau:

```
> prop.table(table(GioiTinh))  # tính tỉ lệ sinh viên nam và
                               nữ
GioiTinh
       Nam         Nu
 0.5333333  0.4666667

> prop.table(table(Diem))      # tính tần suất điểm
Diem
        2.5         4.5           5           6
 0.06666667  0.13333333  0.20000000  0.20000000
        6.5           7         7.5           8
 0.20000000  0.06666667  0.06666667  0.06666667

>   prop.table(table(GioiTinh, # tính tần suất chéo giữa giới
DanhGiaTGTuHoc))               tính và TG tự học
     DanhGiaTGTuHoc
 GioiTinh          BT          It       Nhieu
   Nam     0.20000000  0.13333333  0.20000000
   Nu      0.20000000  0.20000000  0.06666667

>   prop.table(table(GioiTinh, # tính tần suất theo hàng giữa
DanhGiaTGTuHoc), margin = 1)   giới tính và TG tự học
     DanhGiaTGTuHoc
 GioiTinh         BT         It      Nhieu
   Nam     0.3750000  0.2500000  0.3750000
   Nu      0.4285714  0.4285714  0.1428571
```

```
>   prop.table(table(GioiTinh,  # tính tần suất theo cột giữa
DanhGiaTGTuHoc), margin = 2)    giới tính và TG tự học
     DanhGiaTGTuHoc
 GioiTinh    BT    It  Nhieu
   Nam     0.50  0.40   0.75
   Nu      0.50  0.60   0.25
```

Kết quả trong R cho ta các thông tin:

- nam chiếm 53.33% và nữ chiếm 46.67% trong tổng số sinh viên;
- tần suất xuất hiện tương ứng của tám điểm khác nhau, chẳng hạn điểm 4.5 chiếm 13.33% và điểm 6.5 chiếm 20% trong tổng số điểm;
- tần suất chéo giữa các cột, chẳng hạn 20% sinh viên nam có thời gian học bình thường trong tổng số.
- tần suất theo từng mức, chẳng hạn theo mức thời gian tự học là bình thường thì nam chiếm 50% và nữ chiếm 50%.

Tính tần số, tần suất tích lũy của tập dữ liệu

Để tính tần số, tần suất tích lũy của tập dữ liệu, ta dùng hàm: `cumsum()` `(Cumulative Sum)`

```
cumsum(table(x))                  # tính tần số tích lũy
cumsum(prop.table(table(x)))      # tính tần suất tích lũy
```

Để tính tần số và tần suất tích lũy của điểm, ta thực hiện các lệnh sau:

```
> cumsum(table(Diem))                 # tính tần số tích lũy điểm
 2.5  4.5  5  6  6.5   7  7.5   8
   1    3  6  9   12  13   14  15

> cumsum(prop.table(table(Diem)))  # tính tần suất tích lũy điểm
        2.5         4.5           5           6
0.06666667  0.20000000  0.40000000  0.60000000
        6.5           7         7.5           8
0.80000000  0.86666667  0.93333333  1.00000000
```

Kết quả trong R cho ta các thông tin về:

- tần số tích lũy của điểm, chẳng hạn có 6 sinh viên trong tổng số có điểm không vượt quá 5, hay có 12 sinh viên trong tổng số có điểm không vượt quá 6.5;
- tần suất tích lũy của điểm, chẳng hạn có 20% sinh viên có điểm không vượt quá 5 hay có 80% sinh viên có điểm không vượt quá 6.5.

Phân tổ dữ liệu

Để tiến hành phân tổ dữ liệu, trong R ta dùng hàm

```
cut(x, breaks, labels = NULL, right = TRUE, include.lowest
= FALSE, dig.lab = 3, ordered_result = FALSE)
```

trong đó

- `x` là véc tơ dữ liệu dạng số cần được phân tổ;
- `breaks` là tham số dạng véc tơ số (ít nhất hai tọa độ) gồm các điểm chia hoặc là một số nguyên dương (lớn hơn hoặc bằng 2) chỉ số tổ mà tập dữ liệu sẽ phân thành;
- `labels` là tham số gán nhãn cho các khoảng chia, theo mặc định `labels = NULL`, các nhãn được xây dựng dưới dạng nửa khoảng $(a, b]$;
- `right` là tham số dạng logic, nếu `right = TRUE` khoảng chia có dạng $(a, b]$, nếu `right = FALSE` khoảng chia có dạng $[a, b)$;
- `include.lowest` là tham số dạng logic, mặc định `include.lowest = FALSE`, nếu `include.lowest = TRUE` thì trong trường hợp `right = TRUE` khoảng chia đầu tiên chứa giá trị nhỏ nhất của các điểm chia trong `breaks`, còn nếu `right = FALSE` khoảng chia cuối cùng chứa giá trị lớn nhất của các điểm chia trong `breaks`;
- `dig.lab` là tham số dạng số nguyên dương chỉ số chữ số trong điểm chia (trong trường hợp không gán nhãn cho các khoảng chia) và mặc định `dig.lab = 3`.

Ta có điểm thi tuyển sinh toán khối B vào Đại học Thăng Long năm 2008 trong file dữ liệu **DiemToanKhoiB.rda** và tính tần số được bảng sau:

Điểm	0	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5	2.75
Tần số	7	5	15	16	49	23	25	25	43	24	30	23
Điểm	3	3.25	3.5	3.75	4	4.25	4.5	4.75	5	5.25	5.5	5.75
Tần số	52	27	25	19	21	24	18	19	30	22	17	21
Điểm	6	6.25	6.5	6.75	7	7.25	7.5	7.75	8	8.25	8.5	9.5
Tần số	18	7	9	13	19	2	4	2	8	2	1	1

Có nhiều cách để phân chia tập dữ liệu về điểm trên thành các tổ khác nhau, ta có thể thực hiện sự phân chia theo các bước sau:

Bước 1: Xác định số tổ cần chia theo công thức $k = 1 + \log_2(N)$, với N là số giá trị khác nhau của tập dữ liệu. Ở đây thay $N = 36$ là số giá trị điểm khác nhau vào công thức trên ta có $k = 6.169925$. Ta lấy số tổ là $k = 6$.

```
> k = 1 + log(36, base = 2)
> k
[1] 6.169925
```

Bước 2: Xác định khoảng cách tổ $h = \dfrac{x_{\max} - x_{\min}}{k} = \dfrac{9.5 - 0.0}{6} = 1.583333.$ Ta lấy khoảng cách tổ $h = 1.5$.

```
> (max(KhoiBmoi) - min(KhoiBmoi))/6
[1] 1.583333
```

Bước 3: Xác định các tổ, chẳng hạn theo công thức:

Tổ 1: $[x_{\min}, x_{min} + h)$

Tổ 2: $[x_{\min} + h, x_{min} + 2h)$

...

Tổ k: $[x_{\min} + (k-1)h, x_{min} + kh)$

Ta có thể linh hoạt điều chỉnh tổ đầu và tổ cuối để đảm bảo tính khoa học và mỹ thuật. Chẳng hạn để phân tổ điểm toán khối B cho phù hợp, ta chọn các tổ như sau:

Tổ 1: $[0.0, 1.5)$; Tổ 2: $[1.5, 3.0)$; Tổ 3: $[3.0, 4.5)$;

Tổ 4: $[4.5, 6.0)$; Tổ 5: $[6.0, 7.5)$; Tổ 6: $[7.5, 9.5]$.

Trong R, ta dùng hàm `cut()` để xác định các tổ cần chia như sau:

```
> ChiaTo = table(cut(KhoiBmoi, breaks = c(0.0, 1.5,
3.0, 4.5, 6.0, 7.5, 9.5), right = FALSE, include.lowest
= TRUE))
> ChiaTo
 [0,1.5) [1.5,3) [3,4.5) [4.5,6) [6,7.5) [7.5,9.5]
     115     170     168     127      68        18
```

Tương tự như dữ liệu dạng định tính hay dữ liệu định lượng có ít biểu hiện, dữ liệu định lượng có nhiều biểu hiện sau khi phân tổ cũng có thể tính được tần số tích lũy, tần suất hay tần suất tích lũy của các tổ:

```
> cumsum(ChiaTo)                    # tính tần số tích lũy điểm theo
                                    các tổ
 [0,1.5) [1.5,3) [3,4.5) [4.5,6) [6,7.5) [7.5,9.5]
     115     285     453     580     648       666

> prop.table(ChiaTo)                # tính tần suất điểm theo các
                                    tổ
```

```
    [0,1.5)     [1.5,3)     [3,4.5)     [4.5,6)     [6,7.5)   [7.5,9.5]
 0.17267267  0.25525526  0.25225225  0.19069069  0.10210210  0.02702703
```

```
> cumsum(prop.table(ChiaTo))    # tính tần suất tích lũy điểm
                                  theo các tổ
   [0,1.5)    [1.5,3)    [3,4.5)    [4.5,6)    [6,7.5)  [7.5,9.5]
 0.1726727  0.4279279  0.6801802  0.8708709  0.9729730  1.0000000
```

Chú ý: Hàm `cut()` có thể tự chia các tổ khi biết số tổ cần chia, chẳng hạn trong ví dụ về phân tổ cho điểm toán khối B ở trên ta chỉ cần thực hiện:

```
> table(cut(KhoiBmoi, breaks = 6))
```

hoặc đơn giản

```
> table(cut(KhoiBmoi, 6))
 (-0.0095,1.58]  (1.58,3.16]  (3.16,4.75]
            140           197           153
  (4.75,6.34]  (6.34,7.92]  (7.92,9.51]
            115            49            12
```

Tuy nhiên, dùng hàm `cut(x, k)` các điểm chia thường không đẹp, và do cách làm tròn các chữ số trong điểm chia nên nhiều trường hợp ta thấy không phù hợp trong việc đếm tần số của các phần tử thuộc vào một khoảng. Ta lấy một ví dụ rất đơn giản sau:

```
> x = c(1,2,3,4,5,6)
> table(cut(x, 5))
 (0.995,2]  (2,3]  (3,4]  (4,5]  (5,6]
         1      1      2      1      1
```

Lí do dẫn tới việc đếm không chính xác tần số trong mỗi tổ (tổ 1, tổ 3) vì ở đây các chữ số trong điểm chia được làm tròn đến 3 theo mặc định của tham số `dig.lab = 3`. Trong trường hợp này ta phải điều chỉnh lại như sau:

```
> table(cut(x, 5, dig.lab = 4))
 (0.995,1.997]  (1.997,2.999]  (2.999,4.001]  (4.001,5.003]  (5.003,6.005]
             1              1              2              1              1
```

Mặc dù vậy không phải lúc nào ta cũng xác định ngay được `dig.lab` bằng bao nhiêu chữ số mới phù hợp, thậm chí có những trường hợp không thể xác định được. Chính vì những lí do trên mà khi thực hiện phân tổ dữ liệu mà sử dụng hàm `cut()`, ta nên để điểm chia cụ thể trong tham số `breaks` của hàm.

2.2 Tóm tắt dữ liệu bằng biểu đồ và đồ thị

Một số thiết bị đồ họa trong R

R cung cấp cho ta rất nhiều kiểu đồ họa phong phú và đa dạng. Kết quả của một hàm đồ họa không thể gán cho một đối tượng mà được gửi tới một thiết bị đồ họa (một thiết bị đồ họa là một cửa sổ đồ họa hoặc là một file). Ta có thể tìm hiểu danh sách các thiết bị đồ họa sẵn có trong R qua `?device`, ở đây ta chỉ tìm hiểu về một thiết bị đồ họa cho các file bitmap dạng `BMP`, `JPEG` và `PNG`:

```
bmp(filename = "TenFile.bmp", width = 480, height = 480, units = "px", bg = "white")
jpeg(filename = "TenFile.jpg", width = 480, height = 480, units = "px", quality = 75, bg = "white")
png(filename = "TenFile.png", width = 480, height = 480, units = "px", bg = "white")
```

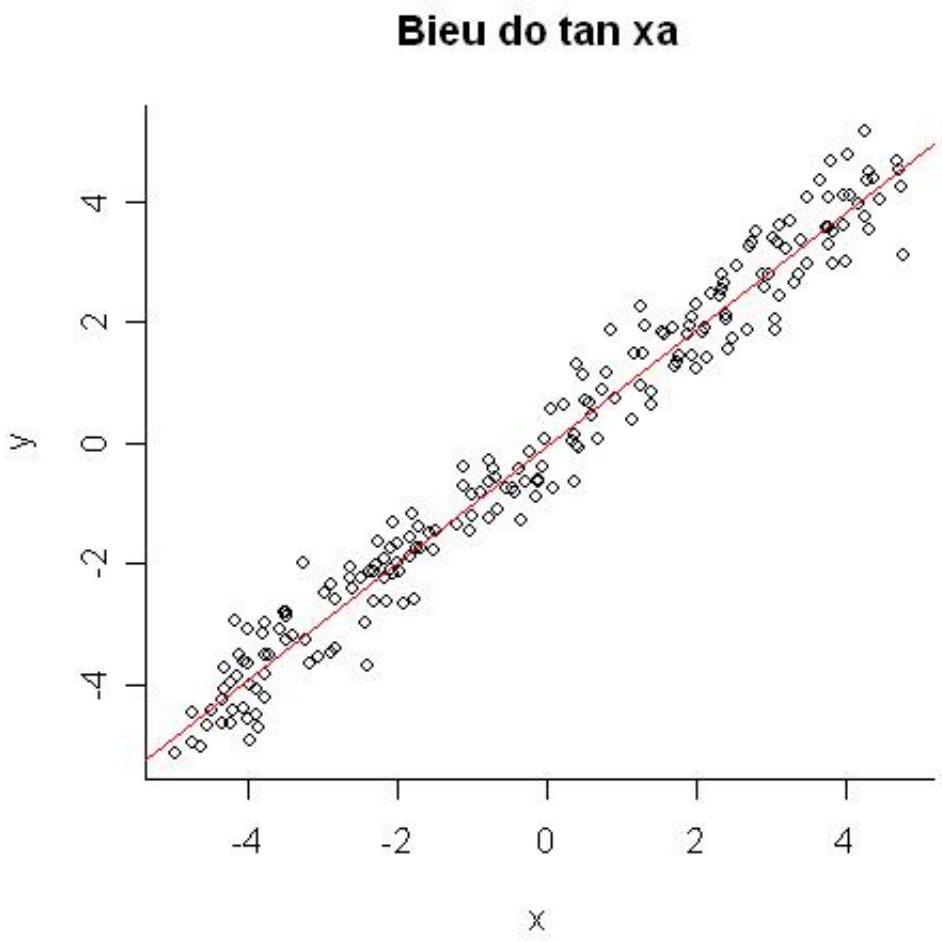
trong đó,

- `filename` là tham số chỉ tên của file đầu ra;
- `width` là tham số chỉ chiều rộng của thiết bị, mặc định là `480px`;
- `height` là tham số chỉ chiều cao của thiết bị, mặc định là `480px`;
- `units` là tham số chỉ đơn vị của chiều cao và chiều rộng, có thể là `inches`, `px(pixels)`, `cm`, `mm`, mặc định đơn vị là `px`;
- `bg(background)` là tham số chỉ màu của nền hình vẽ, mặc định là màu trắng;
- `quality` là tham số dưới dạng phần trăm chỉ chất lượng của ảnh dạng `JPEG`, mặc định là 75%;

Ví dụ ta cần vẽ một biểu đồ có tên là `BieuDoTanXa.jpeg` lưu tại `D:/BaiGiangXSTK`, ta thực hiện như sau:

```
> setwd("D:/BaiGiangXSTK")          # chọn địa chỉ để lưu biểu đồ
> jpeg(filename = "BieuDoTanXa.jpeg", bg="transparent")   # tạo biểu đồ dạng .jpeg
> N = 200                           # thực hiện các thao tác vẽ biểu đồ"
> x = runif(N, -5, 5)
> y = x + 0.5*rnorm(N)
> plot(x, y, main = "Bieu do tan xa", bty = "l")
```

```
> abline(lm(y ~ x), col =
"red")
> dev.off()                          # đóng thiết bị đồ họa
 null device
           1
```



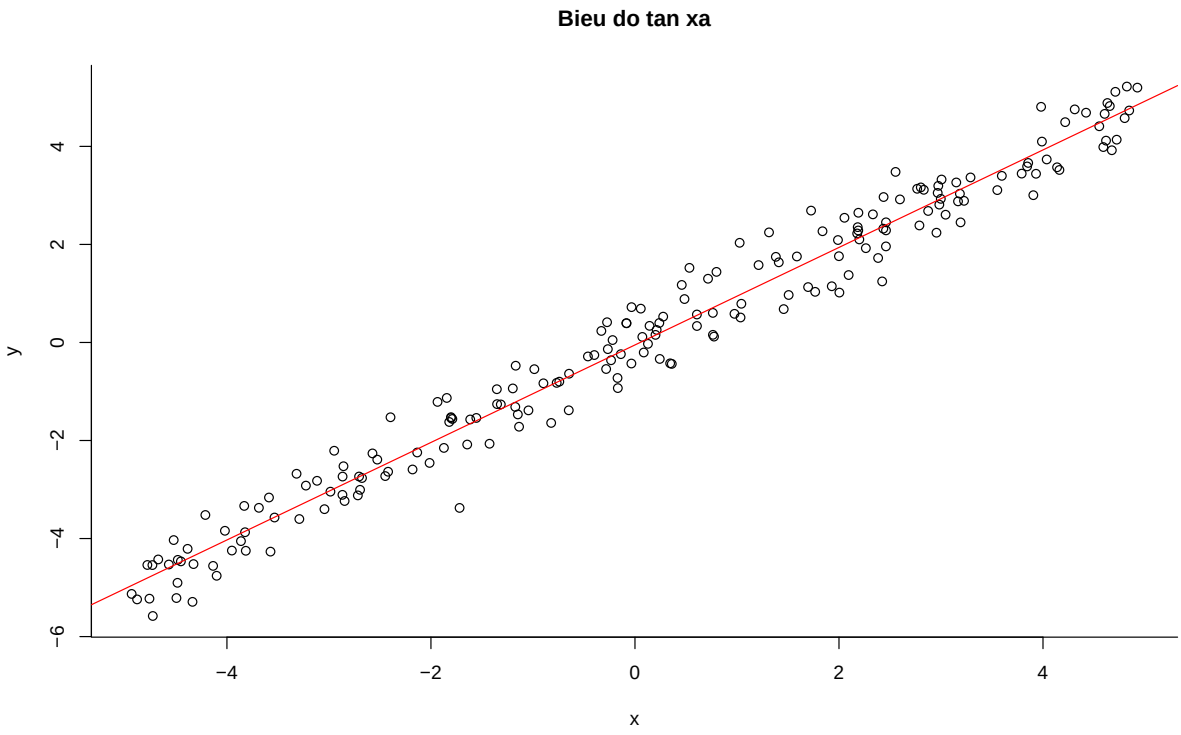
Nếu sử dụng hệ điều hành windows, sau khi thực hiện lệnh vẽ hình:

```
> N = 200
> x = runif(N, -5, 5)
> y = x + 0.5*rnorm(N)
> plot(x, y, main = "Bieu do tan xa", bty = "l")
> abline(lm(y ~ x), col = "red")
```

để lưu hình với tên `BieuDoTanXa.jpeg` tại `D:/BaiGiangXSTK`, ta chọn

- `file` → `save as` → `Jpeg` → `quality` 75%;
- `File name` là `BieuDoTanXa`;
- `Save as type` là `Jpeg files(*.jpeg, *.jpg)`;

- `Save in` tại `D:/BaiGiangXSTK`;
- `Save`



Thao tác trên nhiều cửa sổ đồ họa

Trong R để vẽ và xử lí nhiều hình trên cùng một thiết bị đồ họa, ta phân chia màn hình đồ họa thành nhiều cửa sổ, vẽ, xóa, đóng những hình trên từng cửa sổ thông qua các hàm sau:

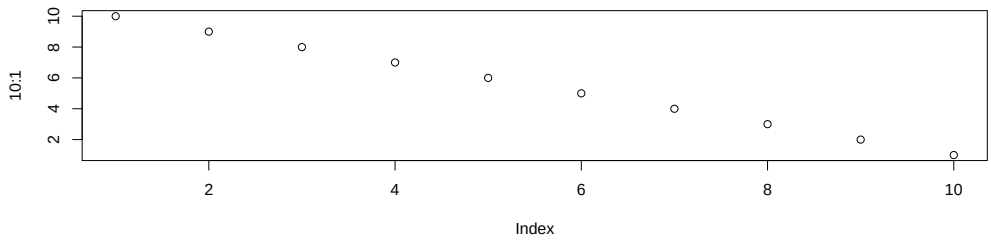
```
split.screen(figs, screen, erase = TRUE)
screen(n = , new = TRUE)
erase.screen(n = )
close.screen(n, all.screens = FALSE)
```

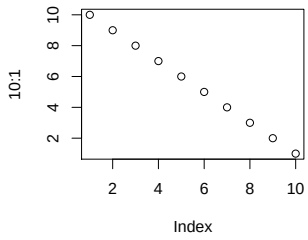
trong đó,

- `figs = c(m, n)` là tham số chỉ việc phân chia cửa sổ đồ họa thành m hàng và n cột;
- `screen` là tham số dạng số chỉ cửa sổ được phân chia;
- `erase = TRUE(FALSE)` là tham số dạng logic chỉ có (không) xóa hình vẽ trên cửa sổ được chọn;

- `n` là tham số dạng số chỉ việc chọn cửa sổ thứ n để thực hiện;
- `new = TRUE(FALSE)` là tham số dạng logic chỉ có (không) xóa hình trong cửa sổ n khi quay lại;
- `all.screens` là tham số dạng logic chỉ xem tất cả các cửa sổ có được xóa hay không, mặc định `all.screens = FALSE`.

```
> split.screen(c(2,1))                  # chia màn hình đồ họa thành hai cửa sổ trên, dưới
[1] 1 2
> split.screen(c(1,3), screen = 2)      #chia tiếp cửa sổ dưới thành 3 cửa sổ nhỏ (3,4,5)
[1] 3 4 5
> screen(1)                             # chọn thao tác trên cửa sổ trên (1)
> plot(10:1)                            # thực hiện vẽ hình
> screen(4)                             # chọn thao tác trên cửa sổ 4
> plot(10:1)                            # thực hiện vẽ hình
> close.screen(all = TRUE)              # màn hình đồ họa trở về bình thường
```





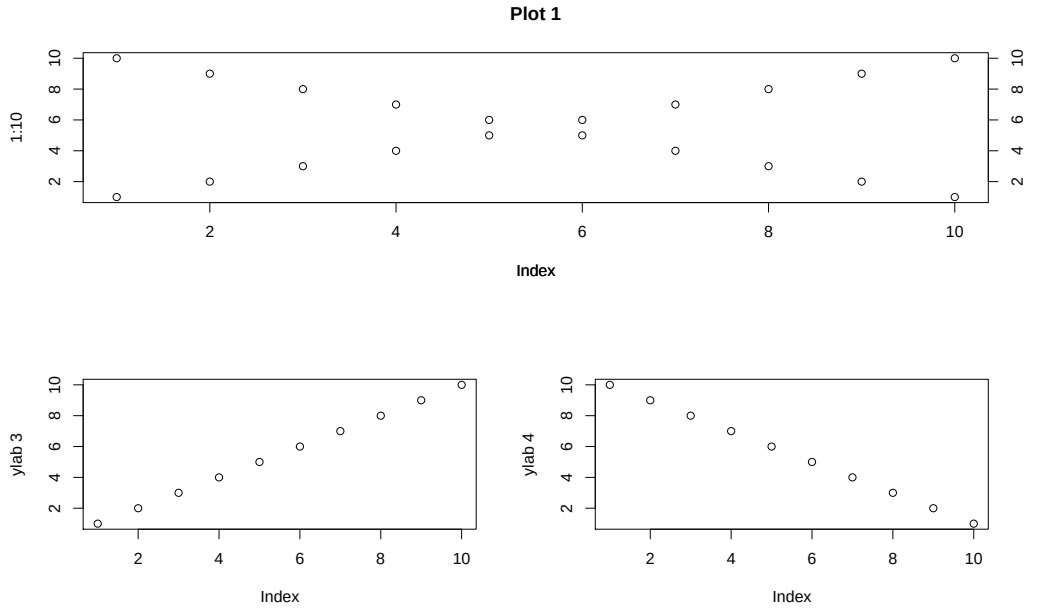

```
> split.screen(c(2,1))                  # chia cửa sổ đồ họa thành hai cửa sổ 1,2
> [1] 1 2
```

```
> split.screen(c(1,2),2)          # chia cửa sổ dưới làm hai cửa sổ 3,4

> [1] 3 4
> plot(1:10)                      # vẽ hình trên cửa sổ hiện hành 3

> plot(1:10, ylab= "ylab 3")      # quên gán nhãn và vẽ lại
> screen(1)                       # chọn cửa sổ 1
> plot(1:10)                      # vẽ hình
> screen(4)                       # chọn cửa sổ 4
> plot(1:10, ylab="ylab 4")       # vẽ hình
> screen(1, new = FALSE)          # quay về cửa sổ 1 nhưng không xóa hình

>    plot(10:1,     axes=FALSE,  # thêm hình vẽ thứ hai
lty=2, ylab="")
> axis(4)                         # thêm bảng chia trục bên phải
> title("Plot 1")                 # thêm tiêu đề
> close.screen(all = TRUE)        # trở về màn hình đồ họa bình thường
```



Hàm đồ họa

Có hai loại hàm đồ họa trong R: hàm vẽ hình bậc cao và hàm vẽ hình bậc thấp. Hàm vẽ hình bậc cao tạo ra hình vẽ mới, còn hàm vẽ hình bậc thấp thêm các yếu tố vào hình đang tồn tại. Các hình vẽ được tạo ra bởi các tham số đồ họa được định nghĩa sẵn trong mỗi hàm và có thể được thay đổi thông qua hàm

`par()`

Hàm vẽ hình bậc cao

Đây là một số hàm vẽ hình bậc cao thường gặp trong R:

`plot(x)`	tạo ra các điểm có tọa độ $(i, x_i), i = \overline{1,n}$ với $x = (x_1, x_2, \ldots, x_n)$
`plot(x, y)`	tạo ra các điểm có tọa độ $(x_i, y_i), i = \overline{1,n}$, với $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$
`hist(x)`	vẽ biểu đồ phân phối tần số của véc tơ dữ liệu x
`barplot(x)`	vẽ biểu đồ thanh thể hiện tần số của các phần tử trong véc tơ dữ liệu x
`pie(x)`	vẽ biểu đồ hình tròn của véc tơ dữ liệu x
`boxplot(x)`	vẽ biểu đồ hộp và râu của véc tơ dữ liệu x
`symbols(x, y)`	vẽ tại tọa độ cho bởi x, y hình tròn, hình chữ nhật, hình vuông, hình sao, hình nhiệt kế hoặc dạng biểu đồ hộp và râu với kích cỡ, màu sắc, ... trong các tham số phù hợp.

Để sử dụng chi tiết những hàm này vẽ hình, ta có thể đọc trong phần help tương ứng với mỗi hàm. Từng hàm có những tham số riêng biệt, tuy nhiên một số hàm đều dùng chung một số tham số sau (với giá trị mặc định cho tương ứng):

`add=FALSE`	nếu `add=TRUE` thì hình đang vẽ sẽ chồng lên hình vẽ trước (nếu có)
`axes=TRUE`	nếu `axes=FALSE` thì không vẽ các trục và hộp bao quanh
`type="p"`	miêu tả kiểu vẽ: `"p"(points)` dạng điểm; `"l"(lines)` dạng đoạn thẳng; `"b"(both points and lines)` dạng các điểm được nối bởi đoạn thẳng; `"o"(overstruck)` dạng các điểm được nối bởi đoạn thẳng nhưng đoạn thẳng đi qua các điểm; `"h"(histogram)` dạng thẳng đứng; `"s"(stair steps)` dạng bậc thang; `"n"(no plot)` không có kiểu gì cả
`xlim, ylim`	giới hạn của trục nằm ngang và trục thẳng đứng
`xlab, ylab`	tên của trục nằm ngang và trục thẳng đứng (kiểu kí tự)
`main`	tiêu đề của hình vẽ (kiểu kí tự)
`sub`	tiêu đề phụ của hình vẽ (kiểu kí tự)

Các hàm vẽ hình bậc thấp

`points(x)`	thêm các điểm $(i, x_i), i = \overline{1,n}$, vào hình vẽ với $x = (x_1, x_2, \ldots, x_n)$
`points(x, y)`	thêm các điểm tọa độ $(x_i, y_i), i = \overline{1,n}$ vào hình vẽ với $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$
`lines(x, y)`	thêm các đoạn thẳng nối các điểm (x_i, y_i) với $(x_{i+1}, y_{i+1}), i = \overline{1, n-1}$ vào hình vẽ
`text(x, y, labels)`	viết đoạn văn bản có nội dung ở `labels` tại tọa độ (x, y); kiểu sử dụng điển hình là `plot(x, y, type="n"), text(x, y, name)`
`mtext(text, side = 3)`	viết đoạn văn có nội dung ở `text` bên lề của cạnh thứ $k, k = 1, 2, 3, 4$ của hình
`segments` (x_0, y_0, x_1, y_1)	thêm đoạn thẳng nối điểm (x_0, y_0) và (x_1, y_1) vào hình
`arrows` $(x_0, y_0, x_1, y_1,$ `code=2)`	thêm đoạn thẳng nối điểm (x_0, y_0) và (x_1, y_1) cùng với mũi tên ở (x_1, y_1) nếu `code=2` và ở (x_0, y_0) nếu `code=1` và cả ở hai nếu `code=3`
`abline(a, b)`	thêm đường thẳng với tung độ gốc là `a` và độ dốc là `b` vào hình
`abline` $(h = y_0)$	vẽ thêm đường thẳng $y = y_0$ song song với trục nằm ngang vào hình
`abline` $(x = x_0)$	vẽ thêm đường thẳng $x = x_0$ song song với trục thẳng đứng vào hình
`abline` $(lm(y \sim x))$	vẽ thêm đường thẳng hồi qui tuyến tính mẫu vào hình
`rect` (x_1, y_1, x_2, y_2)	vẽ thêm hình chữ nhật mà trái, phải, dưới, trên tương ứng giới hạn bởi x_1, x_2, y_1, y_2
`polygon(x, y)`	vẽ thêm một đa giác nối các điểm có tọa độ ở x, y
`legend(x, y, legend)`	điền lời chú thích tại điểm (x, y) với nội dung ở `legend`
`title()`	điền tiêu đề và tiêu đề phụ của hình vẽ
`box()`	vẽ thêm khung bao quanh hình vẽ
`axis(side, vect)`	vẽ thêm trục vào hình vẽ, trục dưới nếu `side=1`, trục trái nếu `side=2` , trục trên nếu `side=3`, trục phải nếu `side=4`, nội dung điền trên mỗi trục qua `vect`

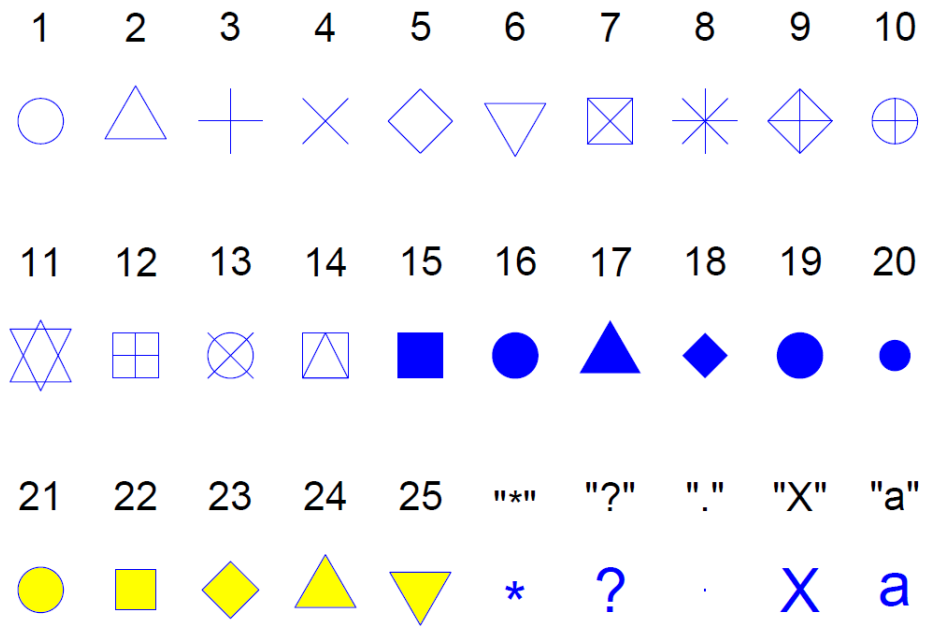
`rug(x)`	vẽ tại dữ liệu x trên trục nằm ngang những đoạn thẳng ngắn thẳng đứng

Các tham số đồ họa

Để điều chỉnh các chi tiết trong một hình vẽ, ta dùng các tham số đồ họa bên trong các hàm đồ họa để tạo ra những điều chỉnh. Có khoảng 73 tham số đồ họa và nhiều tham số dùng chung cho nhiều hàm, những lí giải cho từng tham số này có thể tìm thông qua `?par`. Ở đây ta sẽ liệt kê ra một số tham số đồ họa thường dùng:

`adj`	giá trị của `adj` căn chỉnh đoạn văn bản trong hình vẽ tính từ biên trái của đoạn văn bản: `adj=0` căn trái, `adj=0.5` căn giữa, `adj=1` căn phải, `adj>1` xa hẳn về bên trái, `adj<0` xa hẳn về bên phải
`bg`	điều chỉnh màu của nền màn hình đồ họa `background`, ví dụ `bg="blue", bg="lightyellow",...`
`bty`	điều chỉnh kiểu khung bao quanh hình vẽ, bty="n" không có khung bao quanh hình vẽ, bty="o" vẽ 4 cạnh quanh hình vẽ theo hình chữ `o`, bty="c" vẽ 3 cạnh quanh hình vẽ theo hình chữ `c`, bty="l" vẽ 2 cạnh quanh hình vẽ theo hình chữ `l`, bty="7" vẽ 2 cạnh quanh hình vẽ theo hình số `7`
`cex`	điều chỉnh cỡ chữ của văn bản hoặc kích cỡ của các biểu tượng trong hình vẽ, điều chỉnh số (chữ) trên các trục dùng `cex.axis`, điều chỉnh cỡ của tên trục dùng cex.lab, điều chỉnh cỡ của tiêu đề, tiêu đề phụ hình vẽ dùng `cex.main, cex.sub`
`col`	điều chỉnh màu của các biểu tượng, giống như `cex` có các hàm `col.axis, col.lab, col.main, col.sub`
`font`	số nguyên dương điều chỉnh kiểu của văn bản: `1:bình thường, 2:nghiêng, 3:đậm, 4:nghiêng,đậm`, giống như `cex` ta có `font.axis, font.lab, font.main, font.sub`
`las`	số nguyên dương điều chỉnh hướng tên của các trục: `0: song song với các trục, 1:nằm ngang, 2:vuông góc với các trục, 3:thẳng đứng`
`lty`	điều chỉnh kiểu của đoạn thẳng: `1:liền nét, 2:nét, 3:chấm, 4:chấm,nét, 5:nét dài, 6:hai nét`
`lwd`	dạng số điều chỉnh độ đậm của các đoạn thẳng

`mfcol`	là véc tơ dạng $c(nr, nc)$ chia màn hình đồ họa thành ma trận nr hàng và nc cột, các hình vẽ được vẽ theo cột
`mfrow`	là véc tơ dạng $c(nr, nc)$ chia màn hình đồ họa thành ma trận nr hàng và nc cột, các hình vẽ được vẽ theo hàng
`pch`	điều chỉnh kiểu của kí hiệu, một số kiểu kí hiệu điển hình liệt kê trong bảng 2.1 dưới đây
`xaxt`	nếu `xaxt="n"` thì không vẽ trục x (trục nằm ngang)
`yaxt`	nếu `yaxt="n"` thì không vẽ trục y (trục thẳng đứng)



Hình 2.1: Một số kí hiệu điển hình của tham số `pch`

Ví dụ về vẽ hình trong R

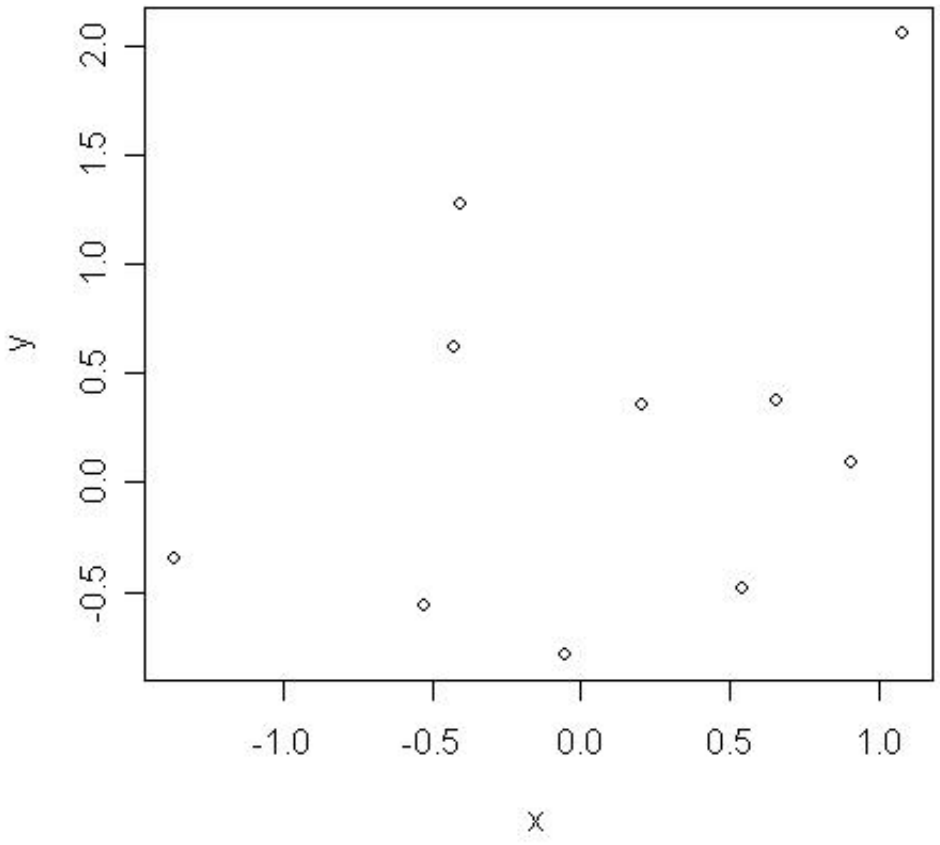
Để minh họa việc sử dụng những hàm đồ họa trong R, ta xét một ví dụ đơn giản về đồ thị của 10 điểm với các tọa độ được chọn ngẫu nhiên theo phân phối chuẩn hóa:

```
x = rnorm(10)
y = rnorm(10)
```

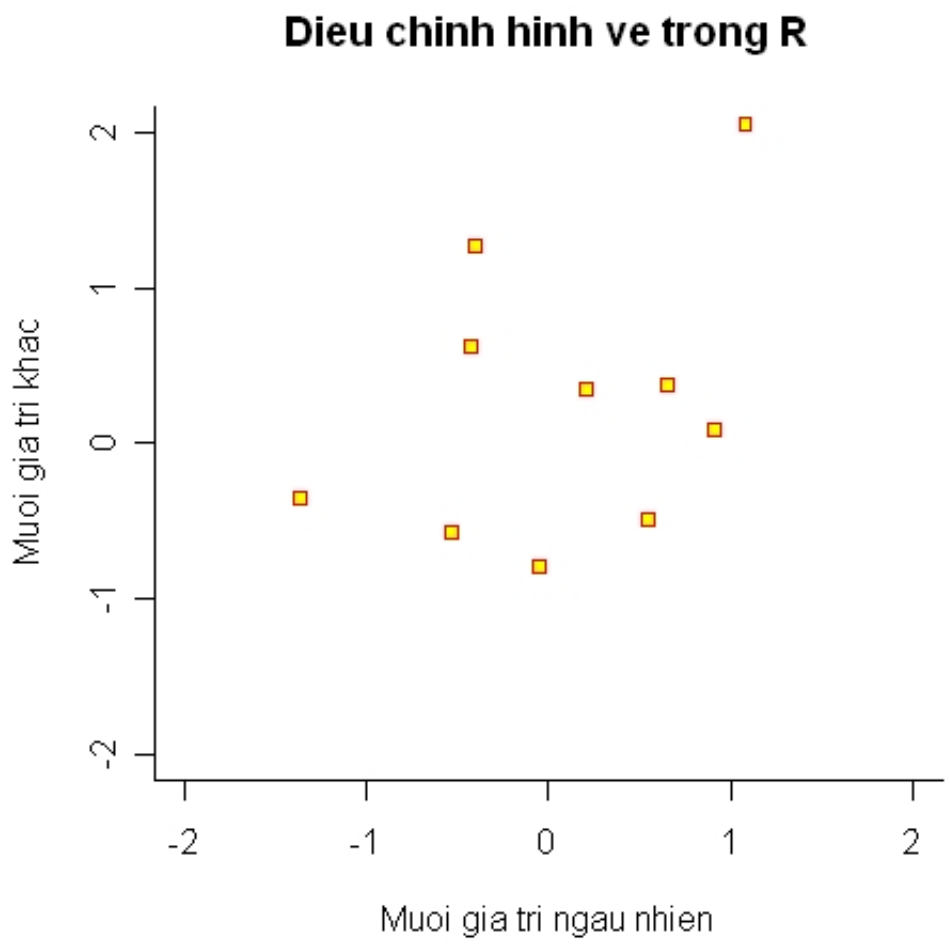
Để vẽ 10 điểm có tọa độ (x, y) trên mặt phẳng tọa độ ta dùng hàm `plot()`:

```
> plot(x, y)
```

và đồ thị ở hình 2.2 sẽ được vẽ theo những mặc định trong R. Theo mặc định, R tính toán và tạo ra các hình vẽ theo một cách "thông minh" nhất có thể.



Hình 2.2: Hình vẽ chưa điều chỉnh



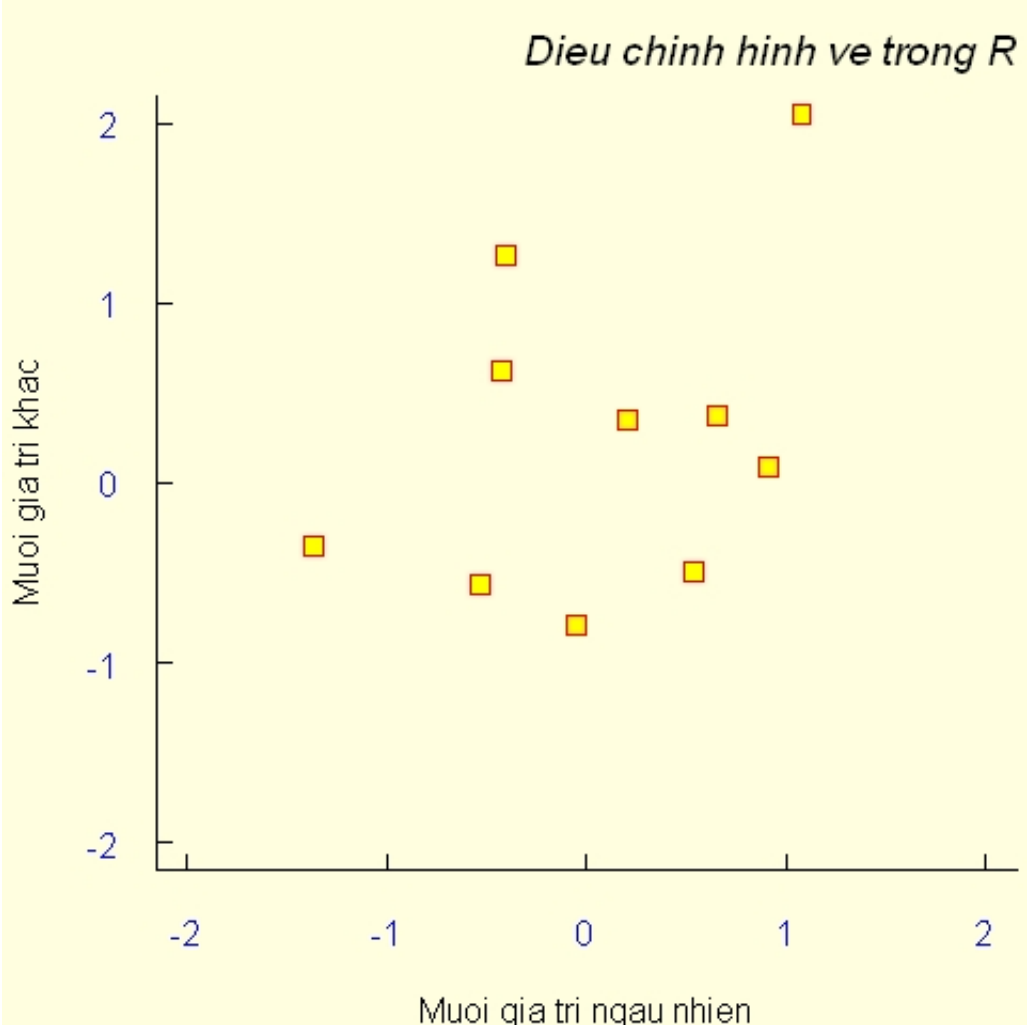
Hình 2.3: Hình vẽ điều chỉnh với hàm `plot`

Tuy nhiên, trong nhiều trường hợp ta cần trình bày hình vẽ theo những ý định riêng của mình và trong trường hợp đó ta cần điều chỉnh các tham số trong những hàm đồ họa để đạt được một hình vẽ theo ý muốn. Chẳng hạn, ta điều chỉnh để được hình vẽ 2.3 một cách có ý nghĩa theo cách sau:

```
> plot(x,y,xlab="Muoi gia tri ngau nhien",ylab="Muoi
gia tri khac",xlim=c(-2, 2),ylim=c(-2,2), pch=22,col="red",
bg="yellow", bty="l", main="Dieu chinh hinh ve trong
R")
```

Bây giờ ta sẽ điều chỉnh một vài tham số với hàm `par()` và hình vẽ 2.4 được tạo ra bởi các lệnh sau:

```
> opar = par()
> par(bg = "lightyellow", col.axis = "blue", mar =
c(4,4,2.5,0.25))
> plot(x, y, xlab = "Muoi gia tri ngau nhien", ylab
```



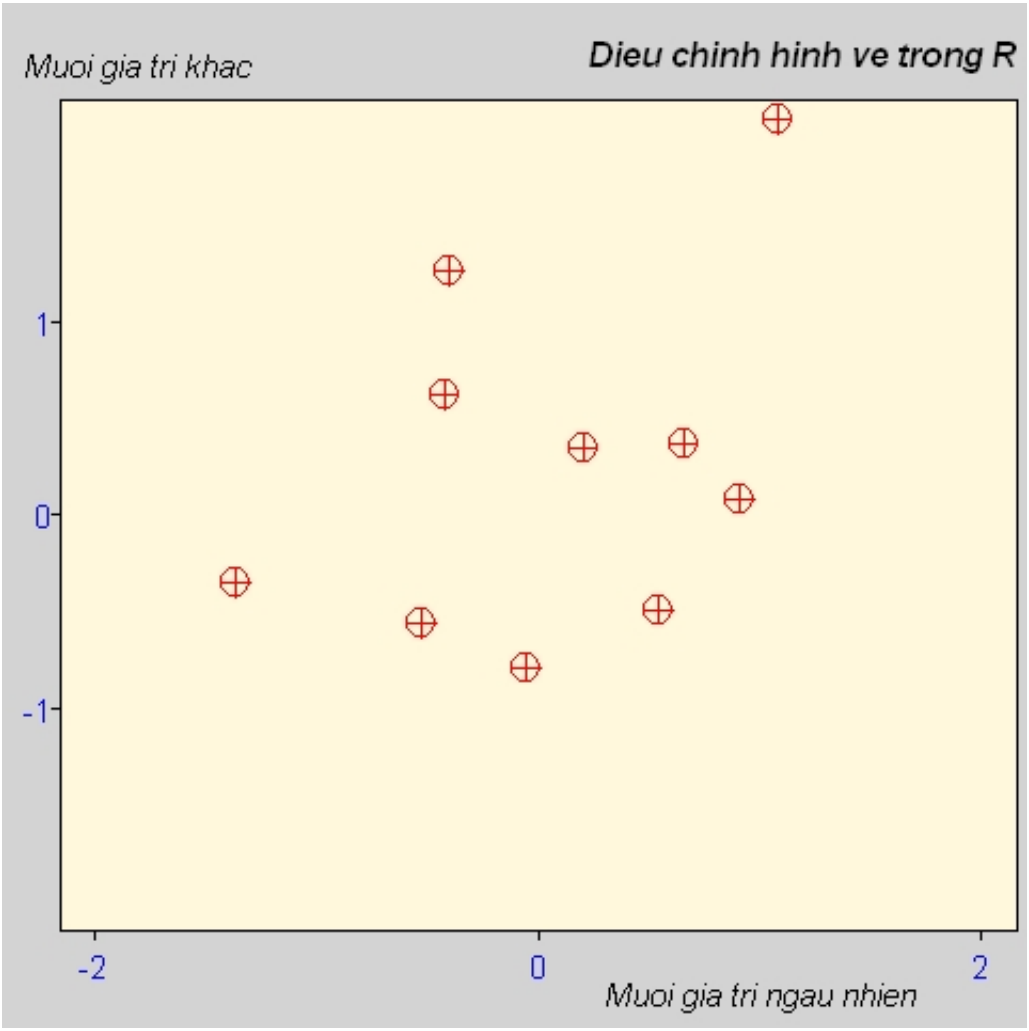
Hình 2.4: Điều chỉnh với hàm `par`, `plot` và `title`

```
= "Muoi gia tri khac", xlim = c(-2, 2), ylim = c(-2,
2), pch = 22, col = "red", bg = "yellow", bty = "l",
tcl = 0.4, las = 1, cex = 1.5)
> title("Dieu chinh hinh ve trong R", font.main = 3,
adj = 1)
> par(opar)
```

Bây giờ ta sẽ thực hiện sự điều chỉnh toàn bộ. Ta bắt đầu bằng việc tạo ra một hình vẽ "trống" và sau đó vẽ trục, đánh dấu các điểm trên trục, viết tên cho trục,... bằng cách sử dụng các hàm đồ họa bậc thấp. Hình vẽ 2.5 được tạo ra bởi các lệnh sau:

```
> opar = par()
> par(bg = "lightgray", mar=c(2.5, 1.5, 2.5, 0.25))
> plot(x, y, type = "n", xlab = "", ylab = "", xlim =
c(-2,2), ylim = c(-2,2), xaxt = "n", yaxt = "n")
```

```
> rect(-3,-3,3,3,col = "cornsilk")
> points(x,y, pch = 10, col=”red”, cex = 2)
> axis(side=1, c(-2,0,2), tcl = -0.2, labels = FALSE)
> axis(side=2, -1:1, tcl = -0.2, labels = FALSE)
> title("Dieu chinh hinh ve trong R", font.main=4,
adj=1, cex.main = 1)
> mtext("Muoi gia tri ngau nhien", side=1, line=1,at=1,
cex = 0.9,font=3)
> mtext("Muoi gia tri khac", line=0.5,at=-1.8, cex =
0.9,font=3)
> mtext(c(-2,0,2), side=1, las=1,at=c(-2,0,2), line=0.3,
col=”blue”, cex=0.9)
> mtext(-1:1,side=2,las=1,at=-1:1,line=0.2,col=”blue”,cex=0.9
> box()
> par(opar)
```



Hình 2.5: Hình vẽ tự tạo

Mô tả hình dáng của phân phối của tập dữ liệu định lượng

Để mô tả hình dáng của phân phối của tập dữ liệu định lượng, ta có thể dùng biểu đồ thân và lá (Stem and Leaf), biểu đồ phân phối tần số (Histogram), đa giác tần số (Frequency Polygon),... Biểu đồ thân và lá thường dùng để miêu tả phân phối của tập dữ liệu khi số quan sát của tập dữ liệu nhỏ (khoảng vài trục đến trăm), biểu đồ phân phối tần số và đa giác tần số thường dùng để miêu tả phân phối của tập dữ liệu khi số quan sát của tập dữ liệu lớn (khoảng vài trăm trở lên).

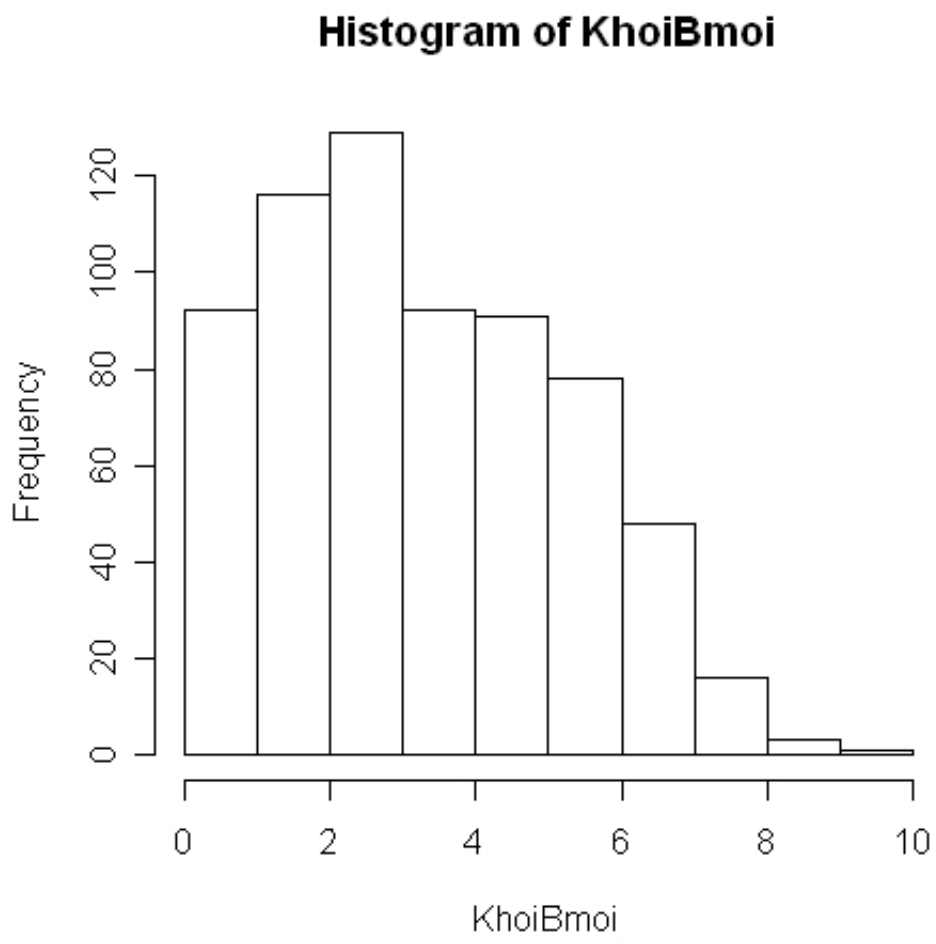
Biểu đồ phân phối tần số

Trong R để minh họa hình dáng của phân phối của tập dữ liệu bằng biểu đồ phân phối tần số, ta dùng hàm `hist()`

```
hist(x, breaks = "Sturges", freq = NULL, probability
= !freq,
include.lowest = TRUE, right = TRUE,
col = NULL, border = NULL,
main = paste("Histogram of" , xname),
xlim = range(breaks), ylim = NULL,
xlab = xname, ylab, labels = FALSE)
```

trong đó,

- `x` là véc tơ dữ liệu dạng số dùng để vẽ biểu đồ;
- `freq` là tham số dạng logic, nếu `freq = TRUE` các cột của biểu đồ biểu thị tần số, nếu `freq = FALSE` các cột của biểu đồ biểu thị tần suất;
- `breaks` là tham số chỉ véc tơ số (ít nhất hai tọa độ) gồm các điểm chia giữa các cột của biểu đồ hoặc là một số nguyên dương (lớn hơn hoặc bằng 2) chỉ số cột của biểu đồ;
- `right` là tham số dạng logic, nếu `right = TRUE` thì cột của biểu đồ lấy phần tử trong khoảng dạng $(a, b]$, nếu `right = FALSE` thì trong khoảng dạng $[a, b)$;
- `include.lowest` là tham số dạng logic, mặc định `include.lowest = TRUE`, nếu `include.lowest = TRUE` thì trong trường hợp `right = TRUE` cột đầu tiên chứa giá trị nhỏ nhất của các điểm chia trong `breaks`, còn nếu `right = FALSE` cột cuối cùng chứa giá trị lớn nhất của các điểm chia trong `breaks`;
- `col` là tham số chỉ màu của các cột;



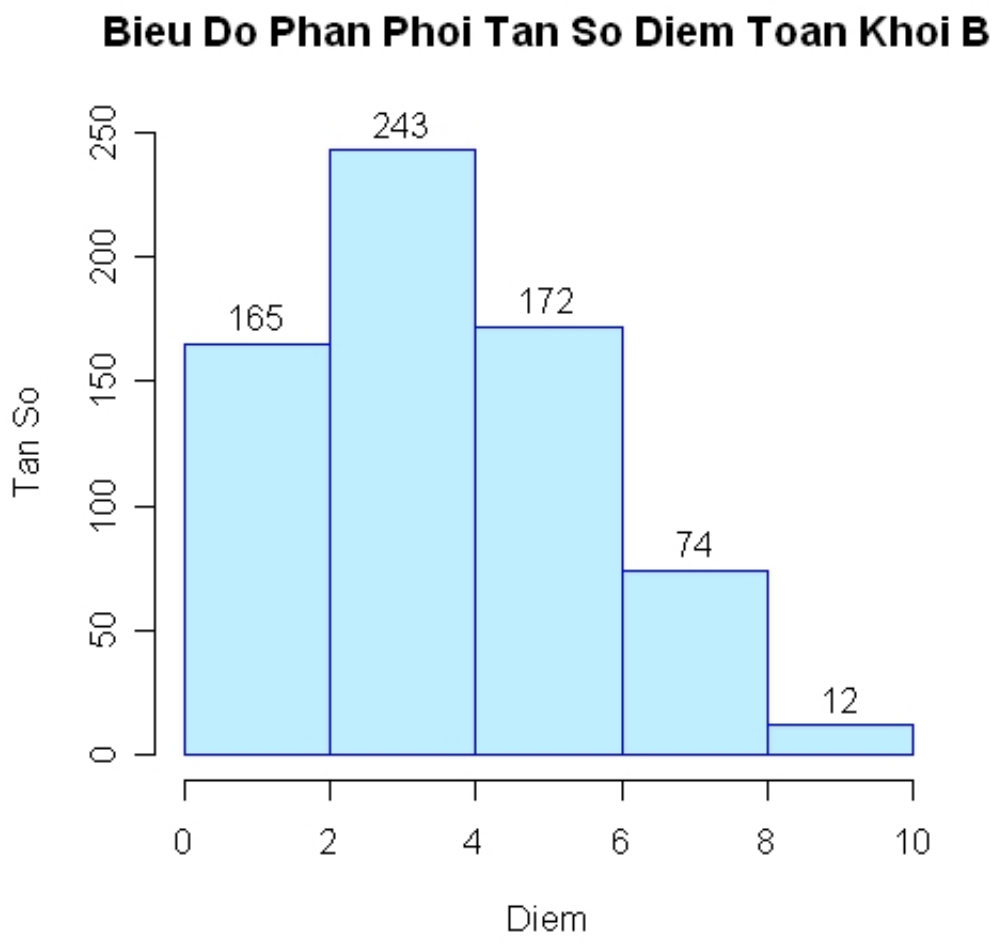
Hình 2.6: Biểu đồ phân phối tần số chưa điều chỉnh

- `border` là tham số chỉ màu của đường biên của các cột;
- `main`, `xlab`, `ylab` là những tham số chỉ tên của biểu đồ, tên trục x, y;
- `xlim`, `ylim` là những tham số giới hạn trên các trục
- `labels` là tham số dạng logic hoặc dạng kí tự điền tên trên đỉnh mỗi cột.

Bây giờ ta sẽ dùng biểu đồ phân phối tần số để minh họa hình dạng của phân phối điểm toán khối B năm 2008. Nếu không cần điều chỉnh biểu đồ theo ý muốn, đơn giản ta chỉ cần dùng lệnh

```
> hist(KhoiBmoi)
```

và được hình 2.6.



Hình 2.7: Biểu đồ phân phối tần số đã điều chỉnh

Ta cũng có thể điều chỉnh biểu đồ để phù hợp với số liệu và sinh động hơn bằng các lệnh

```
> hist(KhoiBmoi, xlim = c(0,10), ylim = c(0,250),
breaks = c(0.0 ,2.0, 4.0, 6.0, 8.0, 10.0), right = F,
xlab = "Diem", ylab = "Tan So", main = "Bieu Do Phan
Phoi Tan So Diem Toan Khoi B", xaxt = "n", yaxt = "n",
col = "lightblue1", border = "blue1", labels = T)
> axis(side = 1, c(0.0 ,2.0, 4.0, 6.0, 8.0, 10.0))
> axis(side = 2, c(0, 50, 100, 150, 200, 250))
```

và đạt được hình vẽ 2.7

Dựa trên hình dáng của biểu đồ phân phối tần số ta có thể biết được:

- Mức độ tập trung tương đối của phân phối của tập dữ liệu;
- Mức độ phân tán tương đối của phân phối của tập dữ liệu;

- Hình dạng tương đối của phân phối của tập dữ liệu là bằng phẳng, lệch hay cân đối.

Chẳng hạn, biểu đồ phân phối tần số của điểm toán khối B năm 2008 cho ta thấy điểm phân phối không đều tập trung ở những giá trị thấp, phân phối không đối xứng, nghiêng sang bên trái và ít phân tán.

Đa giác tần số

Đa giác tần số được vẽ bằng cách nối điểm chia đầu tiên với các trung điểm của các cột rồi nối với điểm chia cuối cùng nên để vẽ được đa giác tần số ta có thể dùng hàm `plot()`. Trong *R* ta có thể thực hiện như sau:

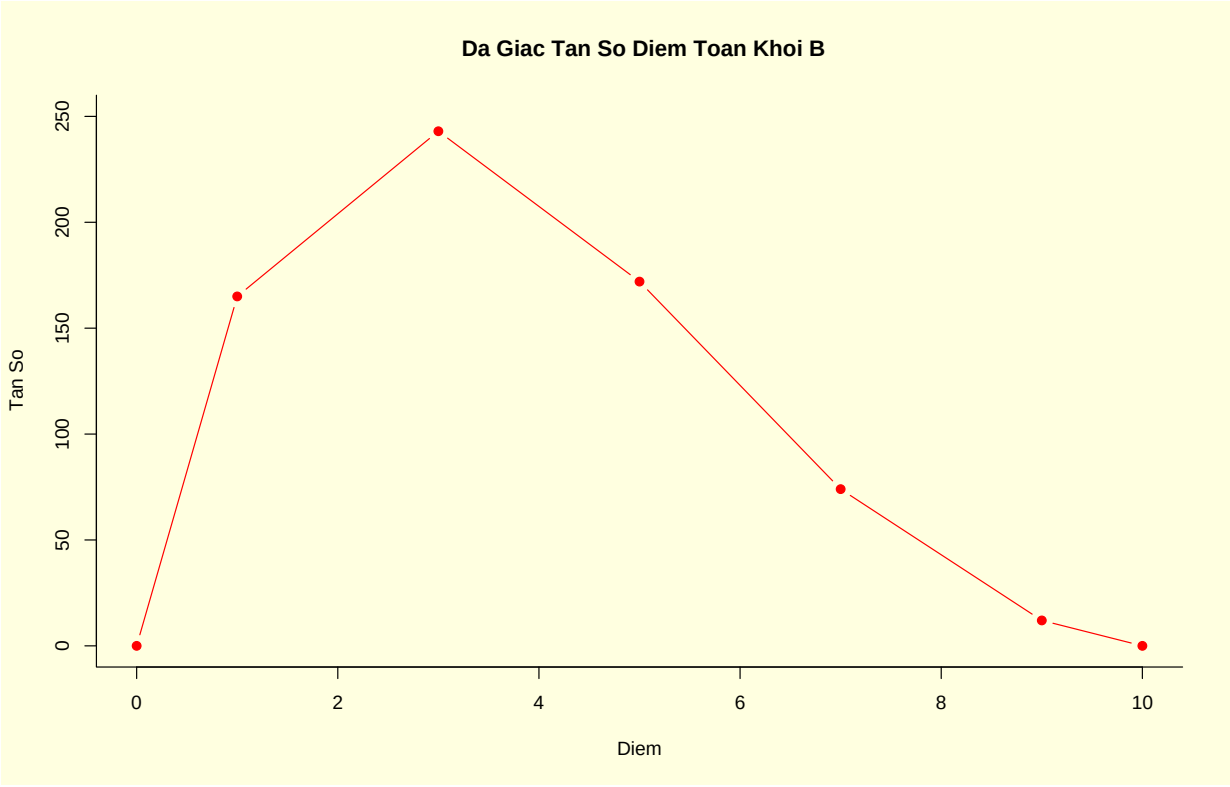
```
> par(bg = "lightyellow")
> B = hist(KhoiBmoi, xlim = c(0,10), ylim = c(0,250),
breaks = c(0.0 ,2.0, 4.0, 6.0, 8.0, 10.0), right = F)
> HoanhDo = c(min(B$breaks), B$mids, max(B$breaks))
> TungDo = c(0, B$counts, 0)
> plot(HoanhDo, TungDo, type = "b", xlim = c(0,10),
ylim = c(0,250), xlab = "Diem", ylab = "Tan So", main
= "Da Giac Tan So Diem Toan Khoi B", xaxt = "n", yaxt
= "n", col = "red", pch = 16, bty = "l")
> axis(side = 1, c(0.0 ,2.0, 4.0, 6.0, 8.0, 10.0))
> axis(side = 2, c(0, 50, 100, 150, 200, 250))
```

và thu được hình 2.8

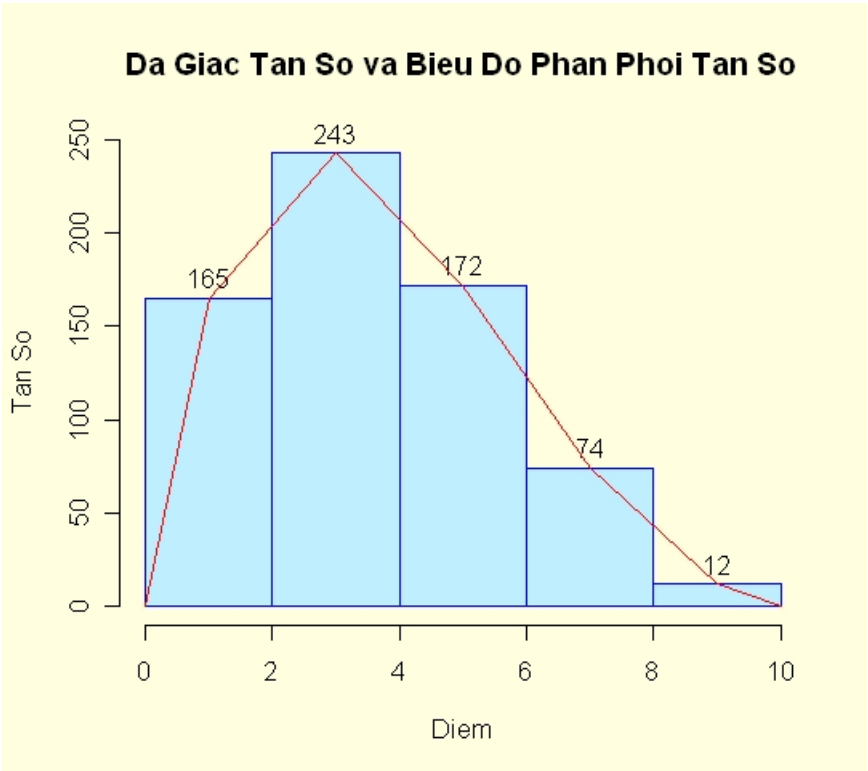
Ta cũng có thể kết hợp vẽ biểu đồ phân phối tần số và đa giác tần số trên cùng một hình bằng cách như sau:

```
par(bg = "lightyellow")
> hist(KhoiBmoi, xlim = c(0,10), ylim = c(0,250),
breaks = c(0.0 ,2.0, 4.0, 6.0, 8.0, 10.0), right = F,
xlab = "Diem", ylab = "Tan So", main = "Da Giac Tan
So Bieu Do Phan Phoi Tan So", xaxt = "n", yaxt = "n",
col = "lightblue1", border = "blue1", labels = T)
> HoanhDo = c(min(B $ breaks), B $ mids, max(B $
breaks))
> TungDo = c(0, B $ counts, 0)
> lines(HoanhDo, TungDo, col = "red")
> axis(side = 1, c(0.0 ,2.0, 4.0, 6.0, 8.0, 10.0))
> axis(side = 2, c(0, 50, 100, 150, 200, 250))
```

2.2. **Tóm tắt dữ liệu bằng biểu đồ và đồ thị**



Hình 2.8: Đa giác tần số



Hình 2.9: Đa giác tần số kết hợp biểu đồ phân phối tần số

Biểu đồ thân và lá

Để vẽ biểu đồ thân và lá trong R, ta sử dụng hàm `stem()`

```
stem(x, scale = 1, width = 80)
```

trong đó,

- `x` là véc tơ dữ liệu dạng số;
- `scale` là tham số điều chỉnh chiều dài của biểu đồ;
- `width` là tham số điều chỉnh chiều dài của biểu đồ theo mong muốn.

Với dữ liệu về tiền nước của 30 hộ gia đình trong một phường:
55, 50, 31, 57, 45, 65, 75, 36, 45, 55, 50, 52, 55, 51, 63, 81, 64, 70, 58, 59, 56, 58, 59, 65, 54, 55, 80, 56, 57, 58
để vẽ biểu đồ thân và lá của dữ liệu về tiền nước ở trên, trong R ta thực hiện như sau:

```
> TienNuoc = c(55, 50, 31, 57, 45, 65, 75, 36, 45, 55,
50, 52, 55, 51, 63, 81, 64, 70, 58, 59, 56, 58, 59,
65, 54, 55, 80, 56, 57, 58)
> stem(TienNuoc)
  The decimal point is 1 digit(s) to the right of the |
  3 | 16
  4 | 55
  5 | 00124555667788899
  6 | 3455
  7 | 05
  8 | 01
```

Do thân 5 quá dài nên ta có thể dùng tham số `scale` để điều chỉnh chiều dài của các thân

```
> stem(TienNuoc, scale = 2)
```

```
  The decimal point is 1 digit(s) to the right of the |
  3 | 1
  3 | 6
  4 |
  4 | 55
  5 | 00124
  5 | 5555667788899
  6 | 34
  6 | 55
  7 | 0
  7 | 5
  8 | 01
```

Biểu đồ thân và lá cũng như biểu đồ phân phối tần số cho ta cái nhìn tương đối về phân phối của tập dữ liệu, tức là ta có thể biết được sự tập trung, sự phân tán hay sự cân đối hay không của tập dữ liệu.

Mô tả hình dáng của phân phối của tập dữ liệu định tính

Để miêu tả hình dáng của phân phối của tập dữ liệu định tính ta có thể dùng biểu đồ thanh, biểu đồ tròn hoặc biểu đồ pareto. Biểu đồ thanh giúp ta có thể so sánh tần số (tần suất) của mỗi biểu hiện trong tập dữ liệu trong khi biểu đồ tròn giúp ta có thể so sánh tần số (tần suất) của mỗi biểu hiện so với toàn bộ các biểu hiện của tập dữ liệu.

Biểu đồ thanh

Trong *R* để vẽ biểu đồ thanh ta dùng hàm `barplot()`:

```
barplot(height, names.arg = NULL, legend.text = NULL,
beside = FALSE, horiz = FALSE,
col = NULL, border = par("fg"), main = NULL, sub =
NULL, xlab = NULL, ylab = NULL, xlim = NULL, ylim =
NULL)
```

- `height` là véc tơ hoặc ma trận dữ liệu dùng để vẽ biểu đồ;
- `names.arg` là tham số chỉ tên được viết dưới mỗi thanh hoặc nhóm các thanh trong biểu đồ;
- `legend.text` là một véc tơ gồm các kí tự hoặc dạng logic dùng để ghi chú thích trong biểu đồ;

- `beside` là tham số dạng logic, nếu `beside = FALSE` thì các cột của biểu đồ được vẽ chồng lên nhau, nếu `beside = TRUE` thì các cột được vẽ cạnh nhau;
- `horiz` là tham số dạng logic, nếu `horiz = FALSE` thì các cột được vẽ vuông góc với trục nằm ngang với cột đầu tiên nằm ở bên trái, nếu `horiz = TRUE` thì các cột được vẽ song song với trục nằm ngang với cột đầu tiên nằm ở dưới cùng;
- `col` là tham số chỉ màu của các cột;
- `border` là tham số chỉ màu của đường biên của các cột;
- `main`, `sub`, `xlab`, `ylab` là những tham số chỉ tên của biểu đồ, tên trục x, y;
- `xlim`, `ylim` là những tham số giới hạn trên các trục.

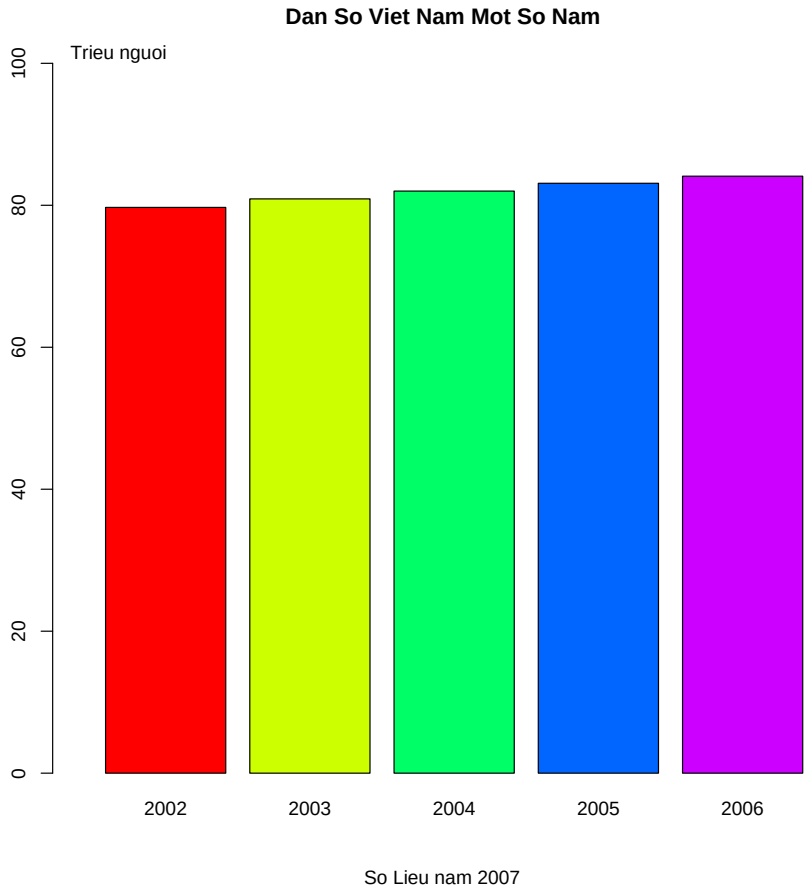
Bây giờ ta dùng biểu đồ thanh để minh họa dân số giữa năm (đơn vị triệu người) của Việt Nam và các nước trong khu vực Đông Nam Á trong một số năm được cho trong bảng sau:

Nước	2002	2003	2004	2005	2006
Vietnam	**79.7**	**80.9**	**82.0**	**83.1**	**84.1**
Brunei	0.3	0.3	0.4	0.4	0.4
Cambodia	13.1	13.3	13.5	13.8	14.2
Indonesia	211.4	214.3	217.1	219.9	222.1
Laos	5.3	5.4	5.5	5.6	5.7
Malaysia	24.5	25.1	25.6	26.1	26.6
Myanmar	52.2	53.2	54.3	55.4	56.5
Philippines	80.2	81.9	83.6	85.8	87.0
Thailand	63.1	63.7	64.2	64.8	65.2
Singapore	4.2	4.2	4.2	4.3	4.5

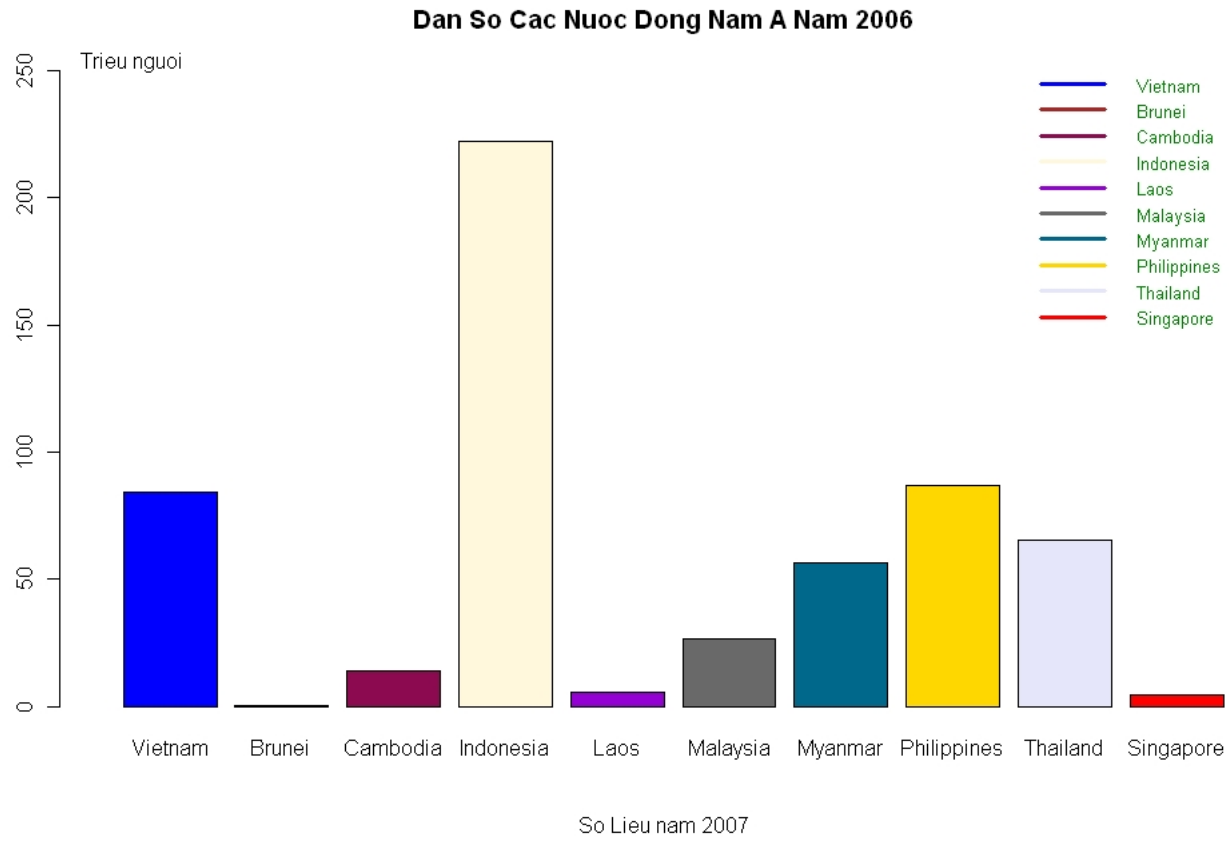
Trước hết ta dùng biểu đồ thanh để minh họa dân số của Việt Nam từ năm 2002 đến năm 2006. Trong R ta thực hiện các lệnh sau:

```
> SoDan = c(79.7, 80.9, 82.0, 83.1, 84.1)
> barplot(SoDan, main = "Dan So Viet Nam Mot So Nam",
sub = "So Lieu nam 2007",
names.arg = c("2002","2003","2004","2005","2006"),
col = rainbow(5), ylim = c(0,100), xlim = c(0,6))
> mtext("Trieu nguoi", at = 0.3)
```

và thu được hình 2.10



Hình 2.10: Biểu đồ thanh minh họa dân số Việt Nam những năm 2002-2006



Hình 2.11: Biểu đồ thanh minh họa dân số một số nước Đông Nam Á năm 2006

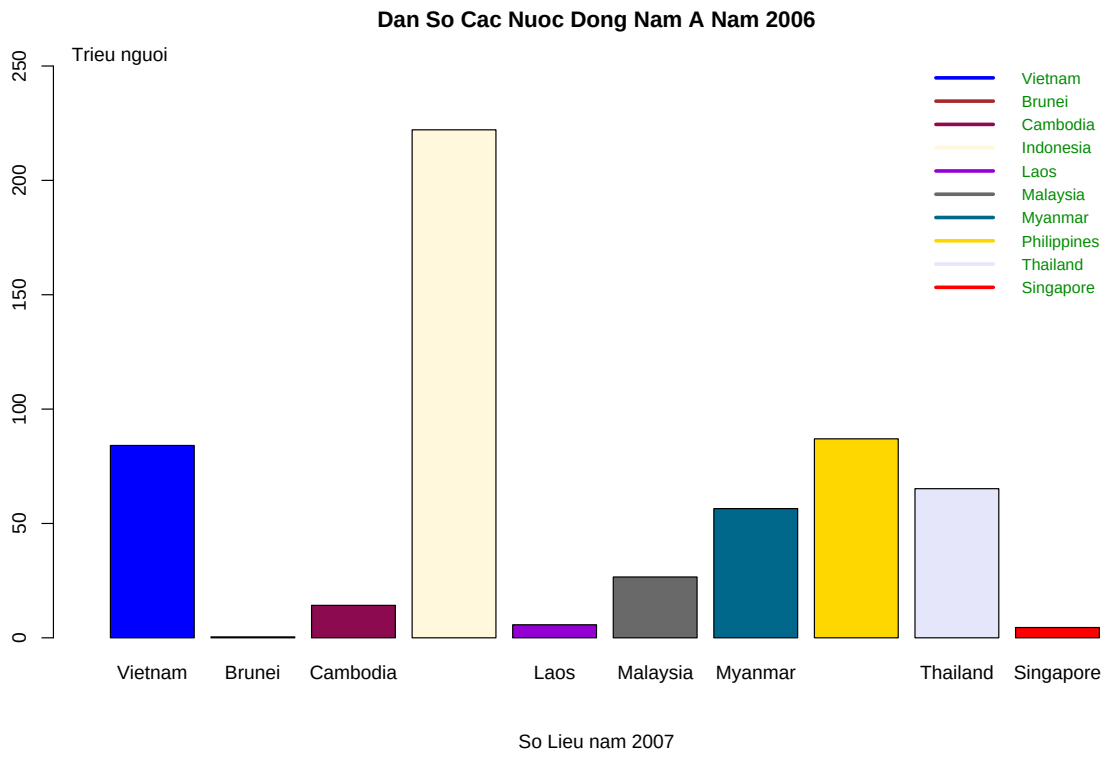
Biểu đồ thanh cũng giúp ta minh họa dân số của 10 nước khu vực Đông Nam Á năm 2006 và dân số của một số nước trong khu vực Đông Nam Á trong những năm 2002, 2004 và 2006 qua các hình 2.12 và 2.13

Biểu đồ tròn

Trong *R* để vẽ biểu đồ tròn ta dùng hàm `pie()`:

```
pie(x, labels = names(x), col = NULL, border = NULL,
lty = NULL, main = NULL)
```

- `x` là véc tơ dạng số thể hiện giá trị của mỗi hình quạt trong biểu đồ;
- `labels` là tham số chỉ tên của những hình quạt trong biểu đồ;
- `col` là tham số chỉ màu của các hình quạt;
- `border` là tham số chỉ màu của đường danh giới giữa các hình quạt;
- `main, sub` là những tham số chỉ tiêu đề và tiêu đề phụ của biểu đồ.



Hình 2.12: Biểu đồ thanh minh họa dân số một số nước Đông Nam Á năm 2006

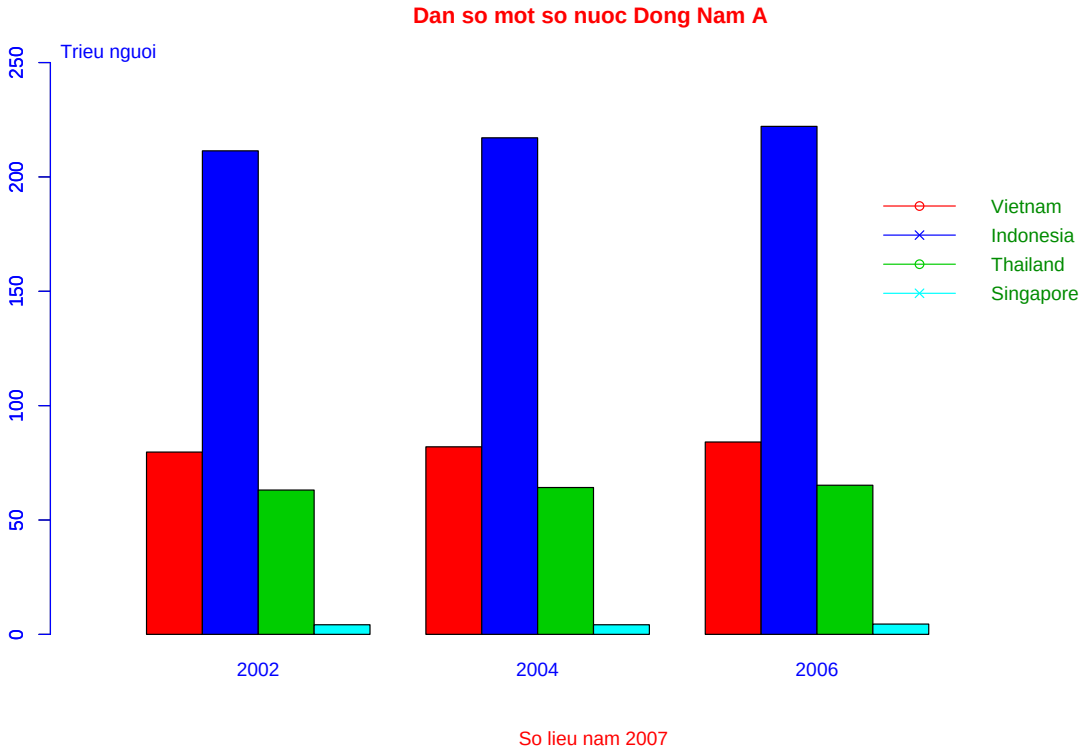
Ta sẽ dùng biểu đồ tròn để minh họa diện tích các châu trên thế giới được cho trong bảng số liệu sau:

Châu	Diện tích
Châu Phi	30306
Châu Mỹ	42049
Châu Á	31764
Châu Âu	22985
Châu Đại Dương	8537

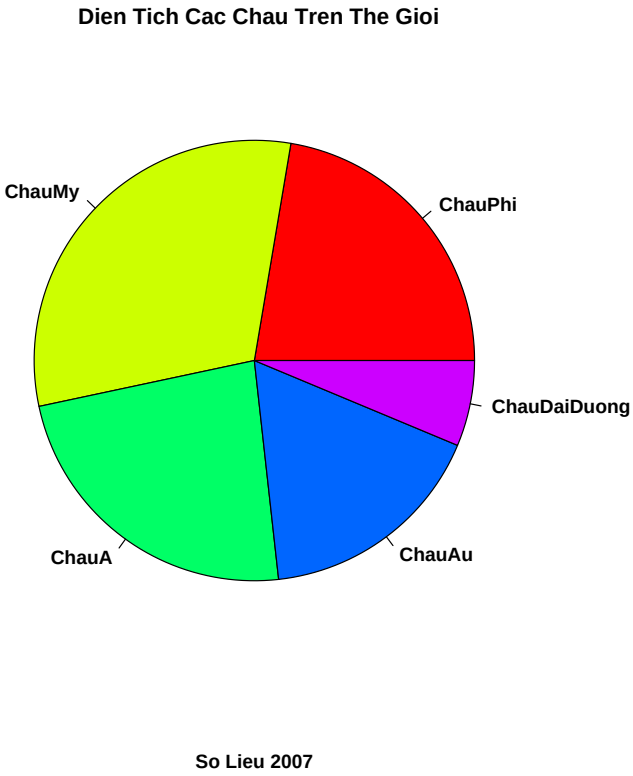
Để vẽ biểu đồ tròn minh họa diện tích của mỗi châu so với toàn bộ diện tích trên thế giới ta thực hiện các lệnh sau:

```
DienTich = c(30306, 42049, 31764, 22985, 8537)
pie(DienTich, col=rainbow(5), labels = c("ChauPhi",
"ChauMy", "ChauA", "ChauAu", "ChauDaiDuong"), main =
"Dien Tich Cac Chau Tren The Gioi", sub = "So Lieu
2007", main.font=4, font=2)
```

và thông tin được thể hiện qua hình 2.14.



Hình 2.13: Biểu đồ thanh minh họa dân số một số nước Đông Nam Á



Hình 2.14: Biểu đồ tròn minh họa diện tích các châu trên thế giới

2.3 Tóm tắt và trình bày dữ liệu bằng các đại lượng thống kê mô tả

Hàm tính các đại lượng thống kê mô tả

Bảng sau cho ta một số hàm trong R dùng để tính những đại lượng mô tả tập dữ liệu:

`mean(x)`	tính trung bình cộng của các giá trị cho trong véc tơ `x`
`median(x)`	tính trung vị của các giá trị cho trong véc tơ `x`
`which(table(x) == max(table(x)))`	cho các giá trị của mode của các giá trị cho trong véc tơ `x` và vị trí theo `table(x)` của những giá trị mode này
`summary(x)`	cho các giá trị lớn nhất, nhỏ nhất, tứ phân vị thứ nhất, thứ hai, thứ ba và trung bình của các giá trị cho trong véc tơ `x`
`quantile(x)`	tính phân vị tùy ý của dữ liệu cho trong véc tơ `x`
`range(x)`	cho giá trị nhỏ nhất và lớn nhất của dữ liệu cho trong véc tơ `x`
`var(x)`	cho phương sai của các giá trị cho trong véc tơ `x`
`sd(x)`	cho độ lệch chuẩn của các giá trị cho trong véc tơ `x`

Để mô tả về phân phối của điểm toán khối B năm 2008 ta có thể xét qua giá trị của các đại lượng thống kê mô tả:

```
> mean(KhoiBmoi)                    # tính điểm trung bình
[1] 3.433934
> median(KhoiBmoi)                  # tính trung vị
[1] 3
>      which(table(KhoiBmoi)==      # tính mode
max(table(KhoiBmoi)))
3                                    # có một giá trị mode là 3
13                                   # giá trị mode ở vị trí thứ 13
                                     khi đã lập bảng tần số
> summary(KhoiBmoi)                 # tính tứ phân vị
   Min.  1st Qu.  Median   Mean 3rd Qu.    Max.
  0.000    2.000   3.000  3.434   5.000  9.500
>                          quan-    # tính phân vị thứ 10 và 90
tile(KhoiBmoi,c(0.1,0.9))
```

```
   10%   90%
  1.00  6.25
> max(KhoiBmoi)-min(KhoiBmoi)   # tính khoảng biến thiên
[1] 9.5
> 5.000 - 2.000                 # tính độ trải giữa
[1] 3
> var(KhoiBmoi)                 # tính phương sai
[1] 3.868373
> sd(KhoiBmoi)                  # tính độ lệch chuẩn
[1] 1.966818
```

Các đại lượng vừa tính toán trên cho ta một số thông tin về phân phối của điểm toán khối B, chẳng hạn

- giá trị của trung bình cộng cho thấy điểm trung bình cộng của toàn bộ thí sinh là 3.4;
- giá trị của mode cho thấy điểm xuất hiện nhiều trong tập bài thi là điểm 3.0 và có 52 bài thi được điểm 3.0;
- giá trị của tứ phân vị cho biết 25% số thí sinh có điểm không vượt quá 2.0, 50% số thí sinh có điểm không vượt quá 3.0 và 75% số thí sinh có điểm không vượt quá 5.0;
- giá trị của phân vị thứ 10 và 90 cho ta thấy 10% thí sinh có điểm không vượt quá 1.0 và điểm thấp nhất trong nhóm 10% thí sinh có điểm cao nhất là 6.25;

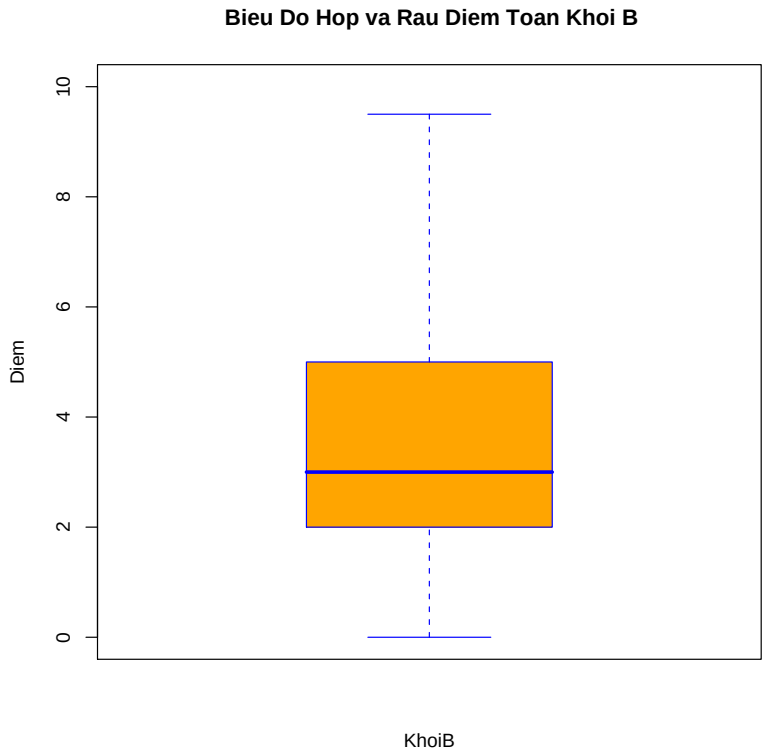
Biểu đồ hộp và râu

Biểu đồ hộp và râu có thể minh họa các đại lượng thống kê như trung vị, tứ phân vị và các giá trị ngoại biên trên cùng một hình vẽ. Để vẽ biểu đồ hộp và râu trong R ta dùng hàm `boxplot()`

```
boxplot(x, names, border, col=NULL, horizontal=FALSE)
```

trong đó,

- `x` là véc tơ dữ liệu số cần vẽ đồ thị;
- `names` là tham số ghi chú thích tên dưới mỗi biểu đồ;
- `border` là tham số chỉ màu của râu, đường biên của hộp và giá trị ngoại biên;



Hình 2.15: Biểu đồ hộp và râu dạng đứng minh họa điểm toán khối B

- `col` là tham số chỉ màu của hộp;
- `horizontal` là tham số logic chỉ cách vẽ biểu đồ, nếu `horizontal=FALSE` thì biểu đồ được vẽ đứng, nếu `horizontal=TRUE` thì biểu đồ được vẽ ngang.

Biểu đồ hộp và râu của điểm toán khối B được thể hiện qua hình 2.15 và được vẽ bởi lệnh sau:

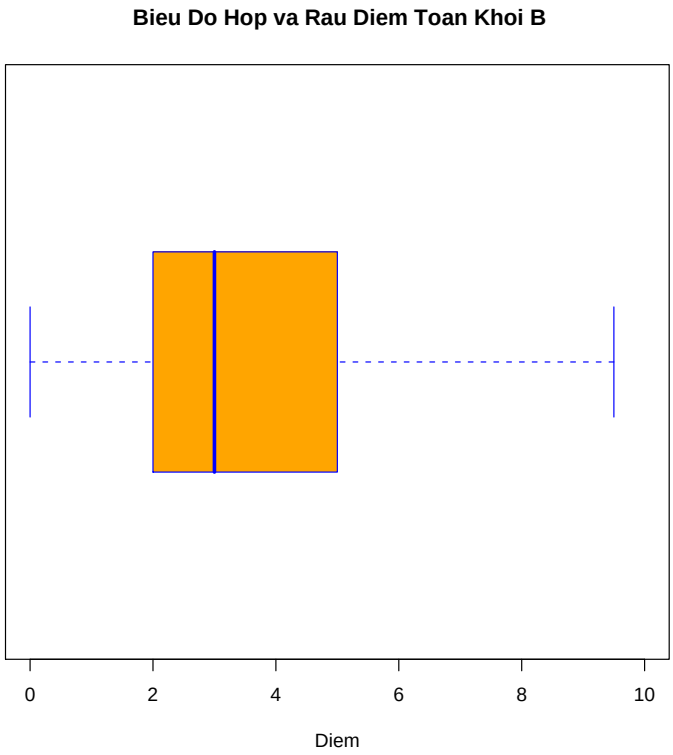
```
boxplot(KhoiBmoi, border="blue", col="orange", main="Bieu Do Hop va Rau Diem Toan Khoi B", ylim=c(0,10), xlab="KhoiB", ylab="Diem")
```

hoặc ta cũng có thể quay ngang biểu đồ trong hình 2.16 bằng cách thực hiện lệnh

```
boxplot(KhoiBmoi, border="blue", col="orange", main="Bieu Do Hop va Rau Diem Toan Khoi B", ylim=c(0,10), xlab="Diem", horizontal=TRUE)
```

Biểu đồ hộp và râu minh họa điểm thi toán khối B cho ta một số thông tin về những đại lượng thống kê mô tả phân phối của điểm toán khối B như:

- Trung vị của điểm khoảng 3.0, tứ phân vị thứ nhất khoảng 2.0, tứ phân vị thứ ba khoảng 5.0;

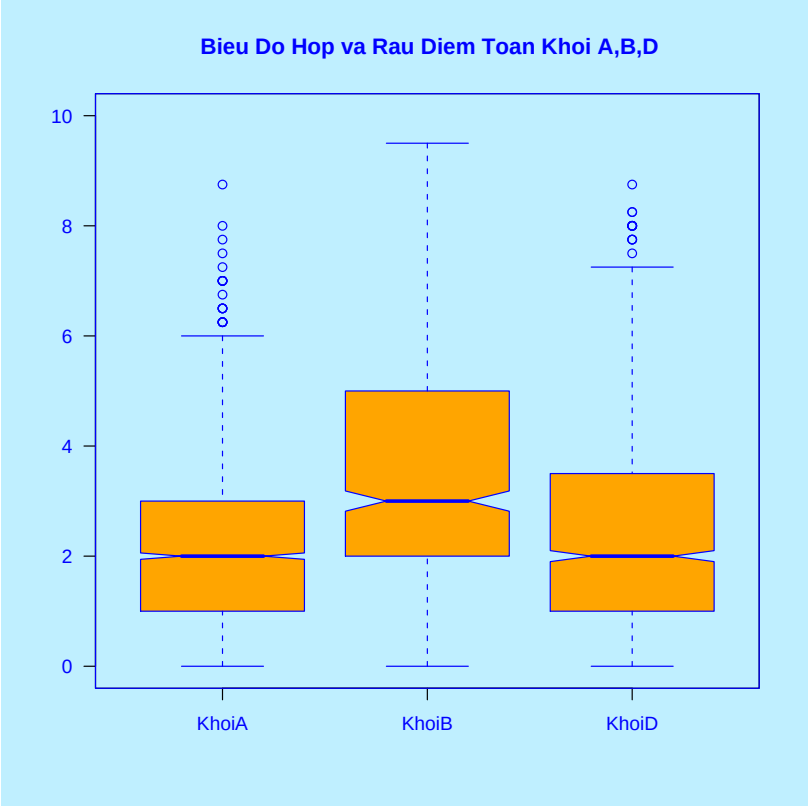


Hình 2.16: Biểu đồ hộp và râu dạng ngang minh họa điểm toán khối B

- Điểm toán khối B không có các điểm ngoại biên, điểm thấp nhất là 0.0 và điểm cao nhất khoảng 9.5;
- Phân phối của điểm toán khối B không đối xứng và nghiêng sang bên trái.

Nhận thấy những thông tin thu được về từ đồ thị hộp và râu cũng phù hợp với những tính toán và minh họa trước của ta bằng các đại lượng thống kê mô tả và biểu đồ phân phối tần số.

Ta cũng có thể so sánh những đại lượng thống kê mô tả của nhiều tập dữ liệu bằng cách vẽ nhiều biểu đồ hộp và râu tương ứng của từng tập dữ liệu trên cùng một hình vẽ. Chẳng hạn, để so sánh các đại lượng thống kê mô tả của điểm toán khối A, B và D ta có thể vẽ ba biểu đồ hộp và râu tương ứng của mỗi khối trên cùng hình 2.17



Hình 2.17: So sánh biểu đồ hộp và râu của khối A, B và D

BÀI TẬP

Bài 1: Trong file dữ liệu có tên là **SoLieu.csv** chứa một số thông tin cá nhân của 30 người về giới tính (GioiTinh), tuổi (Tuoi), khu vực sống (KhuVuc) và tổng thu nhập (đơn vị triệu VND) trong năm 2008 (ThuNhap). Hãy lấy file dữ liệu và thực hiện các yêu cầu sau:

a. Tính số nam sống ở hải đảo và nữ sống ở nông thôn trong nhóm những người được điều tra.

b. Trong số nữ được điều tra, hãy tính tỉ lệ nữ sống ở thành phố và miền núi.

c. Tiến hành phân tổ cột dữ liệu về tuổi thành các tổ với các điểm chia là $20, 30, 40, 50, 60$ và tính tỉ lệ những người được điều tra có độ tuổi không vượt quá 50.

d. Tiến hành phân tổ cột dữ liệu về thu nhập thành các tổ với các điểm chia là $20, 40, 60, 80, 100$ và tính:

i. tỉ lệ những người phải đóng thuế thu nhập nếu biết một người phải đóng thuế thu nhập nếu tổng thu nhập trong năm của người đó vượt quá 60 triệu VND.

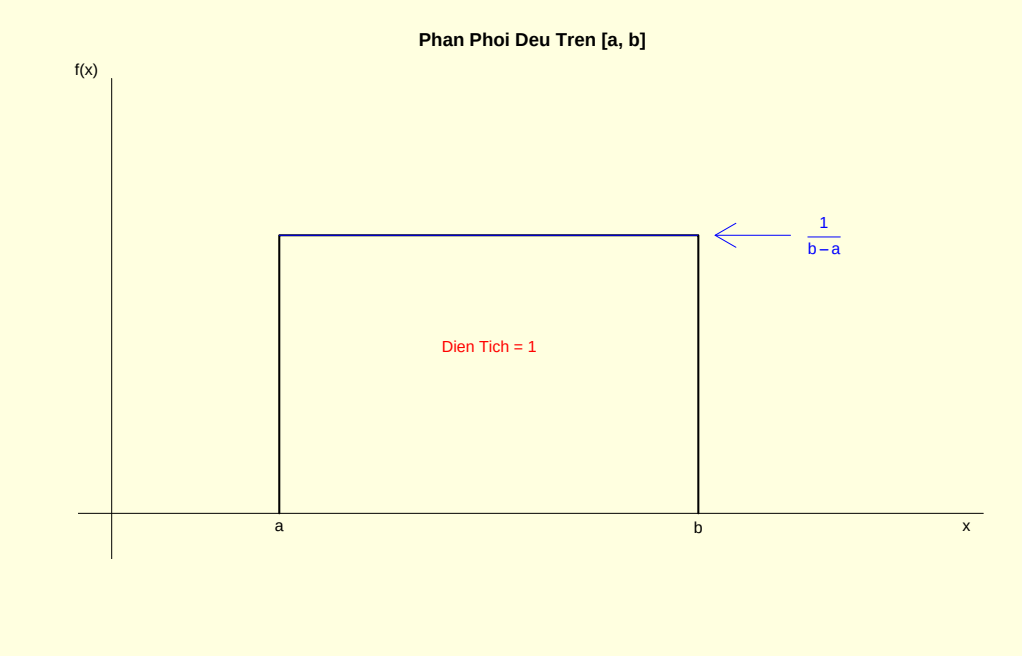
ii. tỉ lệ những người có thu nhập hơn 80 triệu nằm từ độ tuổi 40 đến 50.

Bài 2: Hãy phân tổ cho dữ liệu về điểm toán tuyển sinh khối D năm 2008 trong file **DiemToanKhoiD_2008.rda** với số tổ thích hợp và lập bảng tần số tương ứng với số tổ vừa phân.

Bài 3: Dùng tham số `mfrow` hoặc `mfcol` trong hàm `par` để phân chia cửa sổ đồ họa thành 3 hàng và 2 cột. Trên mỗi phần vừa phân chia này, hãy minh họa các kiểu vẽ khác nhau của tham số `type` thông qua hàm `plot()`

Bài 4: Dùng tham số `mfrow` hoặc `mfcol` trong hàm `par` để phân chia cửa sổ đồ họa thành 3 hàng và 2 cột. Trên mỗi phần vừa phân chia này, hãy minh họa các kiểu vẽ khác nhau của đường bao quanh hình vẽ của tham số `bty` thông qua hàm `plot()`

Bài 5: Sử dụng các hàm đồ họa với các tham số phù hợp để thu được hình vẽ sau:



Bài 6: Cho tập dữ liệu sau:

61	27	26	37	30	47
63	46	67	19	81	47
45	60	65	53	35	28
57	37	45	25	48	60
30	47	60	61	55	48

a. Hãy tính các đại lượng mô tả độ tập trung của tập dữ liệu: trung bình cộng, trung vị và mode. Nêu ý nghĩa của những giá trị này.

b. Tính các đại lượng mô tả độ phân bố của tập dữ liệu như: tứ phân vị, phân vị thứ $10, 60, 90$ của tập dữ liệu. Nêu ý nghĩa của những giá trị này.

c. Tính các đại lượng mô tả độ phân tán của tập dữ liệu: khoảng biến thiên, độ trải giữa, phương sai và độ lệch chuẩn. Những đại lượng này mô tả cho ta thông tin gì về tập dữ liệu?

Bài 7: Cho tập dữ liệu thu gọn sau:

Khoảng giá trị	7-9	9-11	11-13	13-15	15-17	17-19	19-21
Tần số	10	35	20	25	40	60	45

Tính trung bình, phương sai và độ lệch chuẩn của tập dữ liệu trên.

Bài 8: Hình dáng phân phối của dữ liệu sau tuân theo phân phối chuẩn:

61.4	27.3	26.4	37.4	30.4	47.5
63.9	46.8	67.9	19.1	81.6	47.9
73.4	54.6	65.1	53.3	71.6	58.6
57.3	87.8	71.1	74.1	48.9	60.2
54.8	60.5	32.5	61.7	55.1	48.2
56.8	60.1	52.9	60.5	55.6	38.1
76.4	46.8	19.9	27.3	77.4	58.1
32.1	54.9	32.7	40.1	52.7	32.5
35.3	39.1				

a. Lập bảng tần số của tập dữ liệu trên bằng cách phân tập dữ liệu thành các tổ có khoảng cách là 10.

b. Vẽ biểu đồ phân phối tần số và đa giác tần số tương ứng với bảng tần số trên.

c. Vẽ biểu đồ phân phối tần số và đa giác tấn số tương ứng với bảng tần số trên trên cùng một hình.

d. Dựa vào hình dáng của hai biểu đồ trên bạn có nhận xét gì hình dáng của phân phối chuẩn.

e. Tính trung bình $\overline{x}$ và độ lệch chuẩn s_x của tập dữ liệu.

f. Tính tỉ lệ phần trăm những giá trị của tập dữ liệu rơi vào khoảng $[\overline{x}-s_x, \overline{x}+s_x]$, $[\overline{x}-2s_x, \overline{x}+2s_x]$, $[\overline{x}-3s_x, \overline{x}+3s_x]$ và so sánh với qui tắc thực nghiệm.

Bài 9: Hãy vẽ biểu đồ phân phối tần số và đa giác tần số trên cùng một hình để minh họa hình dáng của phân phối điểm thi toán khối B năm 2008 dựa trên dữ liệu cho từ bảng tần số sau:

Khoảng điểm	Tần số
$[0, 1.5)$	115
$[1.5, 3)$	170
$[3, 4.5)$	168
$[4.5, 6)$	127
$[6, 7.5)$	68
$[7.5, 9.5]$	18

Bài 10: Bảng sau đây cho ta cho ta bảng giá của chỉ số chứng khoán công nghiệp Dow Jones trong 30 tuần khác nhau

2656	2301	2975	3002	2468
2742	2830	2405	2677	2990
2200	2764	2337	2961	3010
2976	2375	2602	2670	2922
2344	2760	2555	2524	2814
2996	2437	2268	2448	2460

a. Lập biểu đồ thân và lá cho 30 giá trị trên và đưa ra nhận xét.

b. Lập biểu đồ hộp và râu cho 30 giá trị trên và đưa ra nhận xét.

c. Từ hình dáng của hai biểu đồ ở câu a. và câu b. theo bạn biểu đồ thân và lá hay biểu đồ hộp và râu thể hiện tính đối xứng hay nghiêng trái, phải của tập dữ liệu rõ hơn?

Bài 11: Bảng dữ liệu sau cho số lượng album (triệu bản) được bán trong vài năm gần đây của một số thể loại âm nhạc:

Thể loại	Số lượng
R&B	146.4
Rock	102.6
Rap	73.7
Đồng quê	64.5
Cổ điển	14.8
Latin	14.5

Lập biểu đồ tròn biểu diễn phần trăm của mỗi thể loại nhạc so với toàn bộ các thể loại nhạc được nghiên cứu.

Bài 12: Lấy lại file dữ liệu **SoLieu.csv** và thực hiện các yêu cầu sau:

a. Vẽ biểu đồ thân và lá cho cột tuổi và hãy đưa ra nhận xét về hình dáng của phân phối của tuổi trong nhóm được điều tra.

b. Vẽ biểu đồ phân phối tần số với độ rộng mỗi cột là 10 cho cột thu nhập. Biểu đồ này cho ta thông tin gì về phân phối của thu nhập của nhóm được điều tra.

c. Vẽ biểu đồ thanh minh họa phân phối tần số của khu vực sống và đưa ra nhận xét.

d. Vẽ biểu đồ thanh minh họa phân phối tần số giới tính trong nhóm được điều tra theo khu vực sống và khu vực sống theo giới tính của nhóm được điều tra.

e. Hãy chọn một trong ba đại lượng là trung bình cộng, trung vị và mode mà bạn cho là thích hợp nhất để miêu tả độ tập trung cho mỗi cột dữ liệu và hãy tính những đại lượng này.

Bài 13: Điều tra tổng thu nhập (triệu VND) trong năm 2008 của một số chủ hộ gia đình được chọn ngẫu nhiên trong một phường ở Hà Nội ta thu được bảng số liệu sau:

30	20	35	27	37	30
38	40	42	35	38	44
41	42	47	48	43	50
55	60	70	45	47	49
78	55	60	85	90	100
75	80	85	95	105	47

a. Tính các đại lượng mô tả độ tập trung của tập dữ liệu: trung bình cộng, trung vị và mode. So sánh các giá trị này với nhau.

b. Lập biểu đồ thân và lá cho 30 giá trị trên và đưa ra nhận xét. Những nhận xét về phân phối của tập dữ liệu này có phù hợp với những tính toán ở câu **a.** không?

c. Tính tỉ lệ phần trăm những giá trị của tập dữ liệu rơi vào khoảng $[\overline{x}-s_x, \overline{x}+s_x]$, $[\overline{x}-2s_x, \overline{x}+2s_x]$, $[\overline{x}-3s_x, \overline{x}+3s_x]$, ở đây $\overline{x}$, s_x là trung bình cộng và độ lệch chuẩn của tập dữ liệu. So sánh kết quả với định lí Chebyshev.

Bài 14: Thống kê tiền điện (nghìn VND) trong tháng 8 năm 2008 của một số gia đình trong một quận ở Hà Nội thu được bảng số liệu sau:

5	30	100	150	600	210
85	95	105	120	185	250
310	1500	500	425	480	450
550	175	180	285	350	450
750	65	55	880	1350	270

a. Tính các đại lượng đo độ phân tán của tập dữ liệu: khoảng biến thiên, độ trải giữa và độ lệch chuẩn. So sánh các giá trị này với nhau.

b. Sử dụng đồ thị hộp và râu để minh họa sự phân bố của tập dữ liệu. Đồ thị này cho ta thông tin gì về giá trị ngoại biên của tập dữ liệu.

c. Từ các tính toán ở câu **a.** và **b.** theo bạn khoảng biến thiên hay độ lệch chuẩn đo độ phân tán cho tập dữ liệu trên tốt hơn.

Bài 15: Theo định lí Chebyshev:

a. ít nhất bao nhiêu phần trăm phần tử của tập dữ liệu rơi vào khoảng $[\mu - k\sigma, \mu + k\sigma]$ với những giá trị k sau: $k = 1.8$, $k = 3.5$, $k = 2.5$ và $k = 4$.

b. trong vòng bao nhiêu độ lệch chuẩn từ trung bình chứa ít nhất 80% giá trị của tập dữ liệu.

Bài 16: Một tập dữ liệu số có phân phối xấp xỉ hình chuông đối xứng. Nếu trung bình của các số là 125, độ lệch chuẩn là 12, hãy tìm khoảng giá trị mà:

a. 68% giá trị của tập dữ liệu rơi vào.

b. 95% giá trị của tập dữ liệu rơi vào.

c. 99.7% giá trị của tập dữ liệu rơi vào.

Bài 17: Cho một tập dữ liệu số có phân phối không tuân theo phân phối chuẩn. Nếu trung bình các số là 38 và độ lệch chuẩn là 6, hãy tính xem:

a. bao nhiêu phần trăm giá trị của tập dữ liệu rơi vào khoảng 26 và 50?

b. bao nhiêu phần trăm giá trị của tập dữ liệu rơi vào khoảng 14 và 62?

c. 89% giá trị của tập dữ liệu rơi vào khoảng hai giá trị nào?