Chapter 13: Descriptive Statistics – Thống kê mô tả

Ex2: Life of battery

Bài toán 1: Một nhà sản xuất đang điều tra tuổi thọ hoạt động của pin máy tính xách tay (Battery 1). Các quan sát được liệt kê trong tập tin life_batteries.txt.

1. Tạo life_array từ nội dung tập tin.
2. Vẽ biểu đồ phân phối tần suất của life_array
3. Biểu đồ trên nói lên điều gì?
4. Thống kê cơ bản cho life_array; mean, median, mode (gồm những giá trị nào? số lần là bao nhiêu?), max, min, variance, std
5. Tìm độ nhọn, độ xiên của dữ liệu. Nhận xét kết quả

Bài toán 2: Xem xét một bộ dữ liệu 40 mẫu khác cho một nhãn hiệu pin khác (Battery 2). Các quan sát được liệt kê trong tập tin life_batteries_2.txt.

1. Tạo life2_array từ nội dung tập tin.
2. So sánh 2 nhóm battery 1 (ở bài toán 1) và battery 2 (ở bài toán 2) theo histogram và nhận xét
3. Biểu đồ trên nói lên điều gì?
4. So sánh hai nhóm mẫu dựa trên thống kê chung, nhận xét kết quả.
5. Vẽ boxplot cho cả 2 nhóm batteries_1 và batteries_2 => nhận xét

Gợi ý

Bài toán 1

In [1]:
```
df <- read.csv("life_batteries.txt", sep = "\t", header = F)
print("Content of file:")
print(df)
```

```
[1] "Content of file:"
    V1  V2  V3  V4
1  130 145 126 146
2  164 130 132 152
3  145 129 133 155
4  140 127 139 137
5  131 126 145 148
6  125 132 126 126
7  126 135 131 129
8  147 136 129 136
9  156 146 130 146
10 132 142 132 132
```

In [2]:
```
data <- c(df$V1, df$V2, df$V3, df$V4)
data
```

```
130 164 145 140 131 125 126 147 156 132 145 130 129 127 126 132
135 136 146 142 126 132 133 139 145 126 131 129 130 132 146 152
155 137 148 126 129 136 146 132
```

In [3]:
```
par(mfrow=c(1,2))
hist(data, main = "Life Batteries", xlab = "Hours",
     xlim = c(110, max(data)+10), ylim = c(0, 20), col = "orange",
     border = "blue", breaks = 5)

# Create the histogram.
hist(data, main = "Life batteries",
     xlab = "Hours",
     xlim = c(110, max(data)+10),
     col = "orange",
     breaks = 10,
     border = "blue",
     # so lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(data))
```

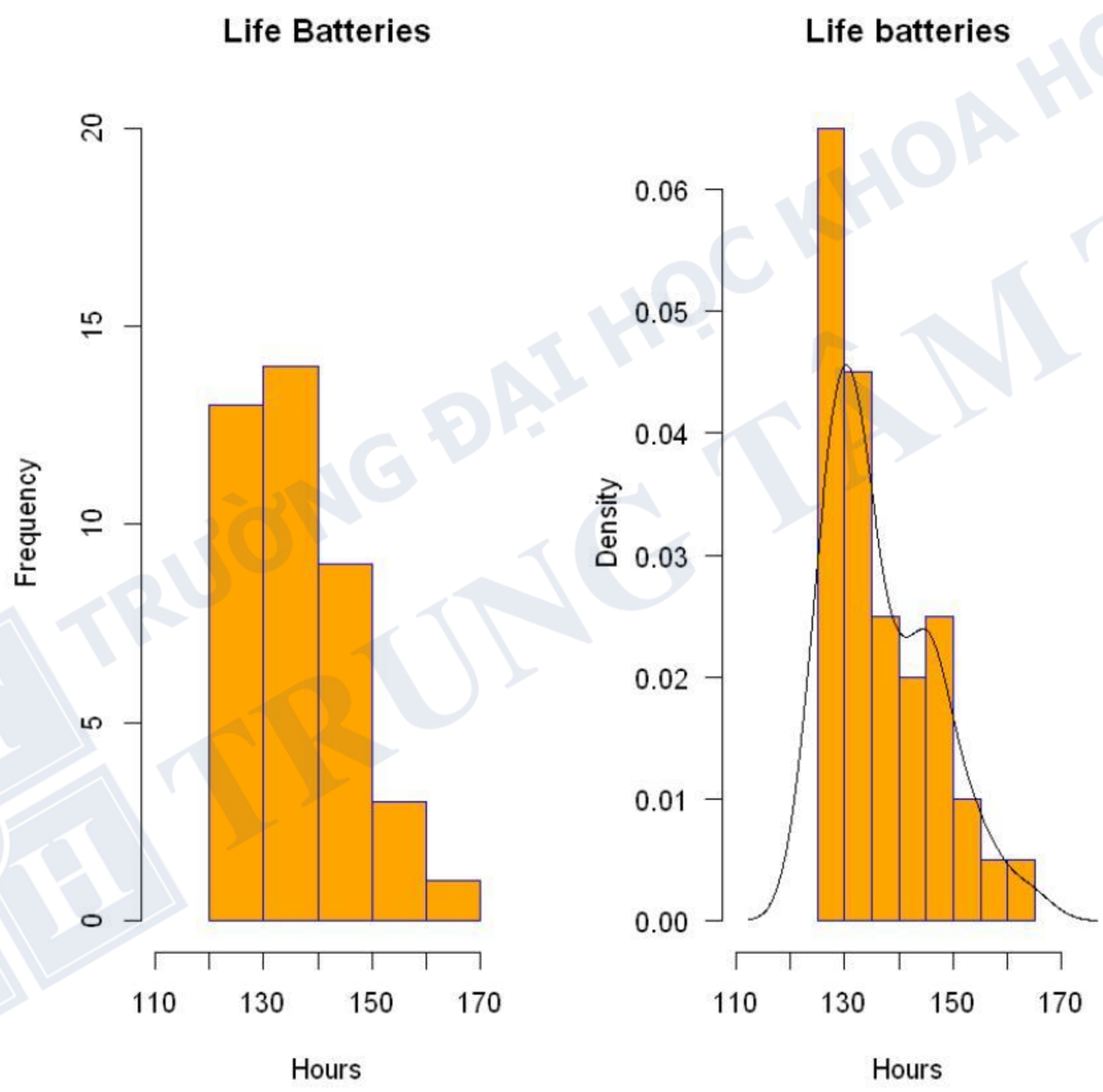


Chart này nói lên điều gì?

- Biểu đồ cho thấy hầu hết các dữ liệu được tập trung trong khoảng 130, với một vài điểm dữ liệu vượt quá 150. Có thể kết luận rằng trung tâm của dữ liệu là một nơi nào đó trong khoảng 130-139.
- Từ hai biểu đồ trên, có thể xác định nhiều biện pháp phân tán và xu hướng trung tâm:

In [4]:
```
summary(data)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  125.0   129.8   132.5   136.8   145.0   164.0 
```

In [5]:
```
var(data)
```

95.874358974359

In [6]:
```
sd(data)
```

9.79154528020777

In [7]:
```
library("moments")
```

In [8]:
```
skewness(data)
```

0.845528704908616

- Phân phối lệch phải

In [9]:
```
kurtosis(data)
```

2.92386761692244

- <3 => phân phối thấp hơn phân phối chuẩn

Bài toán 2

In [10]:
```
df2 <- read.csv("life_batteries_2.txt", sep = "\t", header = F)
print("Content of file:")
print(df2)
```

```
[1] "Content of file:"
    V1  V2  V3  V4
1  134 130 140 151
2  143 134 136 144
3  150 135 160 141
4  143 140 138 141
5  148 146 140 146
6  151 138 151 139
7  151 128 146 147
8  152 142 144 134
9  142 146 142 136
10 122 134 145 147
```

In [11]:
```
data2 <- c(df2$V1, df2$V2, df2$V3, df2$V4)
data2
```

```
134 143 150 143 148 151 151 152 142 122 130 134 135 140 146 138
128 142 146 134 140 136 160 138 140 151 146 144 142 145 151 144
141 141 146 139 147 134 136 147
```

In [12]:
```
par(mfrow=c(1,2))
hist(data2, main = "Life Batteries 2", xlab = "Hours",
     xlim = c(110, max(data2)+10), ylim = c(0, 20), col = "orange",
     border = "blue", breaks = 5)

# Create the histogram.
hist(data2, main = "Life batteries",
     xlab = "Hours",
     xlim = c(110, max(data2)+10),
     col = "orange",
     breaks = 10,
     border = "blue",
     # so lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(data2))
```

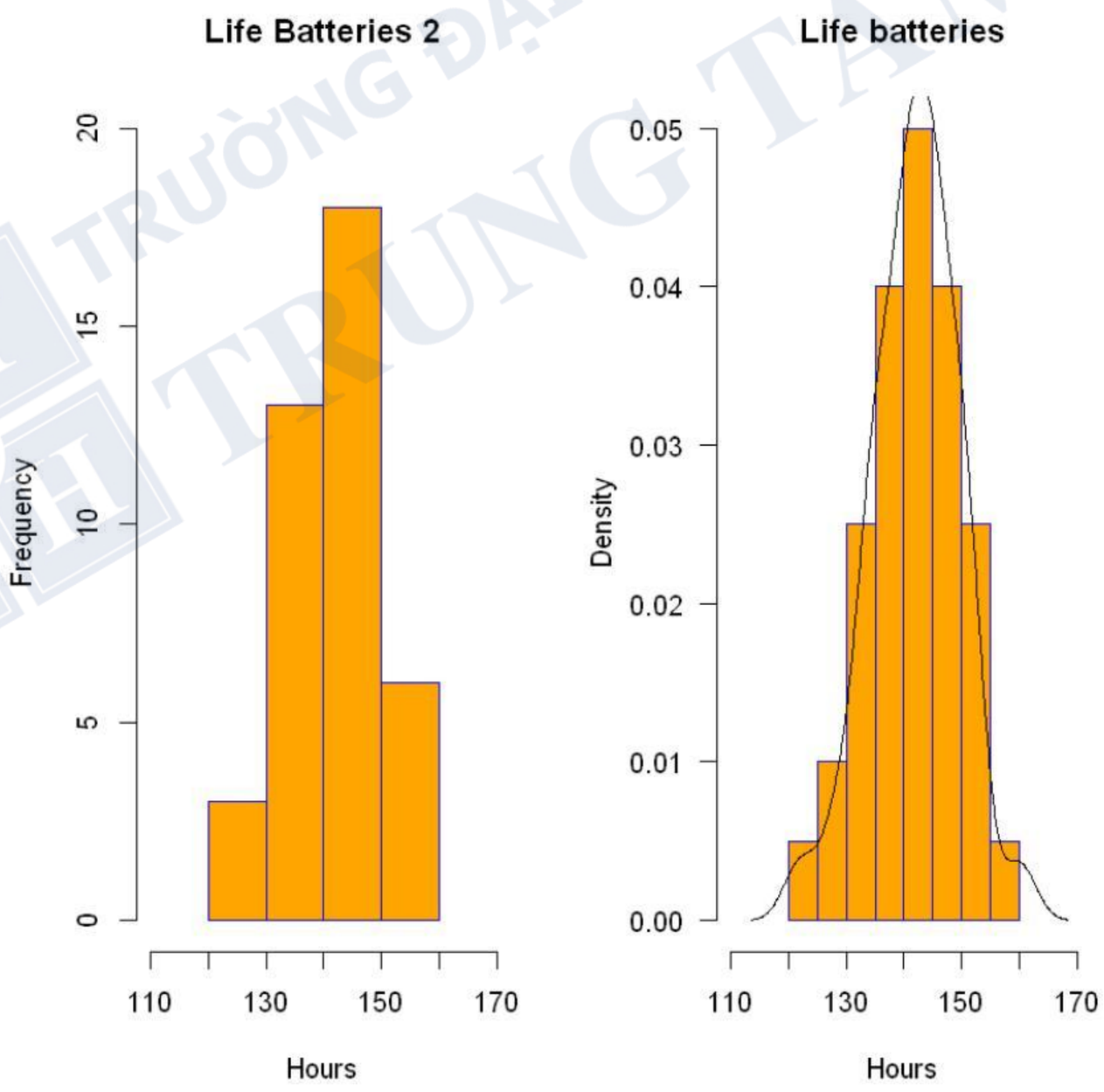


Chart này nói lên điều gì?

- Biểu đồ trên cho thấy rằng có nhiều dữ liệu hơn cho Battery 2 trong khoảng 140 so với Battery 1 trong khoảng 130. Ngoài ra, mức độ biến thiên của Battery 2 ít hơn so với Battery 1.

Dựa trên các kết quả này, có thể kết luận rằng Battery 2 là một nhãn hiệu tốt hơn (trung bình cao hơn và biến thiên thấp hơn). Tuy nhiên, tính hợp lệ của kết luận này còn phụ thuộc vào cách thu thập dữ liệu.

In [13]:
```
summary(data2)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  122.0   137.5   142.0   141.9   146.2   160.0 
```

In [14]:
```
var(data2)
```

55.199358974359

In [15]:
```
sd(data2)
```

7.42962710870195

In [16]:
```
library("moments")
```

In [17]:
```
skewness(data2)
```

-0.251033897416653

- Phân phối lệch trái

In [18]:
```
kurtosis(data2)
```

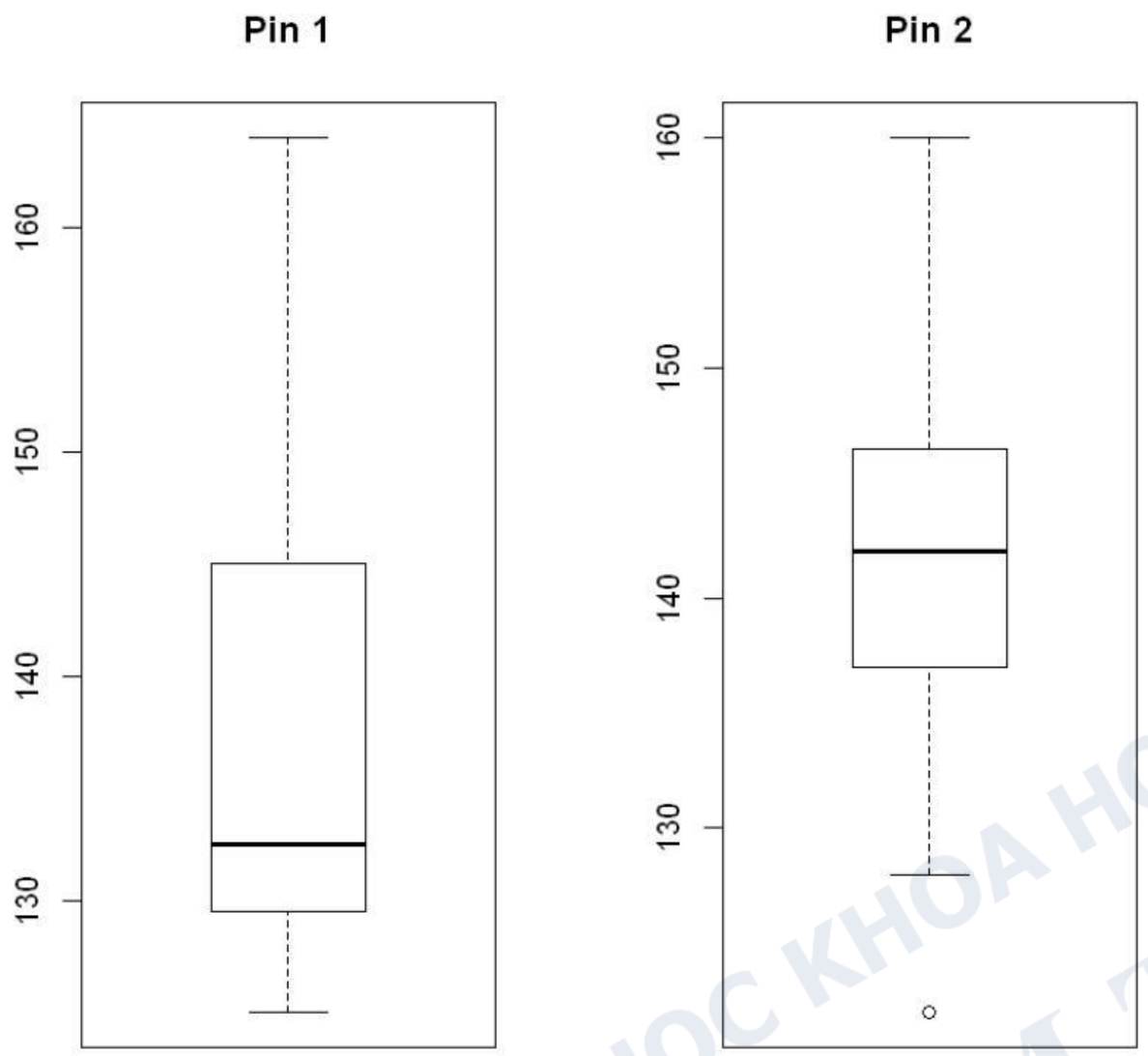
3.35670280005492

- Phân phối cao hơn phân phối chuẩn

Nhận xét:

- Những kết quả này cho thấy rằng Pin 2 có tuổi thọ trung bình cao hơn so với Pin 1 và Pin 2 có phương sai nhỏ hơn.
- Pin 1 có skewness > 0: phân phối lệch phải
- Pin 2 có skewness < 0: phân phối lệch trái
- Pin 1 có kurtosis < 0: phân bố này thấp hơn phân bố chuẩn
- Pin 2 có kurtosis > 0: phân bố này cao hơn phân bố chuẩn

In [19]:
```
par(mfrow=c(1,2))
boxplot(data, main="Pin 1")
boxplot(data2, main="Pin 2")
```



Nhận xét:

- Nhóm Pin 1 không có outliers, nhóm Pin 2 có outliers
- Phân phối Pin 2 tập trung hơn phân phối phin 1
- Biểu đồ ở trên hỗ trợ cho kết luận: phạm vi của Pin 2 ngắn hơn so với Pin 1 (ít thay đổi hơn) và được chuyển sang bên phải (trung tâm cao hơn).