Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh

TRUNG TÂM TIN HỌC

R for Data Science

Bài 13: ***Descriptive Statistics – Thống kê mô tả***

Phòng LT & Mạng

https://csc.edu.vn/lap-trinh-va-csdl/R-Programming-Language-for-Data-Science_190

2019

Giới thiệu

- **Thống kê là một nhánh của toán học liên quan đến việc thu thập, tổ chức, phân tích, giải thích và trình bày dữ liệu. Thống kê liên quan đến mọi khía cạnh của dữ liệu, bao gồm cả việc lập kế hoạch thu thập dữ liệu về mặt thiết kế các khảo sát và thí nghiệm. ~ Wikipedia**
- **Thống kê là một nghệ thuật và khoa học về:**
  - Thu thập
  - Phân tích
  - Trình bày
  - Và giải thích dữ liệu

https://en.wikipedia.org/wiki/Statistics

Giới thiệu

Data Science

❑ Thống kê mô tả (Descriptive Statistics)

- Bao gồm các kỹ thuật được sử dụng để tóm tắt và mô tả dữ liệu số nhằm mục đích giải thích dễ dàng hơn.
  - Ví dụ: Doanh số hàng tháng cho một sản phẩm trong năm qua có thể được mô tả và làm cho ý nghĩa bằng cách chuẩn bị biểu đồ khối hoặc biểu đồ đường.
- **Thống kê mô tả**: Thu thập, tổng kết và mô tả dữ liệu
- **Các phương pháp** được sử dụng để tổng kết dữ liệu: lập bảng, đồ thị, bằng số
- **Thống kê mô tả:** Các tham số thống kê, tần số, phân phối xác suất

Nội dung

Data Science

1. **Centrality measures**
2. **Measures of dispersion**
3. **Percentile, IQR**
4. **Varience, Standard deviation , skewness, kurtosis**
5. **Covariance, correlation**

❑ Centrality measures (Đo lường trung tâm)

- Centrality measure cho chúng ta ước tính về trung tâm phân phối.
- Nó cho chúng ta cảm giác về một giá trị điển hình mà chúng ta mong đợi sẽ thấy.
- Ba phương thức chính của đo lường trung tâm là: mean (trung bình), median (trung vị) và mode.

❑ Mean

- Công thức: $\mu = \frac{1}{n}\sum_i x_i$
- Giá trị trung bình đại diện cho một quan sát điển hình trong bộ dữ liệu. Nếu chúng ta chọn một trong những quan sát một cách ngẫu nhiên, thì chúng ta có thể nhận được một giá trị gần với giá trị trung bình.
- Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
```

```
mean(age)
```

```
17.5714285714286
```

Centrality measures

- Đôi khi giá trị trung bình là một mô tả tốt về một tập hợp các giá trị. Ví dụ, tất cả các quả táo đều có cùng kích thước (ít nhất là những quả được bán trong siêu thị). Vì vậy, nếu mua 6 quả táo và tổng trọng lượng là 900gr, thì đó là một tóm tắt hợp lý để nói rằng chúng chỉ khoảng 150gr.
- Nhưng bí đỏ thì đa dạng hơn. Giả sử bạn trồng nhiều loại bí đỏ trong vườn và một ngày thu hoạch ba quả bí đỏ: 200gr, 600gr và 12kg => giá trị trung bình của mẫu này là 4.27 kg, nhưng nếu nói bí đỏ trung bình trong vườn là 4.27 kg, đó là sai, hoặc ít nhất là sai lệch. Trong ví dụ này, không có trung bình có ý nghĩa vì không có bí đỏ điển hình.
- Giá trị trung bình đại diện cho giá trị điển hình đóng vai trò là thước đo cho tất cả các quan sát.

Centrality measures

❑ Median

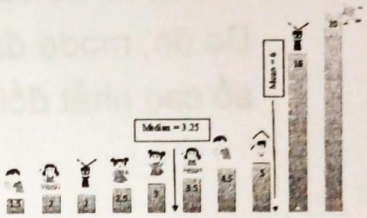
- Giá trị trung vị là giá trị có vị trí ở giữa các item đã được sắp xếp tăng dần
- Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
median(age)
16.5
```

Centrality measures

- Dùng median tốt hơn mean khi:
  - Median là một lựa chọn tốt hơn khi chỉ báo có thể bị ảnh hưởng bởi một số ngoại lệ (outliers).

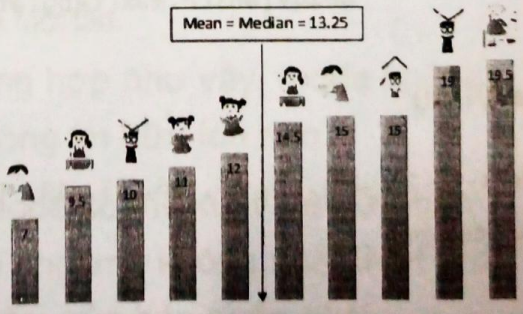
Nhìn vào bức tranh dưới đây, chúng ta có thể thấy rằng Học sinh thứ 9 và thứ 10 ghi được nhiều điểm hơn so với phần còn lại và điểm số của chúng được đưa vào tính toán trung bình, làm cho nó ít đại diện hơn cho quan sát điển hình. Mặt khác, trung vị là giá trị trung bình không bị ảnh hưởng bởi các ngoại lệ này. Nó cho chúng ta một ước tính tốt hơn về điểm số điển hình. Tuy nhiên, điều này có nghĩa là sự hiện diện của Sinh viên thứ 9 và thứ 10 có điểm số tương đối cao hơn.



Mean versus Median in case of outliers

Centrality measures

- Dùng mean tốt hơn median khi:
  - Mean là một lựa chọn tốt hơn khi không có giá trị cực đoan (outlier) nào có thể ảnh hưởng đến nó.



Mean versus Median in case of no extreme values

Centrality measures

❑ Mode

- Mode là giá trị xuất hiện thường xuyên nhất trong bộ dữ liệu. Mode có một lý do rõ ràng: nếu một giá trị xuất hiện liên tục trong toàn bộ dữ liệu => nó sẽ ảnh hưởng đến mức trung bình. Giá trị xuất hiện càng nhiều, nó sẽ càng ảnh hưởng đến giá trị trung bình. Do đó, mode đại diện cho yếu tố đóng góp có trọng số cao nhất đối với giá trị trung bình.

Centrality measures

- Trong R không có hàm mode, do đó cần phải viết function này và gọi sử dụng

```
# Create the function.
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

- Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
getmode(age)
15
```

Centrality measures

- Tại sao dùng mode?
  - Mode có ý nghĩa khi chúng ta không có tập dữ liệu có giá trị số được yêu cầu trong trường hợp mean và median.
    - Ví dụ, nếu một công ty muốn tìm ra nơi phổ biến nhất mà nhân viên của họ cư trú, thì mode sẽ là một thành phố (không thể được thực hiện trong trường hợp mean và median).

Centrality measures

- Tại sao không dùng mode?
  - Mode không hữu ích khi phân phối phẳng; tức là, tần số của tất cả các nhóm là tương tự nhau,
    - Ví dụ, trong bài kiểm tra giữa kỳ cho trường hợp Subject 1, phân phối không thay đổi vì không có số cụ thể nào xuất hiện nhiều hơn một lần.
  - Trong những trường hợp như vậy, mode không cung cấp bất kỳ thông tin hữu ích nào.
  - Ngoài ra, đôi khi mode có thể xuất hiện ở phần đuôi của bản phân phối mà không nhất thiết phải ở hoặc gần trung tâm của bản phân phối.

Subject 1
10
14
11
7
9.5
15
19

Nội dung

1. Centrality measures
2. Measures of dispersion
3. Percentile, IQR
4. Varience, Standard deviation , skewness, kurtosis
5. Covariance, correlation

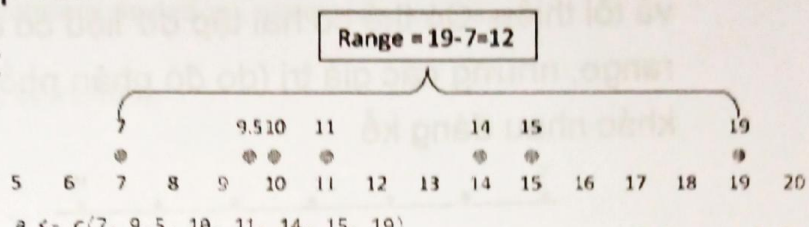
Measures of dispersion

- **Measures of dispersion (Đo lường phân tán)**
  - Là các giá trị mô tả cách dữ liệu thay đổi.
  - Nó cho chúng ta cảm giác về việc dữ liệu có xu hướng phân kỳ khỏi giá trị điển hình, trong khi các đo lường trung tâm cho chúng ta ý tưởng về giá trị điển hình của phân phối.

Measures of dispersion

❑ Range

- Là sự khác biệt giữa các giá trị tối đa (maximum) và tối thiểu (minimum) trong phân phối.
- Ví dụ:




```
a <- c(7, 9.5, 10, 11, 14, 15, 19)

ptp <- range(a)

ptp
7  19
```

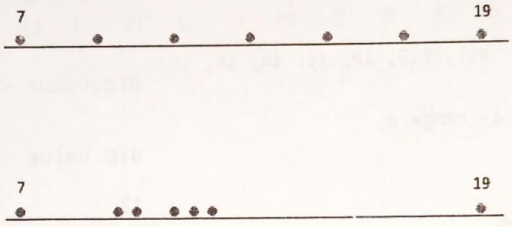
```
ptp.value <- ptp[2] - ptp[1]

ptp.value
12
```


Measures of dispersion

- Tại sao dùng range?
  - Range mang lại cảm giác nhanh chóng về sự lan truyền của phân phối đến những ai chỉ yêu cầu một dấu hiệu sơ bộ của dữ liệu.

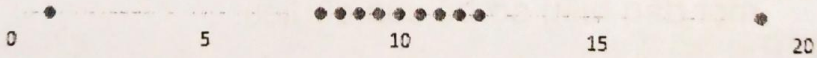
Measures of dispersion

- Tại sao không dùng range?
  - Có một số nhược điểm của việc sử dụng range như một đo lường lây lan. Một là nó không cung cấp bất kỳ thông tin nào của dữ liệu ở giữa tối đa và tối thiểu. Có thể có hai tập dữ liệu có cùng range, nhưng các giá trị (do đó phân phối) có thể khác nhau đáng kể



Measures of dispersion

- Ngoài ra, range rất nhạy cảm với các giá trị outlier vì nó là sự khác biệt giữa các giá trị tối đa và tối thiểu.



- Phân phối ở trên có range = 18, nhưng có thể thấy rõ rằng dữ liệu được phân cụm khoảng 8 đến 12, cho thấy range = 4. Điều này không đưa ra một dấu hiệu tốt về các quan sát trong tập dữ liệu.

Nội dung

1. Contrality measures
2. Measures of dispersion
3. Percentile, IQR
4. Varience, Standard deviation , skewness, kurtosis
5. Covariance, correlation

Percentile, IQR

❑Percentile (phần trăm/phân vị)

- Trả về phân vị thứ k của các giá trị trong phạm vi. Có thể dùng hàm này để thiết lập ngưỡng chấp nhận.
  - Ví dụ: Chẳng hạn, phân vị (percentile) thứ 35 (P35) là điểm số dưới 35% số điểm dữ liệu có thể được tìm thấy. Trung vị đại diện cho phân vị thứ 50. Tương tự, có thể có phần trăm thứ 0 đại diện cho phần trăm tối thiểu và phần trăm thứ 100 đại diện cho mức tối đa của tất cả các điểm dữ liệu.

Percentile, IQR

- Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
quantile(age, c(0.35, 0.5, .98))
```

35%	15.55
50%	16.5
98%	24.74

Percentile, IQR

- Tại sao dùng Percentile?
  - Percentile đưa ra vị trí tương đối của một giá trị cụ thể trong tập dữ liệu. Nếu chúng ta quan tâm đến các vị trí tương đối, thì mean và độ lệch chuẩn (standard deviation) không có ý nghĩa. Phần trăm sẽ phản ánh ý nghĩa cho nhiều thứ. Ví dụ: điểm GRE và GMAT đều tính theo phần trăm.
  - Một đặc tính tốt khác của percentile là nó có một cách giải thích tổng quát; tức là, điều đó không phụ thuộc vào việc chúng ta đang xem điểm thi hay chiều cao của các cầu thủ trong một vài đội bóng rổ. Phân vị thứ 55 luôn có nghĩa là 55% sẽ luôn được tìm thấy giá trị. Nó giúp so sánh các tập dữ liệu có trung bình và độ lệch khác nhau.

Percentile, IQR

Data Science

- **Quartile (Tứ phân vị)**
  - Là cách dùng ba điểm chia bộ dữ liệu thành bốn phần bằng nhau sao cho mỗi nhóm bao gồm một phần tư dữ liệu. Với phần trăm thứ 25 là phần tư thứ nhất (Q1), phần trăm thứ 50 là phần tư thứ hai (Q2) và phần trăm thứ 75 là phần tư thứ ba (Q3).
  - Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
quantile(age)
```

0%	15
25%	15
50%	16.5
75%	18
100%	25

Percentile, IQR

Data Science

- **Interquartile Range(IQR)**
  - Sự khác biệt giữa phần tư thứ ba (Q3) và phần tư thứ nhất (Q1).
  - Công thức: IQR=Q3−Q1
  - Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
quantile(age)
```

0%	15
25%	15
50%	16.5
75%	18
100%	25

```
IQR(age)
```

3

Percentile, IQR

- Tại sao dùng IQR?
  - Interquartile range là một lựa chọn tốt hơn range vì nó không bị ảnh hưởng bởi các ngoại lệ. Nó loại bỏ các ngoại lệ bằng cách chỉ tập trung vào khoảng cách trong khoảng 50% giữa của dữ liệu.

Nội dung

1. Centrality measures
2. Measures of dispersion
3. Percentile, IQR
4. Varience, Standard deviation , skewness, kurtosis
5. Covariance, correlation

❑Variance (Phương sai)

- Phương sai là "trung bình của bình phương khoảng cách của mỗi điểm dữ liệu tới điểm trung bình"
- Với X={a_1,a_2,…,a_n} , và mean là $\bar{x}$

Thì phương sai: $Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
var(age)
```

10.5714285714286

- Tại sao dùng variance?
  - Đây là một biện pháp quan trọng trong thống kê mô tả bởi vì nó cho phép chúng ta đo lường sự lan truyền của một tập dữ liệu xung quanh giá trị trung bình của nó. Các quan sát có thể có hoặc không có ý nghĩa nếu các quan sát trong các tập dữ liệu có tính lan truyền cao.

❑Standard deviation (Độ lệch chuẩn)

- Là căn bậc hai của phương sai

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

  - Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
sd(age)
```

3.25137333621173

- So sánh phương sai và độ lệch chuẩn
  - Điểm giống nhau duy nhất giữa phương sai và độ lệch chuẩn là cả hai đều không âm.
  - Sự khác biệt quan trọng nhất là độ lệch chuẩn nằm trên cùng một tỷ lệ (scale) với các giá trị trong tập dữ liệu. Do đó, nó được thể hiện trong cùng một đơn vị với các giá trị.
  - Trong khi phương sai có tỷ lệ lớn hơn. Vì vậy, nó không được thể hiện trong cùng đơn vị với các giá trị.

❑ Skewness (Độ xiên)

- Độ xiên là một đại lượng đo lường mức độ bất đối xứng của phân phối xác suất của một biến ngẫu nhiên; còn gọi là hệ số bất đối xứng
- Công thức: skewness = $\frac{1}{n}\frac{\sum_{i=1}^{n}(X_i-\bar{X})^3}{s^3}$
  - Với n là số lượng mẫu, s là độ lệch chuẩn, $\bar{X}$ là mean
- install.packages("moments")

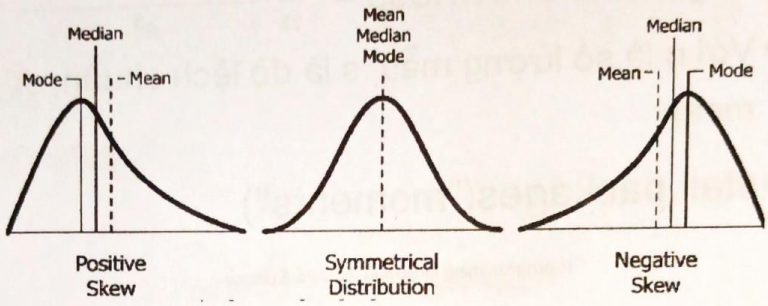
- Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
library("moments")

skewness(age)
```

1.36048829961259

Covariance, correlation

- Nếu hệ số này bằng 0, thì phân phối là đối xứng. Mean = median = mode
- Nếu hệ số này lớn hơn 0, thì phân phối lệch phải. Mode < median, và median < mean
- Nếu hệ số này nhỏ hơn 0, thì phân phối lệch trái. Mean < median, và median < mode



Covariance, correlation

❑ Kurtosis (Độ nhọn)

- Là một chỉ số để đo lường về đặc điểm hình dáng của một phân phối xác suất. Cụ thể hơn, nó so sánh độ cao phần trung tâm của một phân phối so sánh với một phân phối chuẩn.
- Phần trung tâm càng cao và nhọn, chỉ số Kurtosis của phân phối đó càng lớn.
- Công thức: Kurtosis = $\frac{1}{n}\frac{\sum_{i=1}^{n}(X_i-\bar{X})^4}{s^4}$
  - Với n là số lượng mẫu, s là độ lệch chuẩn, $\bar{X}$ là mean

Covariance, correlation

- Ví dụ:

```
age <- c(15, 16, 15, 15, 16, 25, 19, 18, 15, 24, 17, 18, 18, 15)
kurtosis(age)
3.69765431511793
```

  - Với Pearson's measure of kurtosis: nếu giá trị = 3: phân phối bằng normal distribution, <3: thấp hơn, >3: cao hơn

- Tại sao dùng skewness và kurtosis
  - Skewness và kurtosis được sử dụng để mô tả một số khía cạnh của tính đối xứng và hình dạng của sự phân bố các quan sát trong một tập dữ liệu.

Covariance, correlation

❑ Covariance (Hiệp phương sai)

- Là đại lượng phản ánh mức độ tương quan tuyến tính của hai biến số và được tính bằng công thức: $\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})$
- Covariance có thể âm hoặc dương, phản ánh mối quan hệ thuận hay nghịch giữa hai biến số. Giá trị tuyệt đối của nó càng nhỏ, độ mạnh (mức độ chặt chẽ) của mối quan hệ càng thấp. Nếu Covariance = 0 thì *hai biến độc lập.*

Covariance, correlation

- Ví dụ

```
years <- c(7, 28, 23, 18, 19, 15, 24, 13, 2, 8, 20, 21, 18, 7, 2)
salary <- c(26075, 79370, 65726, 41983, 62308, 41154, 53610, 33697,
            22444, 32562, 43076, 56000, 58667, 22210, 20521)

cov(years, salary)
```

135894.071428571

Covariance, correlation

❑ Correlation

$$r = \frac{1}{n-1}\Sigma\left(\frac{X-\overline{X}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right) \qquad S = \sqrt{\frac{\Sigma(X-\overline{X})^2}{N}}$$

where S = the standard deviation of a sample
Σ means "sum of."
X = each value in the data set
$\overline{X}$ = mean of all values in the data set
N = number of values in the data set

- Correlation có giá trị trong khoảng -1 đến 1. Giá trị tuyệt đối của nó càng gần 1 thì độ mạnh (mức độ chặt chẽ) của mối quan hệ càng cao, càng gần 0 thì càng thấp. Nếu Correlation = 0 thì *hai biến độc lập*.

Covariance, correlation

- Ví dụ:

```
years <- c(7, 28, 23, 18, 19, 15, 24, 13, 2, 8, 20, 21, 18, 7, 2)
salary <- c(26075, 79370, 65726, 41983, 62308, 41154, 53610, 33697,
            22444, 32562, 43076, 56000, 58667, 22210, 20521)

cor(years, salary)
```

0.923665165962887

Covariance, correlation

Covariance	Correlation
Một thước đo được sử dụng để chỉ ra mức độ mà hai biến ngẫu nhiên thay đổi song song liên quan đến nhau được gọi là hiệp phương sai.	Một thước đo được sử dụng để biểu thị sức mạnh của hai biến ngẫu nhiên có liên quan được gọi là tương quan.
Giá trị của hiệp phương sai nằm giữa -∞ và +. Vì vậy, phạm vi hiệp phương sai là không xác định. Kết quả khó diễn giải	Giá trị tương quan nằm giữa -1 và +1. Vì vậy, nó là một phạm vi xác định. Kết quả dễ diễn giải.
Hiệp phương sai bị ảnh hưởng bởi sự thay đổi tỷ lệ, tức là nếu tất cả giá trị của một biến được nhân với một hằng số và tất cả giá trị của một biến khác được nhân bởi một hằng số tương tự hoặc khác nhau, thì hiệp phương sai bị thay đổi.	Mối tương quan không bị ảnh hưởng bởi sự thay đổi bởi sự thay đổi tỉ lệ.

Chapter 13: Descriptive Statistics – Thống kê mô tả

Exercise 1: PimaIndiansDiabetes

- Cài package "mlbench"
- Load dataset PimaIndiansDiabetes
- In 10 dòng đầu của dataset
- In thông tin của dataset
- In thống kê chung của dataset
- Cho biết số dòng, cột của dataset
- Với diabetes, hãy cho biết có bao nhiêu mẫu "pos"/ bao nhiêu mẫu "neg". Nhận xét.
- Tính phương sai của tất cả các thuộc tính trong dữ liệu
- Tính standard deviation của tất cả các thuộc tính trong dữ liệu
- Cài package "e1071"
- Tính skewness của tất cả các thuộc tính trong dữ liệu
- Tính kurtosis của tất cả các thuộc tính trong dữ liệu
- Thực hiện correlation matrix để xem xét mối tương quan giữa các thuộc tính trong dữ liệu. Trực quan hóa kết quả.
- Vẽ boxplot của các thuộc tính kiểm tra dữ liệu có outlier hay không?
- Cho biết giá trị ở phân vị thứ 0.05, 0.15, 0.25, 0.50, 0.75, 0.90 của dữ liệu. Biểu diễn phân vị và giá trị tương ứng trên biểu đồ.

```
In [1]: install.packages("mlbench")

        Installing package into 'C:/Users/ktphuong/Documents/R/win-library/3.3'
        (as 'lib' is unspecified)


        package 'mlbench' successfully unpacked and MD5 sums checked

        The downloaded binary packages are in
                C:\Users\ktphuong\AppData\Local\Temp\RtmpSg00cI\downloaded_packages
```

In [2]:
```
# Load the library
library(mlbench)
# Load the dataset
data(PimaIndiansDiabetes)
# display first 10 rows of data
head(PimaIndiansDiabetes, n=10)
```

pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
6	148	72	35	0	33.6	0.627	50	pos
1	85	66	29	0	26.6	0.351	31	neg
8	183	64	0	0	23.3	0.672	32	pos
1	89	66	23	94	28.1	0.167	21	neg
0	137	40	35	168	43.1	2.288	33	pos
5	116	74	0	0	25.6	0.201	30	neg
3	78	50	32	88	31.0	0.248	26	pos
10	115	0	0	0	35.3	0.134	29	neg
2	197	70	45	543	30.5	0.158	53	pos
8	125	96	0	0	0.0	0.232	54	pos

In [3]:
```
str(PimaIndiansDiabetes)
```

```
'data.frame':   768 obs. of  9 variables:
 $ pregnant: num  6 1 8 1 0 5 3 10 2 8 ...
 $ glucose : num  148 85 183 89 137 116 78 115 197 125 ...
 $ pressure: num  72 66 64 66 40 74 50 0 70 96 ...
 $ triceps : num  35 29 0 23 35 0 32 0 45 0 ...
 $ insulin : num  0 0 0 94 168 0 88 0 543 0 ...
 $ mass    : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ pedigree: num  0.627 0.351 0.672 0.167 2.288 ...
 $ age     : num  50 31 32 21 33 30 26 29 53 54 ...
 $ diabetes: Factor w/ 2 levels "neg","pos": 2 1 2 1 2 1 2 1 2 2 ...
```

In [4]: summary(PimaIndiansDiabetes)

```
    pregnant         glucose          pressure         triceps
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    insulin           mass           pedigree           age          diabetes
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   neg:500
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   pos:268
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

In [5]: dim(PimaIndiansDiabetes)

```
768  9
```

In [6]:
```
# distribution of class variable
y <- PimaIndiansDiabetes$diabetes
cbind(freq=table(y), percentage=prop.table(table(y))*100)
```

	freq	percentage
neg	500	65.10417
pos	268	34.89583

In [7]:
```
# calculate variance for all attributes
sapply(PimaIndiansDiabetes[,1:8], var)
```

pregnant	11.3540563206215
glucose	1022.24831425196
pressure	374.647271227184
triceps	254.473245328118
insulin	13281.1800779552
mass	62.1599839573827
pedigree	0.109778637873139
age	138.303045890374

In [8]:
```
# calculate standard deviation for all attributes
sapply(PimaIndiansDiabetes[,1:8], sd)
```

pregnant	3.36957806269887
glucose	31.9726181951362
pressure	19.3558071706448
triceps	15.9522175677276
insulin	115.244002351338
mass	7.88416032037545
pedigree	0.331328595012775
age	11.7602315406787

In [9]:
```
remove.packages("e1071")
```

```
Removing package from 'C:/Users/ktphuong/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
```

In [10]:
```
install.packages("e1071")
```

```
Installing package into 'C:/Users/ktphuong/Documents/R/win-library/3.3'
(as 'lib' is unspecified)


package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ktphuong\AppData\Local\Temp\RtmpSg00cI\downloaded_packages
```

In [11]:
```
# calculate skewness for each variable
library(e1071)
skew <- apply(PimaIndiansDiabetes[,1:8], 2, skewness)
# display skewness
skew
```

pregnant	0.898154872604808
glucose	0.173075366346189
pressure	-1.83641264105356
triceps	0.10894563103235
insulin	2.2633825833595
mass	-0.427307333732463
pedigree	1.91241792381955
age	1.12518804431459

In [12]:
```
kur <- apply(PimaIndiansDiabetes[,1:8], 2, kurtosis)
# display kurtosis
kur
```

pregnant	0.142183957122032
glucose	0.619369451373168
pressure	5.11750995409889
triceps	-0.530936920676741
insulin	7.13313491538188
mass	3.2449626779631
pedigree	5.52853885711983
age	0.621726907802014

In [13]:
```
# calculate a correlation matrix for numeric variables
correlations <- cor(PimaIndiansDiabetes[,1:8])
# display the correlation matrix
correlations
```

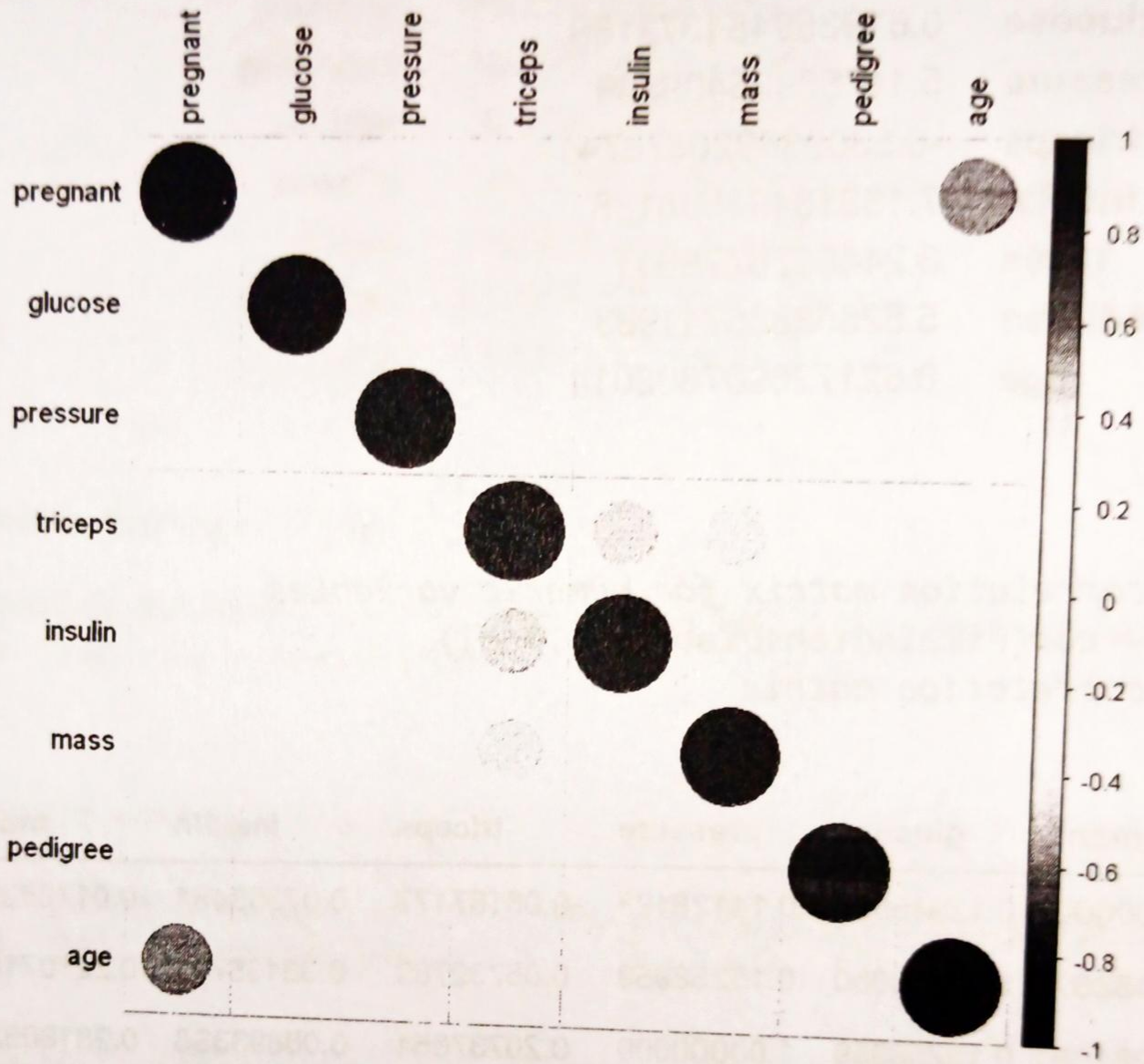
	pregnant	glucose	pressure	triceps	insulin	mass	pedigree
pregnant	1.00000000	0.12945867	0.14128198	-0.08167177	-0.07353461	0.01768309	-0.03352267
glucose	0.12945867	1.00000000	0.15258959	0.05732789	0.33135711	0.22107107	0.13733730
pressure	0.14128198	0.15258959	1.00000000	0.20737054	0.08893338	0.28180529	0.04126495
triceps	-0.08167177	0.05732789	0.20737054	1.00000000	0.43678257	0.39257320	0.18392757
insulin	-0.07353461	0.33135711	0.08893338	0.43678257	1.00000000	0.19785906	0.18507093
mass	0.01768309	0.22107107	0.28180529	0.39257320	0.19785906	1.00000000	0.14064695
pedigree	-0.03352267	0.13733730	0.04126495	0.18392757	0.18507093	0.14064695	1.00000000
age	0.54434123	0.26351432	0.23952795	-0.11397026	-0.04216295	0.03624187	0.03356131

In [14]:
```
install.packages("corrplot")
```

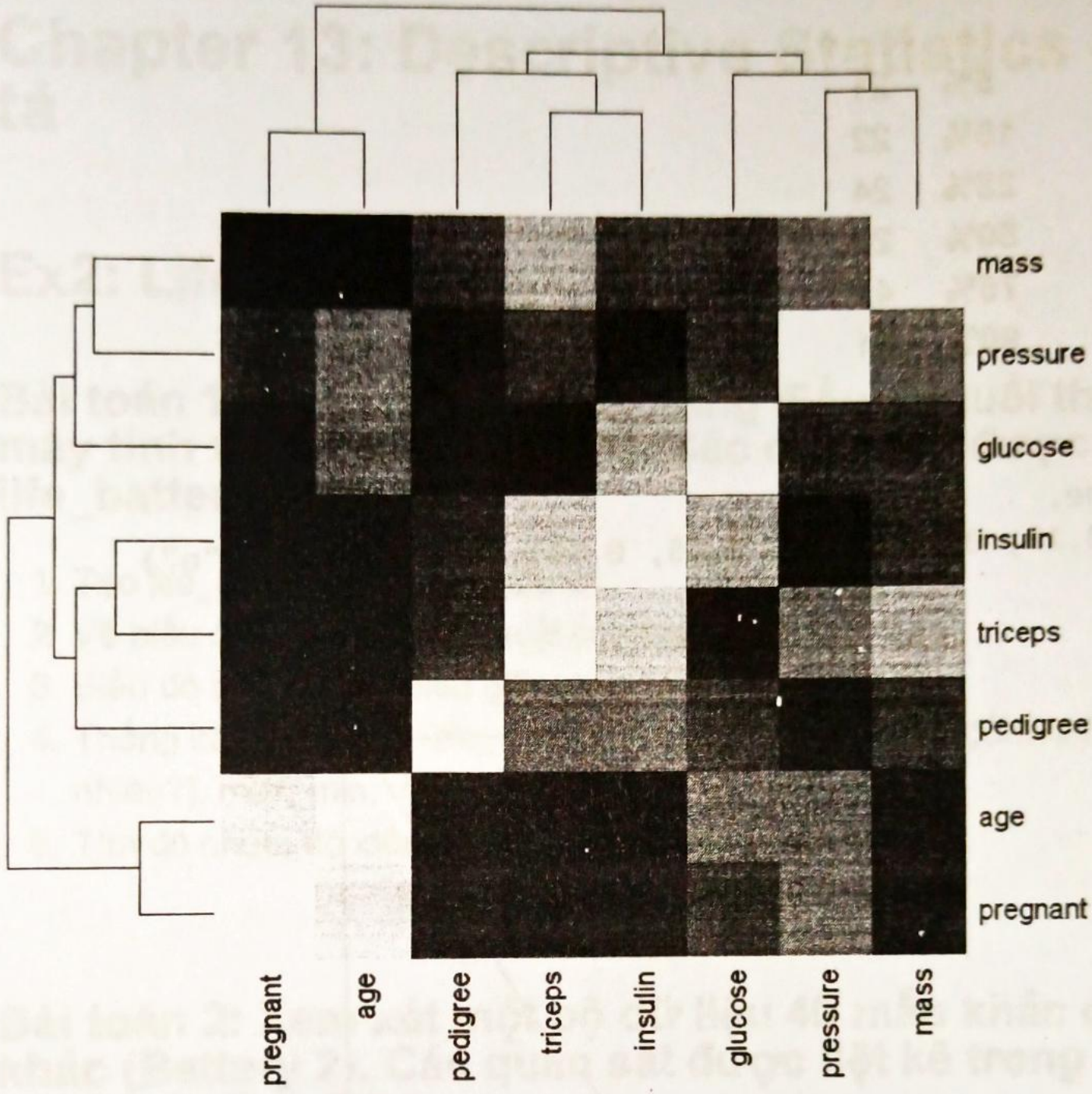
```
Installing package into 'C:/Users/ktphuong/Documents/R/win-library/3.3'
(as 'lib' is unspecified)

package 'corrplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ktphuong\AppData\Local\Temp\RtmpSg00cI\downloaded_packages
```

In [15]:
```
# link: http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-cor
```

```
In [16]: library(corrplot)
         corrplot(correlations)
```



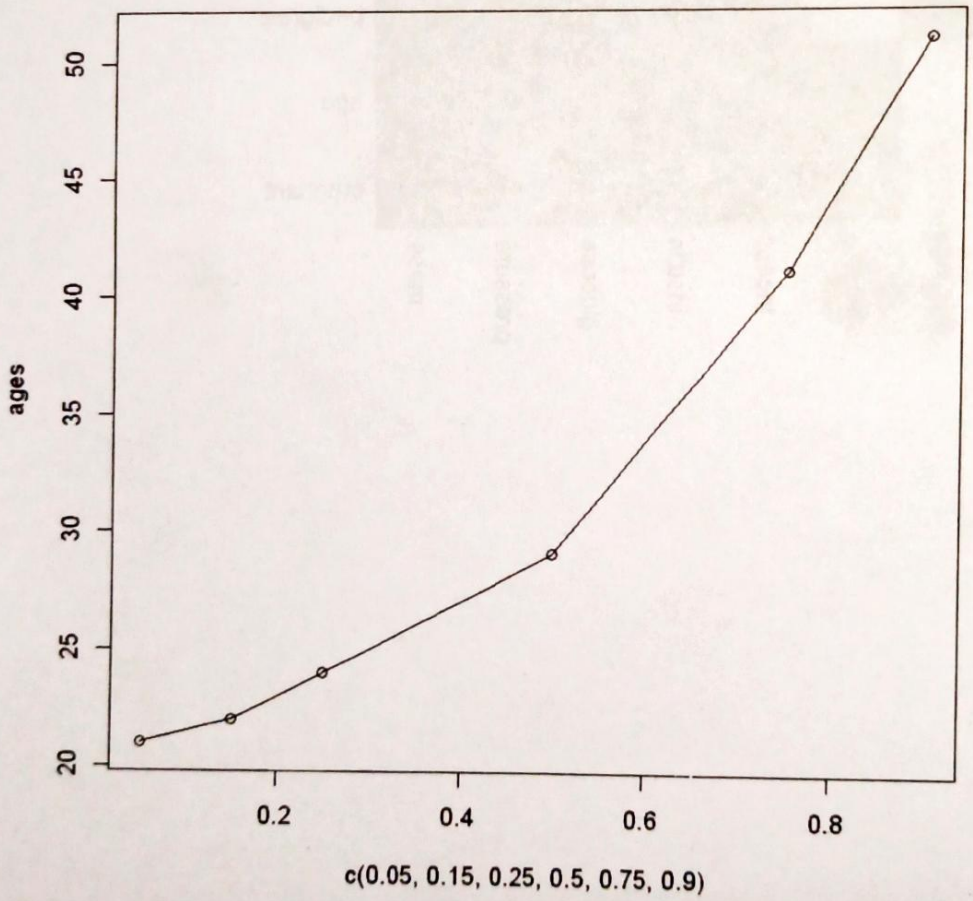
```
In [17]: heatmap(x = correlations)
```



In [18]:
```
ages = quantile(PimaIndiansDiabetes$age, c(0.05, 0.15, 0.25, 0.50, 0.75, 0.90))
ages
```

5%	21
15%	22
25%	24
50%	29
75%	41
90%	51

In [19]:
```
# Plot the line.
plot(c(0.05, 0.15, 0.25, 0.50, 0.75, 0.90), ages, type = "o")
```



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))
```

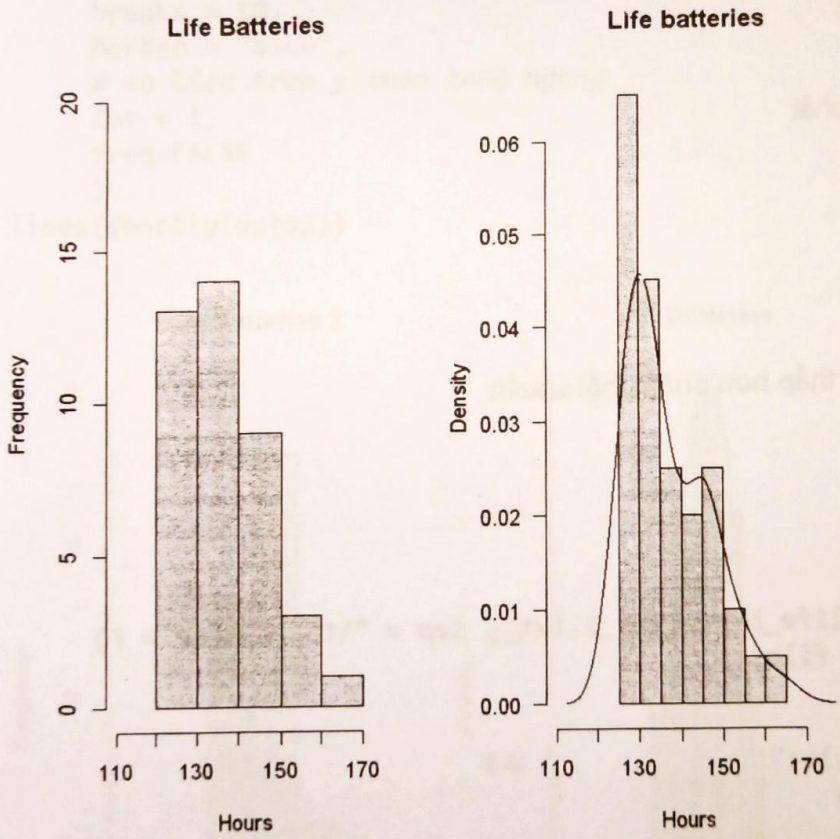


Chart này nói lên điều gì?

- 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))
```

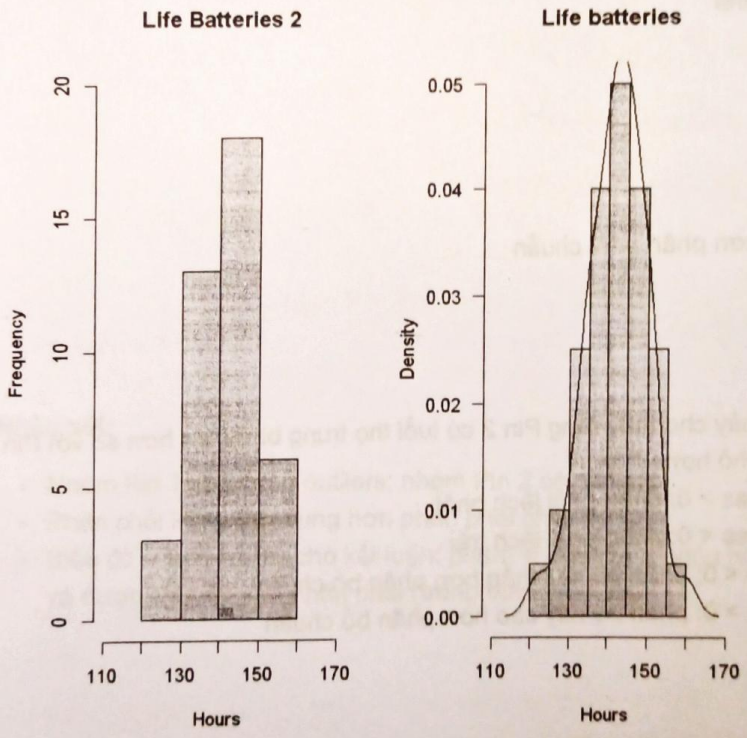


Chart này nói lên điều gì?

- 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)
```

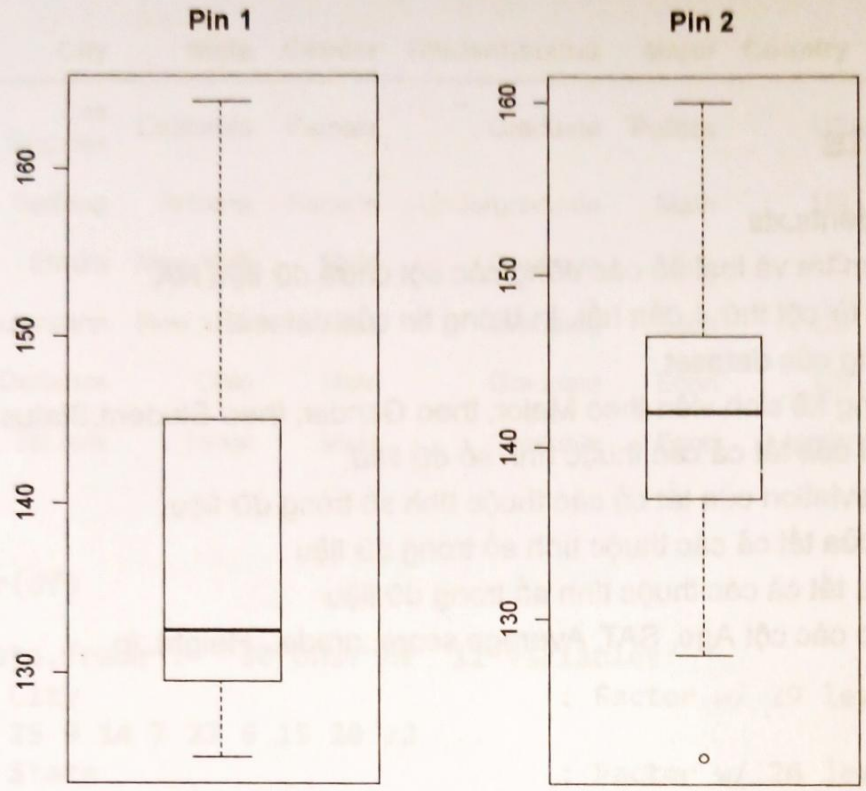
3.35670280005492

- 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).

Chapter 13: Descriptive Statistics – Thống kê mô tả

Ex3: Students

- Cho dữ liệu students.xls
- Đọc dữ liệu, kiểm tra và loại bỏ các dòng/các cột chứa dữ liệu NA
- Tạo dữ liệu mới từ cột thứ 4 đến hết. In thông tin của dataset
- In thống kê chung của dataset
- Vẽ pie chart thống kê sinh viên theo Major, theo Gender, theo Student.Status
- Tính phương sai của tất cả các thuộc tính số dữ liệu
- Tính standard deviation của tất cả các thuộc tính số trong dữ liệu
- Tính skewness của tất cả các thuộc tính số trong dữ liệu
- Tính kurtosis của tất cả các thuộc tính số trong dữ liệu
- vẽ histogram cho các cột Age, SAT, Average.score..grade., Height..in.

Gợi ý

In [1]:
```
# Load the library into R workspace.
library("xlsx")
```

```
Loading required package: rJava

Loading required package: xlsxjars
```

In [2]:
```
df <- read.xlsx("students.xls", sheetName ="Full")
print("Content of file:")
head(df)
```

```
[1] "Content of file:"
```

ID	LastÂ.Name	First.Name	City	State	Gender	Student.Status	Major	Country	Age
1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30
2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19
3	DOE01	JOE01	Elmira	New York	Male	Graduate	Math	US	26
4	DOE02	JOE02	Lackawana	New York	Male	Graduate	Econ	US	33
5	DOE03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37
6	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25

In [3]:
```
df <- df[colSums(!is.na(df)) > 0]
```

```
In [4]: df <- na.omit(df)
```

```
In [5]: df <- df[, 4:14]
        head(df)
```

City	State	Gender	Student.Status	Major	Country	Age	SAT	Average.score..grade.
Los Angeles	California	Female	Graduate	Politics	US	30	2263	67.00000
Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63.00000
Elmira	New York	Male	Graduate	Math	US	26	2221	78.11328
Lackawana	New York	Male	Graduate	Econ	US	33	1716	77.80859
Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65.00000
Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69.00000

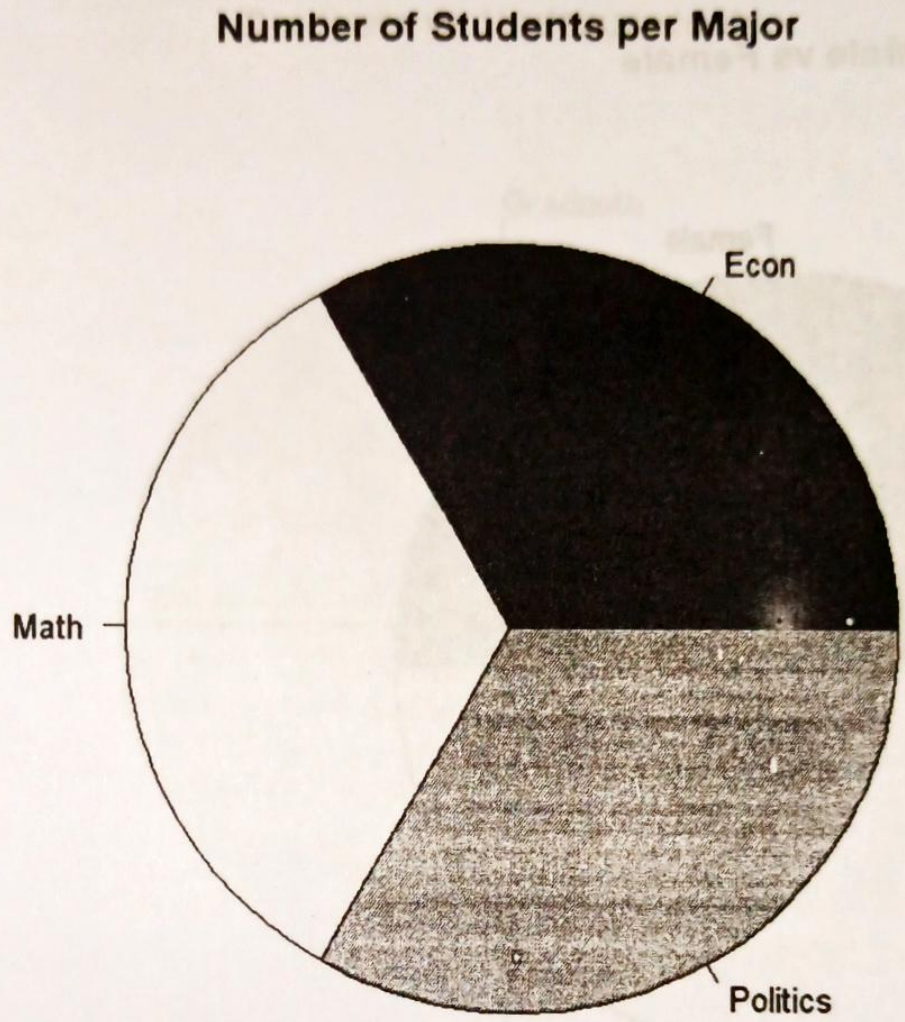
```
In [6]: str(df)

'data.frame':   30 obs. of  11 variables:
 $ City                          : Factor w/ 29 levels "Acme","Amsterdam",..:
17 25 9 14 7 27 6 15 20 22 ...
 $ State                         : Factor w/ 26 levels "Argentina","Arizon
a",..: 4 2 15 15 17 8 16 9 5 15 ...
 $ Gender                        : Factor w/ 2 levels "Female","Male": 1 1 2 2
2 2 2 1 1 1 ...
 $ Student.Status                : Factor w/ 2 levels "Graduate","Undergraduat
e": 1 2 1 1 1 1 1 2 2 1 ...
 $ Major                         : Factor w/ 3 levels "Econ","Math",..: 3 2 2
1 1 1 3 3 2 2 ...
 $ Country                       : Factor w/ 11 levels "Argentina","Bulgari
a",..: 10 10 10 10 10 6 10 10 3 10 ...
 $ Age                           : num  30 19 26 33 37 25 39 21 18 33 ...
 $ SAT                           : num  2263 2006 2221 1716 1701 ...
 $ Average.score..grade.         : num  67 63 78.1 77.8 65 ...
 $ Height..in.                   : num  61 64 73 68 71 67 70 62 62 66 ...
 $ Newspaper.readership..times.wk.: num  5 7 6 3 6 5 5 5 6 5 ...
```

In [7]:
```
summary(df)
```

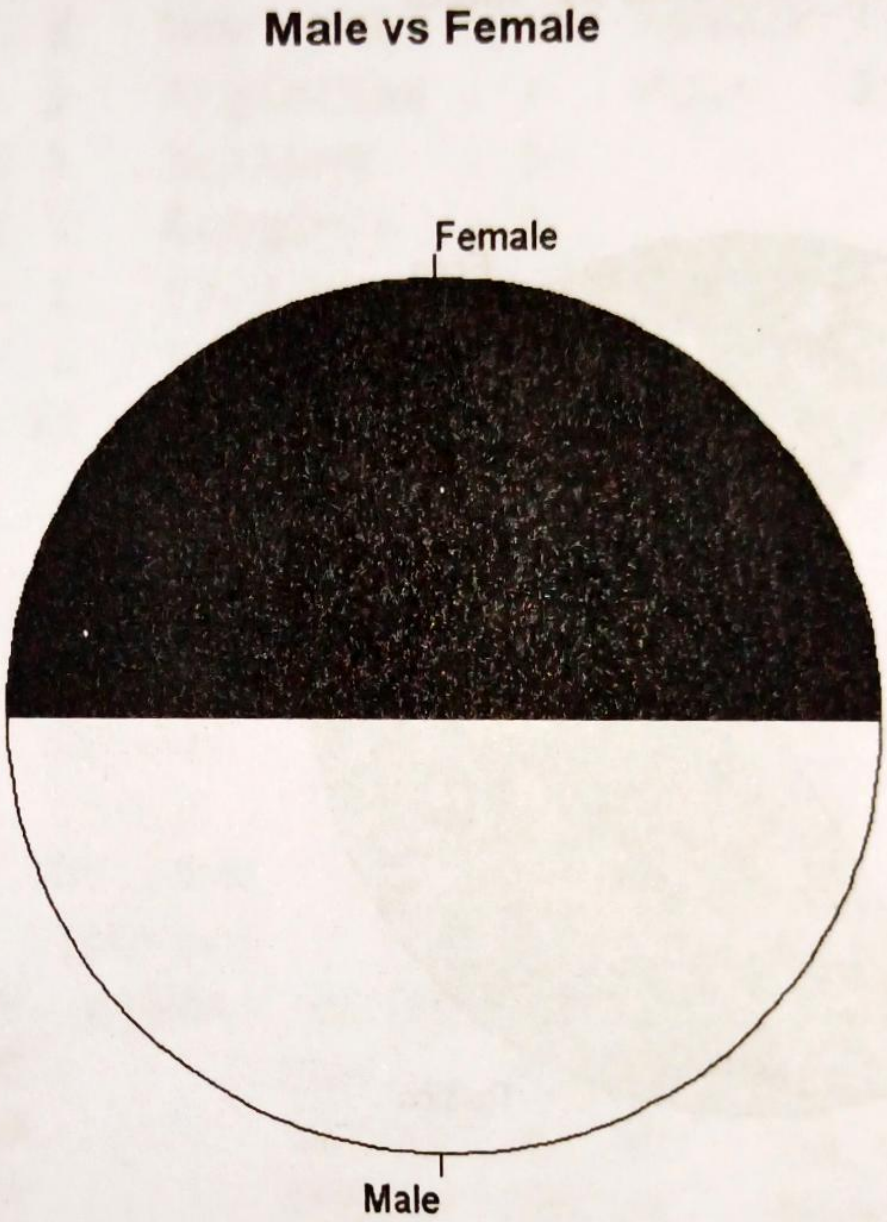
```
                City            State        Gender          Student.Status
 New York     : 2   New York  : 5   Female:15   Graduate      :15
 Acme         : 1   Argentina : 1   Male  :15   Undergraduate:15
 Amsterdam    : 1   Arizona   : 1
 Beijing      : 1   Bulgaria  : 1
 Buenos Aires: 1   California: 1
 Caracas      : 1   Canada    : 1
 (Other)      :23   (Other)   :20
     Major          Country         Age             SAT
 Econ    :10   US        :20   Min.   :18.0   Min.   :1338
 Math    :10   Argentina: 1   1st Qu.:19.0   1st Qu.:1658
 Politics:10   Bulgaria : 1   Median :23.0   Median :1817
               Canada   : 1   Mean   :25.2   Mean   :1849
               China    : 1   3rd Qu.:30.0   3rd Qu.:2032
               Holland  : 1   Max.   :39.0   Max.   :2309
               (Other)  : 5
 Average.score..grade.  Height..in.    Newspaper.readership..times.wk.
 Min.   :63.00         Min.   :59.00   Min.   :3.000
 1st Qu.:72.00         1st Qu.:63.00   1st Qu.:4.000
 Median :79.75         Median :66.50   Median :5.000
 Mean   :80.40         Mean   :66.43   Mean   :4.867
 3rd Qu.:88.00         3rd Qu.:70.75   3rd Qu.:6.000
 Max.   :95.88         Max.   :75.00   Max.   :7.000
```

In [8]:
```
major <- table(df$Major)
colors1 <- c("red", "yellow", "green")

gender <- table(df$Gender)
colors2 <- c("blue", "pink")

status <- table(df$Student.Status)
colors3 <- c("violet", "cyan")
```

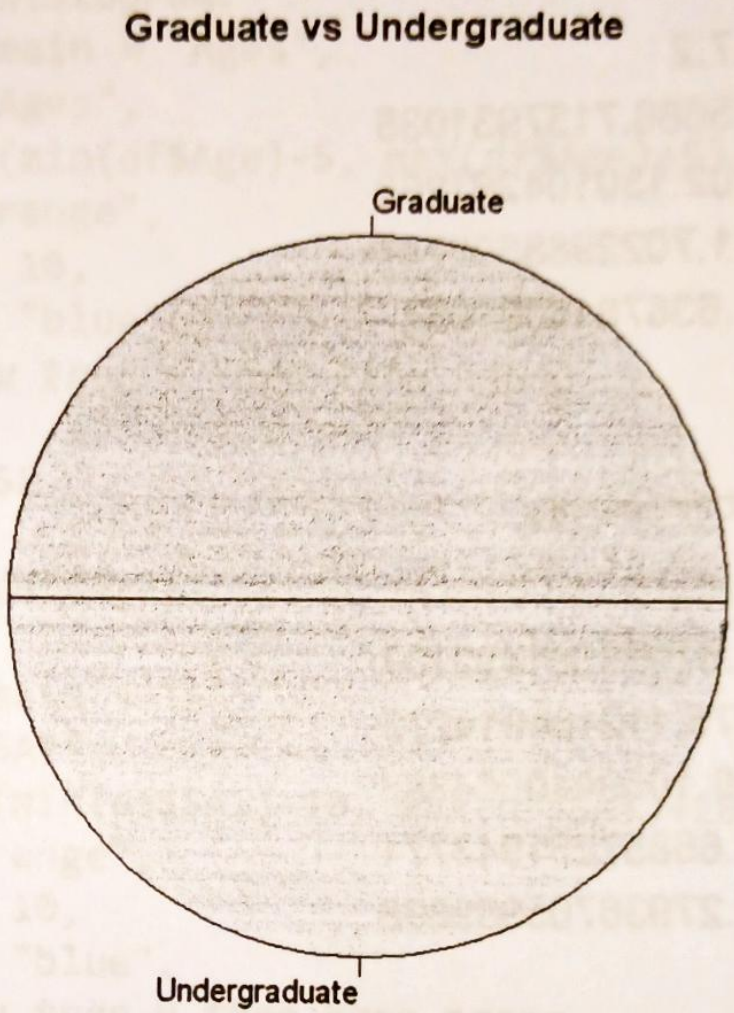
```
In [9]: pie(major, col=colors1, main="Number of Students per Major")
```



```
In [10]: pie(gender, col=colors2, main="Male vs Female")
```



```
In [11]: pie(status, col=colors3, main="Graduate vs Undergraduate")
```



In [12]:
```
# calculate variance for all attributes
sapply(df[7:11], var)
```

Age	47.2
SAT	75686.7137931035
Average.score..grade.	102.130104261266
Height..in.	21.7022988505747
Newspaper.readershi...	1.6367816091954

In [13]:
```
# calculate sd for all attributes
sapply(df[7:11], sd)
```

Age	6.87022561492707
SAT	275.112184014273
Average.score..grade.	10.1059440064383
Height..in.	4.65857261943771
Newspaper.readershi...	1.27936765989898

In [14]:
```
#install.packages("e1071")
```

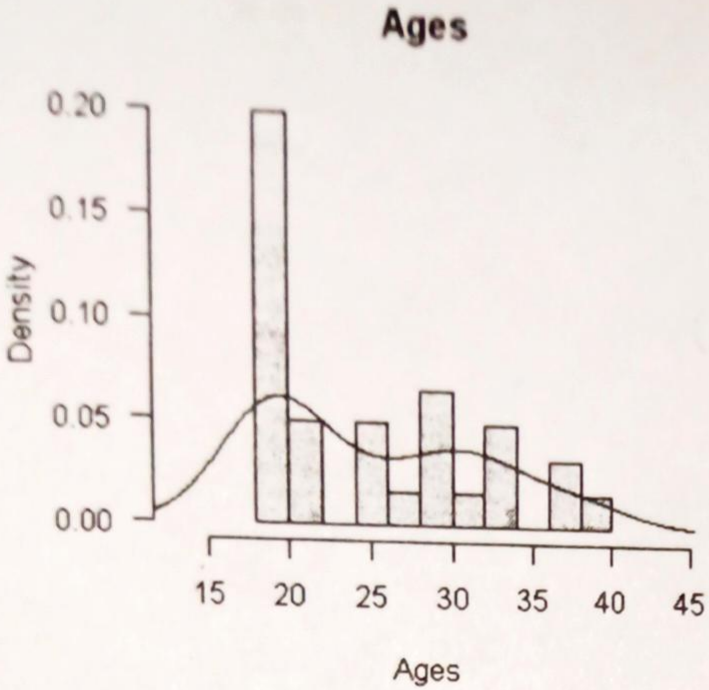
In [15]:
```
# calculate skewness for each variable
library("e1071")
skew <- apply(df[,7:11], 2, skewness)
# display skewness
skew
```

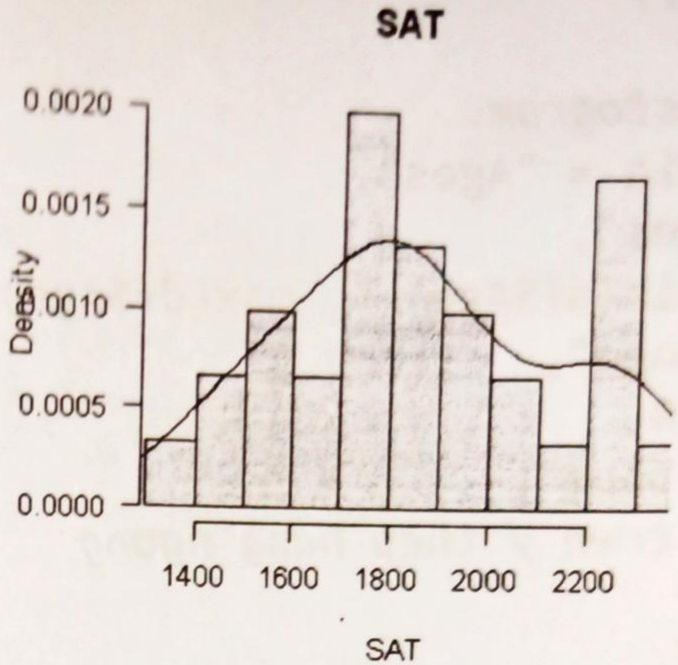
Age	0.502709664447605
SAT	0.14044712797574
Average.score..grade.	-0.101374236291567
Height..in.	0.155085443150385
Newspaper.readershi...	-0.0468347397810847

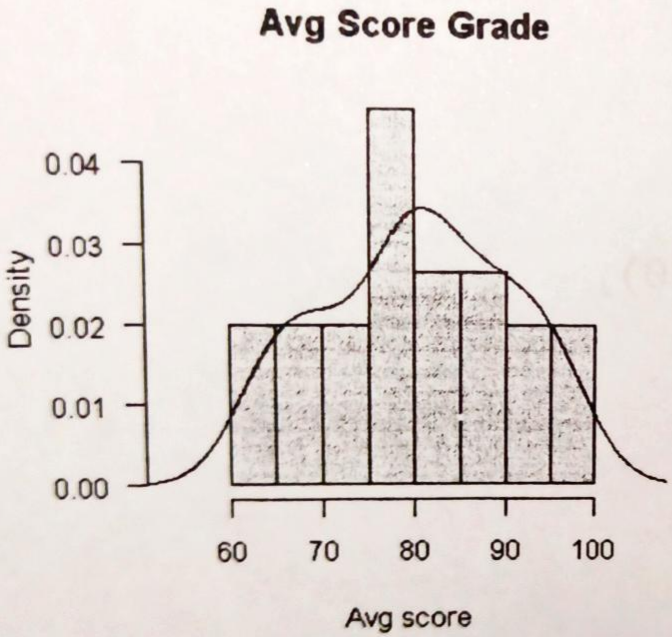
In [16]:
```
kur <- apply(df[,7:11], 2, kurtosis)
kur
```

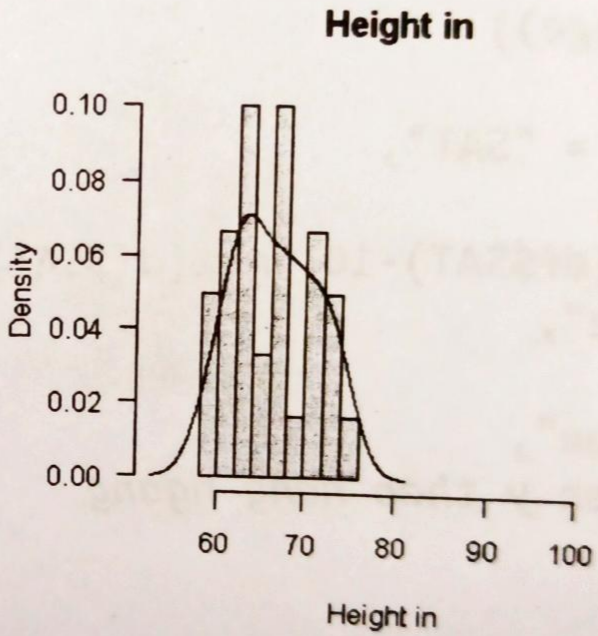
Age	-1.20242842095183
SAT	-1.04281692418247
Average.score..grade.	-1.15697430829857
Height..in.	-1.21584757091849
Newspaper.readershi...	-1.14165472583428

In [17]:
```
par(mfrow=c(2,2))

# Create the histogram.
hist(df$Age, main = "Ages",
     xlab = "Ages",
     xlim = c(min(df$Age)-5, max(df$Age)+5),
     col = "orange",
     breaks = 10,
     border = "blue",
     # so Lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$Age))

hist(df$SAT, main = "SAT",
     xlab = "SAT",
     xlim = c(min(df$SAT)-10, max(df$SAT)+10),
     col = "orange",
     breaks = 10,
     border = "blue",
     # so Lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$SAT))

hist(df$Average.score..grade., main = "Avg Score Grade",
     xlab = "Avg score",
     xlim = c(min(df$Average.score..grade.)-10, max(df$Average.score..grade.)+10
     col = "orange",
     breaks = 10,
     border = "blue",
     # so Lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$Average.score..grade.))

hist(df$Height..in., main = "Height in",
     xlab = "Height in",
     xlim = c(min(df$Average.score..grade.)-10, max(df$Average.score..grade.)+10
     col = "orange",
     breaks = 10,
     border = "blue",
     # so Lieu tren y theo hang ngang
     las = 1,
     freq=FALSE
     )
lines(density(df$Height..in.))
```


Ages
Density
0.20
0.15
0.10
0.05
0.00
15
20
25
30
35
40
45
Ages


SAT
Density
0.0020
0.0015
0.0010
0.0005
0.0000
1400
1600
1800
2000
2200
SAT


Avg Score Grade
Density
0.04
0.03
0.02
0.01
0.00
60
70
80
90
100
Avg score


Height in
Density
0.10
0.08
0.06
0.04
0.02
0.00
60
70
80
90
100
Height in