ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH

TOÁN RỜI RẠC 1

Bài tập lớn đề tài MÃ ĐỀ

Thống kê & phân tích dữ liệu bằng R

GVHD: NNN
Mã nhóm: N
Nhóm SV: Nguyen Van A- 22102134
Tran Van B - 88471334

TP. HỒ CHÍ MINH, THÁNG 1/2015

Mục lục

Nhật ký làm việc nhóm

Nhật ký làm việc nhóm

Bài báo cáo này trình bày về thống kê và phân tích dữ liệu chiều cao của ca sĩ ở New York Choral Society năm 1979, được chia thành 4 cột lần lượt theo giọng nữ cao, nữ trầm, nam cao và nam trầm.

1 Giới thiệu bài toán

Ta cần phải phân tích dữ liệu để cung cấp các thông tin xác thực, trực quan, mô tả cụ thể, dễ hiểu vấn đề đang phân tích để phục vụ nghiên cứu khoa học. Đặc biệt trong các vấn đề kinh tế-xã hội và khi nghiên cứu số lớn chúng ta cần phải quan tâm đến các công cụ kỹ thuật về phân tích số liệu và biểu đồ.

Phân tích số liệu và biểu đồ thường được tiến hành bằng các phần mềm thông dụng như SAS, SPSS, Stata, Statistica, và S-Plus. Đây là những phần mềm được các công ti phần mềm phát triển và giới thiệu trên thị trường khoảng ba thập niên qua, và đã được các trường đại học, các trung tâm nghiên cứu và công ti kĩ nghệ trên toàn thếgiới sửdụng cho giảng dạy và nghiên cứu. Nhưng vì chi phí để sửdụng các phần mềm này tương đối đắt tiền (có khi lên đến hàng trăm ngàn đô-la mỗi năm). Do đó, các nhà nghiên cứu thống kê trên thế giới đã hợp tác với nhau để phát triển một phần mềm mới, với chủ trương mã nguồn mở, sao cho tất cảcác thành viên trong ngành thống kê học và toán học trên thếgiới có thể sử dụng một cách thống nhất và hoàn toàn miễn phí.

Năm 1996, trong một bài báo quan trọng vềtính toán thống kê, hai nhà thống kê học Ross Ihaka và Robert Gentleman [lúc đó] thuộc Trường đại học Auckland, New Zealand phát hoạmột ngôn ngữmới cho phân tích thống kê mà họ đặt tên là R. Nói một cách ngắn gọn, R là một phần mềm sử dụng cho phân tích thống kê và vẽ biểu đồ. Thật ra, về bản chất, R là ngôn ngữ máy tính đa năng, có thể sửdụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí (recreational mathematics), tính toán ma trận (matrix), đến các phân tích thống kê phức tạp. Vì là một ngôn ngữ, cho nên người ta có thể sử dụng R để phát triển thành các phần mềm chuyên môn cho một vấn đề tính toán cá biệt.

Sơ lược về đề tài : Phân tích đề tài chiều cao của các ca sĩ trong các hội hợp xướng New York vào năm 1979. Với giọng hát từ cao nhất đến thấp nhất với thứ tự Soprano, Alto, Tenor, Bass. Trong đó hai cột đầu tiên là giọng nữ còn hai cột sau là giọng nam. Các dữ liệu ban đầu bao gồm hai bộ phận cho từng phần. Bộ dữ liệu này chỉ báo cáo 1 Soprano, 1 Alto, 1 Tenor, 1 Bass. Với số trường hợp ban đầu là 39.

2 Cơ sở lý thuyết

2.1 Thống kê mô tả

Nói đến thống kê mô tả là nói đến việc mô tả dữ liệu bằng các phép tính và chỉ số thống kê thông thường mà chúng ta đã làm quen qua từ thuở trung học như số trung bình (mean), số trung vị (median), số lớn nhất (max), số nhỏ nhất (min), phương sai (variance), độ lệch chuẩn (standard deviation)...

Trong đó ta làm quen các định nghĩa chưa biết :

- Phương sai của một biến ngẫu nhiên là một độ đo sự phân tán thống kê của biến đó, nó hàm ý các giá trị của biến đó thường ở cách giá trị kỳ vọng bao xa.

- Độ lệch chuẩn, hay độ lệch tiêu chuẩn, là một đại lượng thống kê mô tả dùng để đo mức độ phân tán của một tập dữ liệu đã được lập thành bảng tần số. Có thể tính ra độ lệch chuẩn bằng cách lấy căn bậc hai của phương sai.

- số trung vị (tiếng Anh: median) là một số tách giữa nửa lớn hơn và nửa bé hơn của một mẫu, một quần thể, hay một phân bố xác suất. Nó là giá trị giữa trong một phân bố, mà số số nằm trên hay dưới con số đó là bằng nhau. Điều đó có nghĩa rằng 1/2 quần thể sẽ có các giá trị nhỏ hơn hay bằng số trung vị, và một nửa quần thể sẽ có giá trị bằng hoặc lớn hơn số trung vị.

2.2 Công cụ R

Như đã nói ở trên, R là một công cụ miễn phí dùng để phân tích dữ liệu. Chúng ta có thể sử dụng R để thực hiện các phép toán từ đơn giản đến phức tạp. Những bài toán tiêu biểu: các phép kiểm định thống kê, tính toán trên ma trận, hồi quy tuyến tính, gom cụm dữ liệu, bài toán phân lớp... Và vì R là một ngôn ngữ nên chúng ta có thể viết ứng dụng trên R để giải quyết các vấn đề cụ thể.

- Các hàm của R để tính toán thống kê mô tả:

```
> option (width=100)
# chuyển directory
> setwd ("C:/works/stats")

# đọc dữ liệu vào R
> igfdata <- read.table ("igf.txt", hearder = TRUE, na.string = ".")
> attach (igfdata)

# xem xét các cột số trong dữ liệu
> names (igfdata)
hoặc
> igfdata

# tính trung bình
> mean (age)

# phương sai và độ lệch chuẩn
> var (age)
> sd (age)
```

3 Kết quả phân tích dữ liệu

3.1 Tập dữ liệu

- Tập dữ liệu được chia thành 4 cột lần lượt theo giọng nữ cao, nữ trầm, nam cao và nam trầm.
- Đọc dữ liệu bằng R : nhập dữ liệu vào excel và lưu dưới dạng csv (coma delimited).
- Dùng R để nhập dữ liệu dạng csv: giả sử lưu dữ liệu có tên excel.csv trong directory "D:/trr"
- Vào R và gõ lệnh :

```
>setwd(‘‘D:/trr’’) # dẫn R đến thư mục chứa file excel.csv.
>a<-read.csv(‘‘excel.csv’’, header = TRUE) # đọc số liệu bằng R và lưu vào object  có tên là a.
>save (a, file=’’a.rda’’ ) # lưu a dưới dạng R để xử lý.
```

- Sau đó ta kiểm tra lại:

```
> setwd("D:/trr")
> a<- read.csv("excel.csv", header = TRUE)
```

```
> a
```

- Kết quả:

Soprano	Alto	Tenor	Bass
1	64	65	69	72
2	62	62	72	70
3	66	68	71	72
4	65	67	66	69
5	60	67	76	73
6	61	63	74	71
7	65	67	71	72
8	66	66	66	68
9	65	63	68	68
10	63	72	67	71
11	67	62	70	66
12	65	61	65	68
13	62	66	72	71
14	65	64	70	73
15	68	60	68	73
16	65	61	73	70
17	63	66	66	68
18	65	66	68	70
19	62	66	67	75
20	65	62	64	68
21	66	70	NA	71
22	62	65	NA	70
23	65	64	NA	74
24	63	63	NA	70
25	65	65	NA	75
26	66	69	NA	75
27	65	61	NA	69
28	62	66	NA	72
29	65	65	NA	71
30	66	61	NA	70
31	65	63	NA	71
32	61	64	NA	68
33	65	67	NA	70
34	66	66	NA	75
35	65	68	NA	72
36	62	NA	NA	66
37	NA	NA	NA	72
38	NA	NA	NA	70
39	NA	NA	NA	69

```
>a <- na.omit(a)  # loại bỏ những dòng có giá trị NA.
> save(a, file="a.rda")   # lưu a dưới dạng R
> attach(a) # dẫn cho R biết chúng ta muốn xử lí a.
>a
```

	Soprano	Alto	Tenor	Bass
1	64	65	69	72
2	62	62	72	70
3	66	68	71	72
4	65	67	66	69
5	60	67	76	73
6	61	63	74	71
7	65	67	71	72
8	66	66	66	68
9	65	63	68	68
10	63	72	67	71
11	67	62	70	66
12	65	61	65	68
13	62	66	72	71
14	65	64	70	73
15	68	60	68	73
16	65	61	73	70
17	63	66	66	68
18	65	66	68	70
19	62	66	67	75
20	65	62	64	68

3.2 Kết quả phân tích

- Thuộc tính thứ 1 - Soprano:

```
> min(Soprano)
[1] 60
> max(Soprano)
[1] 68
> mean(Soprano)
[1] 64.2
> median(Soprano)
[1] 65
> var(Soprano)
[1] 4.168421
> sd(Soprano)
[1] 2.041671
```

Nhận xét: Qua số liệu được phân tích ở trên ta thấy: chiều cao thấp nhất của đối tượng alto là 60 inch, chiều cao cao nhất là 68 inch,phương sai của Soprano thấp (4.7) cho thấy khoảng cách để đạt đến chiều cao kì vọng gần,ở đây số trung vị cho thấy chiếu cao của đối tượng này nằm chủ yếu ở 65 inch,độ lệch chuẩn cho thấy các đối tượng có chênh lệch chiều cao so với chiều cao trung bình khoảng hơn 2 inch.
- Thuộc tính thứ 2 - Alto:

```
> min(Alto)
[1] 60
> max(Alto)
```

```
[1] 72
> mean(Alto)
[1] 64.7
> median(Alto)
[1] 65.5
> var(Alto)
[1] 8.747368
> sd(Alto)
[1] 2.957595
```

Nhận xét: Qua số liệu được phân tích ở trên ta thấy: chiều cao thấp nhất của đối tượng alto là 60 inch, chiều cao cao nhất là 72 inch,phương sai của Alto khá lớn (8.7) cho thấy khoảng cách để đạt đến chiều cao kì vọng khá xa,ở đây số trung vị cho thấy chiếu cao của đối tượng này nằm chủ yếu ở 65.5 inch,độ lệch chuẩn cho thấy các đối tượng có chênh lệch chiều cao so với chiều cao trung bình khoảng 3 inch.
- Thuộc tính thứ 3 - Tenor:

```
> min(Tenor)
[1] 64
> max(Tenor)
[1] 76
> mean(Tenor)
[1] 69.15
> median(Tenor)
[1] 68.5
> var(Tenor)
[1] 10.34474
> sd(Tenor)
[1] 3.216323
```

Nhận xét: Qua số liệu được phân tích ở trên ta thấy: chiều cao thấp nhất của đối tượng alto là 64 inch, chiều cao cao nhất là 76 inch,phương sai của Tenor lớn (10.3) cho thấy khoảng cách để đạt đến chiều cao kì vọng rất xa,ở đây số trung vị cho thấy chiếu cao của đối tượng này nằm chủ yếu ở 69.15 inch,độ lệch chuẩn cho thấy các đối tượng có chênh lệch chiều cao o với chiều cao trung bình khoảng hơn 3 inch.
- Thuộc tính thứ 4 - Bass:

```
>  min(Bass)
[1] 66
> max(Bass)
[1] 75
> mean(Bass)
[1] 70.4
> median(Bass)
[1] 70.5
> var(Bass)
[1] 5.305263
> sd(Bass)
[1] 2.303316
```

Nhận xét: Qua số liệu được phân tích ở trên ta thấy: chiều cao thấp nhất của đối tượng alto là 66 inch, chiều cao cao nhất là 75 inch,phương sai của Bass mức trung bình (5.3) cho thấy khoảng

cách để đạt đến chiều cao kì vong,ở đây số trung vị cho thấy chiếu cao của đối tượng này nằm chủ yếu ở 70.4 inch,độ lệch chuẩn cho thấy các đối tượng có chênh lệch chiều cao so với chiều cao trung bình khoảng hơn 2 inch.

4 Kết luận

Trong báo cáo này chúng tôi đã trình bày về R với định nghĩa, ứng dụng về R. Sử dụng các hàm của R để thực hiện việc thống kê mô tả tập dữ liệu là phân tích chiều cao của nam và nữ trong dàn hợp xướng New York vào năm 1979. Qua đó đã làm rõ được các thông số về chiều cao min, max, phương sai, độ lệch chuẩn... Và cũng đã chỉ ra được ý nghĩa tầm quan trọng của ngôn ngữ R va ứng dụng của nó để phân tích dữ liệu.

Tài liệu

[1] Giáo sư Nguyễn Văn Tuấn "<**http://www.nguyenvantuan.net/**>", *xem ngày :* 24-29/05/2012.

[2] wikipedia. "**link: http://vi.wikipedia.org/**", *phương sai,độ lệch chuẩn, số trung vị*, lần truy cập cuối: 29/05/2012.