Bài giảng 2:
Đọc dữ liệu vào R

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Đọc dữ liệu vào R

- Tạo dữ liệu trực tiếp
- Đọc text file
- Đọc từ mạng

R có thể đọc dữ liệu dạng nào?

- Trước khi phân tích, chúng ta cần có dữ liệu (data)
- Dữ liệu có thể lưu trữ dưới nhiều dạng
- R có thể đọc bất cứ dữ liệu dạng nào
- Những dạng phổ biến: text, excel, SAS, SPSS, stata, v.v.

Tạo dữ liệu trực tiếp

Tạo dữ liệu trực tiếp

- Dùng lệnh **`c(), data.frame`**
- Ví dụ: chúng ta có dữ liệu 2 cột như sau:

```
50   16.5
62   10.8
60   32.3
40   19.3
48   14.2
47   11.3
57   15.5
70   15.8
48   16.2
67   11.2
```

Tạo dữ liệu trực tiếp

```
age = c(50,62, 60,40,48,47,57,70,48,67)

insulin = c(16.5,10.8,32.3,19.3,14.2,11.3,
   15.5,15.8,16.2,11.2)

tuan = data.frame(age, insulin)

setwd("c:/works/stats")

save(tuan, file="tuan.rda")
```

50	16.5
62	10.8
60	32.3
40	19.3
48	14.2
47	11.3
57	15.5
70	15.8
48	16.2
67	11.2

Đọc ASCII file
`read.table()`

American Standard Code for Information Interchange

Đọc dữ liệu từ text (ASCII file)

- Xác định folder (directory) + tên file
- Dùng lệnh **`read.table()`**

Đọc dữ liệu từ text (ASCII file)

- Folder: document/bai giang online/datasets
- File: bodyfat.txt

```
setwd("/Users/tuannguyen/Documents/Bai giang
Online/Datasets")

dd = read.table("bodyfat.txt", header=T)


dd = read.table("/Users/tuannguyen/Documents/
Bai giang Online/Datasets/bodyfat.txt",
header=T)
```

Đọc từ internet

Đọc dữ liệu từ internet

```
data = read.csv("http://statistics.vn/data/
ExampleData.csv", header=T)
```

```
osteo = read.csv("http://statistics.vn/data/
does_vn07.csv", header=T)
```

Tóm lược: đọc dữ liệu vào R

- R có thể đọc bất cứ dữ liệu dưới dạng nào
- Thông dụng

  text file: **`read.table()`**

  Đọc từ internet

Bổ sung

Dữ liệu dưới dạng tóm lược (summary data)

	Ung thư	Không ung thư
Phơi nhiễm	11	17
Không phơi nhiễm	36	127

Dữ liệu dưới dạng tóm lược (summary data)

	Ung thư	Không ung thư
Phơi nhiễm	11	17
Không phơi nhiễm	36	127

```
cancer = c(1, 0, 1, 0)
exposure = c(1, 1, 0, 0)
nn = c(11, 17, 36, 127)
data = data.frame(exposure, cancer, nn)
```

Hoán chuyển từ character sang numeric

- fat đọc vào như là biến "character"
- chúng ta muốn chuyển sang numeric để dễ tính toán
- dùng lệnh as.numeric

```
fat1 = as.numeric(fat)
```

Đọc từ Excel / csv

`read.csv()`

Dữ liệu từ Excel

- Excel có lẽ là nhu liệu phổ biến để lưu trữ số liệu
- Cấu trúc "flat"
  - cột = columns
  - dòng = rows

Hai cách đọc dữ liệu từ excel vào R

- Đọc trực tiếp (từ .xls)

- Đọc gián tiếp (từ .xls → .csv)

Đọc trực tiếp từ excel

`read.xls()`

Đọc trực tiếp

- Dùng package "**`gdata`**"
- Lệnh **`read.xls`** trong gdata

- Ví dụ: file "FTO gene.xls"

Đọc trực tiếp

```
setwd("/Users/tuannguyen/Documents/Bai
giang Online/Datasets")

library(gdata)

fto = read.xls("FTO gene.xls", header=T)

attach(fto)

fix(fto)
```

Đọc gián tiếp từ csv

Đọc qua read.csv

Bước 1

- Folder: **document/bai giang online/datasets**
- File: **FTO gene.xls**
- Export sang "FTO gene.csv"

Bước 2

```
fto = setwd("/Users/tuannguyen/Documents/Bai giang Online/Datasets/fto gene.csv", header=T)
```

Đọc dữ liệu từ excel vào R

- Xác định folder và tên file
- Lưu trữ excel file dưới dạng .csv
- Dùng lệnh **`read.csv()`**

Đọc dữ liệu từ excel

- Folder: **document/bai giang online/datasets**

```
setwd("/Users/tuannguyen/Documents/Bai giang Online/Datasets")
```

- File: **FTO gene.xls**
- Export to "**FTO gene.csv**"

Đọc từ phần mềm thống kê

Đọc dữ liệu từ SPSS

- Xác định folder và tên file
- Dùng lệnh **`read.spss()`**

Tóm lược: đọc dữ liệu vào R

- R có thể đọc bất cứ dữ liệu dưới dạng nào
- Thông dụng

  excel file: **`read.xls`** (từ gdata)

  excel file: **`read.csv()`**

  SPSS file: **`read.spss()`**

  `STATA : read.stata()`