

Ngôn ngữ lập trình



Thống kê mô tả sử dụng package `dplyr` và `table1`
Xuất kết quả ở định dạng bảng trong Word và Excel

Duc Nguyen

www.tuhocr.com

R Programming Course for Data Science
K2-RPC | T7 8/8/2026 | Session 4

I. Các lệnh thống kê tần số | Frequency table [\[link\]](#)

- `table(b)` # trả về tần số xuất hiện các thành phần của `b`

```
> b <- c(5, 7, 9, 1, 2, 3, 3, 5, 6, 6, 7, 6, 9, NA, 10, 2, NA)
> table(b)
b
 1  2  3  5  6  7  9 10
 1  2  2  2  3  2  2  1
> table(b, useNA = "always")
b
 1  2  3  5  6  7  9 10 <NA>
 1  2  2  2  3  2  2  1  2
```

Lệnh `table()` hai tham số, thống kê tổ hợp, tạo bảng 2x2 contingency

Lệnh `unique()` trả về danh sách các thành phần xuất hiện ít nhất 1 lần trong vector

```
> unique(b)
[1] 5 7 9 1 2 3 6 NA 10
```

II. Các lệnh summary dữ liệu

- ```
aggregate(Age ~ Gender,
 data = df,
 FUN = mean
)
```
- # Function tự định nghĩa
- ```
aggregate(Age ~ Gender,  
          data = df,  
          FUN = function(x) {  
            mean(x, na.rm = TRUE)  
          }  
          )
```
- Các lệnh `apply()`, `sapply()`, `lapply()`

III. Các lệnh trong package `dplyr`

- `dplyr::group_by()`
- `dplyr::summarise()` <https://rpubs.com/hadley/dplyr-programming>
- `dplyr::arrange()` <https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html>
- `dplyr::mutate()`
- `dplyr::case_when()`
- Dấu pipe operator `%>%` và `|>`

Khi dùng các lệnh package `dplyr`, cần xác định xem các lệnh này có thể code bằng các lệnh R cơ bản hay không, nếu được, thì ưu tiên lệnh R cơ bản (không dùng package).

Khi gọi câu lệnh trong package, lưu ý namespace bị trùng, sử dụng `::` để gọi chính xác đến tên function trong package cụ thể.

IV. Package `table1`

- Các lệnh xuất format table bằng package `table1`

<https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html>

- Cách xuất dữ liệu sang Word và Excel

```
library(table1)
library(flextable)
library(magrittr)
library(openxlsx)
```

```
table_freq <- table1::table1.formula( ~ . | Gender,
                                     topclass = "Rtable1-grid",
                                     data = df)
```

```
table1::t1flex(table_freq) %>%
  flextable::save_as_docx(path = "table_freq.docx")
```

```
openxlsx::write.xlsx(table_freq, file = "table_freq.xlsx")
```

IV. Lưu ý khi sử dụng package table1

```
table_freq <- table1:::table1.formula( ~ . ,  
                                         data = df,  
                                         digits = 4)
```

```
> head(df)
```

```
  year
```

```
1 2003
```

```
2 2000
```

```
3 1984
```

```
4 2001
```

```
5 1967
```

```
6 1980
```

```
> dim(df)
```

```
[1] 8444 1
```

```
summary(df)
```

```
  year
```

```
Min.   :1917
```

```
1st Qu.:1955
```

```
Median :1962
```

```
Mean    :1965
```

```
3rd Qu.:1972
```

```
Max.    :2010
```

```
NA's    :751
```

Khi dùng package table1 hoặc các lệnh thống kê mô tả, cần lưu ý đến tham số làm tròn.

`digits = 3`
(mặc định) sẽ tính mean và median không chính xác vì đã được làm tròn.

Overall (N= 8444)	
year	
Mean (SD)	1965 (13.43)
Median [Min, Max]	1962 [1917, 2010]
Missing	751 (8.9%)

Overall (N=8444)	
year	
Mean (SD)	1960 (13.4)
Median [Min, Max]	1960 [1920, 2010]
Missing	751 (8.9%)

V. Lưu ý lệnh làm tròn số trong R

```
> options()$digits
[1] 7
> options()$scipen
[1] 0
```

```
> a <- c(0.2224677112,
+        0.00003,
+        0.00000000424,
+        16002320000,
+        112.2335,
+        1.5,
+        2.5,
+        3.4)
> a
[1] 2.224677e-01 3.000000e-05
[3] 4.240000e-09 1.600232e+10
[5] 1.122335e+02 1.500000e+00
[7] 2.500000e+00 3.400000e+00
```

- `digits` định nghĩa số lượng số sau dấu thập phân hiển thị trên màn hình.
- `scipen` chuyển từ dạng rút gọn (nhân với 10) qua dạng hiển thị toàn bộ số lượng số.
- `?options` xem chi tiết

Việc tùy chỉnh hai thông số này chỉ là “hiển thị” về mặt format, không làm thay đổi bản chất giá trị thực của vector.

```
> options(digits = 20)
> a
[1] 2.2246771120000000121e-01
[2] 3.0000000000000000759e-05
[3] 4.2400000000000002474e-09
[4] 1.6002320000000000000e+10
[5] 1.1223350000000000648e+02
[6] 1.5000000000000000000e+00
[7] 2.5000000000000000000e+00
[8] 3.399999999999999112e+00
> options(digits = 2)
> a
[1] 2.2e-01 3.0e-05 4.2e-09
[4] 1.6e+10 1.1e+02 1.5e+00
[7] 2.5e+00 3.4e+00
```

```
> options(scipen = 999)
> a
[1] 0.22246771120
[2] 0.00003000000
[3] 0.00000000424
[4] 16002320000.00000000000
[5] 112.23350000000
[6] 1.50000000000
[7] 2.50000000000
[8] 3.40000000000
```

V. Lưu ý lệnh làm tròn số trong R

Lệnh `round()` mặc định của R sẽ theo quy ước làm tròn về số chẵn gần nhất. (round to even)

Để thực hiện việc làm tròn lên (≥ 0.5 thì làm tròn lên 1 đơn vị) ta dùng function `roundHalfUp()` trong package `clinUtils`

Các giá trị sau khi được làm tròn này là giá trị thực (được trim, được làm tròn), không phải chỉ là hiển thị format như việc điều chỉnh số lượng số thập phân trong lệnh `options(digits)`.

```
> round(0.5, digits = 0)
[1] 0
> round(1.5, digits = 0)
[1] 2
> round(2.5, digits = 0)
[1] 2
> round(3.5, digits = 0)
[1] 4
>
> library(clinUtils)
> clinUtils::roundHalfUp(0.5, digits = 0)
[1] 1
> clinUtils::roundHalfUp(1.5, digits = 0)
[1] 2
> clinUtils::roundHalfUp(2.5, digits = 0)
[1] 3
> clinUtils::roundHalfUp(3.5, digits = 0)
[1] 4
```

V. Lưu ý lệnh làm tròn số trong package dplyr

```
> options(pillar.sigfig = 2)
> PlantGrowth |> dplyr::group_by(group) |>
+   dplyr::summarise(weight_mean = mean(weight),
+                     weight_sd = sd(weight))
# A tibble: 3 × 3
  group weight_mean weight_sd
<fct>    <dbl>    <dbl>
1 ctrl      5.0      0.58
2 trt1      4.7      0.79
3 trt2      5.5      0.44
```

Ta điều chỉnh số lượng số xuất hiện sau dấu thập phân bằng lệnh `options(pillar.sigfig)`.

Tuy nhiên bản chất đây là con số hiển thị mà thôi, khi chuyển qua data frame thì ta thấy đầy đủ số lượng con số sau dấu thập phân.

```
> options(pillar.sigfig = 2)
> PlantGrowth |> dplyr::group_by(group) |>
+   dplyr::summarise(weight_mean = mean(weight),
+                     weight_sd = sd(weight)) |> as.data.frame()
  group weight_mean weight_sd
1  ctrl      5.032 0.5830914
2  trt1      4.661 0.7936757
3  trt2      5.526 0.4425733
```

V. Các package khác

- Cách xuất bảng kết quả hồi quy qua Word dùng package `jtools`

<https://jtools.jacob-long.com/articles/summ>

- Tính thống kê Odds Ratio dùng package `compareGroups`

https://cran.r-project.org/web/packages/compareGroups/vignettes/compareGroups_vignette.html

- Xuất bảng table dùng package `gtsummary`

<https://www.danielsjoberg.com/gtsummary/>

VI. Homework

Sử dụng dataset `students.csv` để thống kê mô tả toàn bộ các cột dữ liệu sử dụng package `table1`.

Tham khảo: <https://www.tuhocr.com/descriptive-statistics>