

Xử lý và mô tả số liệu nghiên cứu bằng ngôn ngữ lập trình R

TS. Lưu Phúc Lợi,
Trưởng Phòng Nghiên cứu khoa học, Viện Ariha
Bệnh viện Thống Nhất

Jan 11, 2026

GIỚI THIỆU NGÔN NGỮ LẬP TRÌNH R

- Môi trường tính toán cho thống kê, mô hình hóa và trực quan hóa dữ liệu.
- Miễn phí, mã nguồn mở.
- Hỗ trợ phân tích tái lập (R Markdown/Quarto).
- Hệ sinh thái gói lớn: CRAN (đa lĩnh vực) và Bioconductor (omics/bioinformatics).
- Sử dụng rộng rãi trong lĩnh vực học thuật
- Có thể tích hợp vào các phần mềm khác
- Dễ học



<https://www.r-project.org/>

Thống kê & đồ họa

Tái lập nghiên cứu

Gói phong phú

ỨNG DỤNG NGÔN NGỮ LẬP TRÌNH R

1) Dữ liệu & trực quan

- readr/haven: nhập dữ liệu
- dplyr/tidyr: làm sạch & biến đổi
- ggplot2: biểu đồ chất lượng xuất bản

2) Thống kê lâm sàng

- Hồi quy (GLM), mô hình hỗn hợp
- Sống sót (survival), ROC/AUC
- Thiết kế & phân tích thử nghiệm

3) Omics & bioinformatics

- Bioconductor: RNA-seq, microarray, single-cell
- Hạ tầng dữ liệu chuẩn hoá (SummarizedExperiment, ...)
- Vignettes giúp tái lập workflow

4) Báo cáo & triển khai

- R Markdown/Quarto: báo cáo & slide tái lập
- Shiny: dashboard/ứng dụng tương tác
- Chia sẻ: Git + dữ liệu + môi trường

Data Visualization with R

Starting from 13:30 to 15 Wed (19/11/2025 - 25/03/2026)

Location: Auditorium at level 6th

00.0 Introduction to Data Visualizaion [Loi] - 19/11/2025

- [PDF](#)
- [Book](#)

00.1 Practice [Dan] - 19/11/2025

- [PDF](#)
- Installation R & Rstudio
- Basic R
- Introduction to Google Colab
- Practice:  [Open in Colab](#)

01.0 Boxplot and its Variants [Loi] - 10/12/2025

- [PDF](#)

01.1 Practice [Dan] - 10/12/2025

- [PDF](#)
- [Data for Practice 1](#)
- Practice 1:  [Open in Colab](#)
- [Data for Practice 2](#)

```
library(ggplot2)

df <- read.csv("/content/BÀI_TÂP_1_sheet2.csv",
               header = TRUE,
               check.names = FALSE)

# Tính p-value (t-test 2 nhóm)
test_res <- t.test(Weight ~ SEX, data = df)
p_val <- test_res$p.value

b <- ggplot(df, aes(x = SEX, y = Weight)) +
  geom_boxplot(staplewidth = 1, fill = "violet", color = "black") +
  geom_jitter(
    width = 0,
    size = 6,
    alpha = 1,
    color = "brown"
  ) +
  annotate("text",
           x = 1.5,
           y = max(df$Weight) * 1,
           label = paste0("p = ", signif(p_val, 3)),
           size = 8) +
  theme(
    axis.text.y = element_text(size = 20),
    axis.text.x = element_text(size = 20)
  )

b
```

CÁC THÀNH PHẦN CHÍNH TRONG NGÔN NGỮ R

- **Base R**: tập các hàm và cấu trúc dữ liệu mặc định sau khi cài đặt.
- R vận hành qua gõ lệnh (trong R, lệnh = hàm = function)
- **Packages**: mở rộng tính năng (vd: tidyverse, ggplot2, survival, Bioconductor).
- R vận hành dựa trên 'đối tượng': **object** (bao gồm dữ liệu, biến số, hình ảnh, bảng, v.v.)
- Dữ liệu (dataset) trong R gọi là **data.frame**
- Mỗi data frame có nhiều biến số gọi là **variable**
- **Function** phải có **arguments** / đối số (đầu vào)

$$R = \text{Base} + \text{Packages}$$

Ví dụ: cài và gọi package để phân tích dữ liệu

```
install.packages("ggplot2")
library(ggplot2)
```

Dữ liệu nghiên cứu dạng excel (csv format)

Heart Failure Prediction Dataset

	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
2	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
3	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
4	37	M	ATA	130	283	0	ST	98	N	0	Up	0
5	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
6	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
7	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
8	45	F	ATA	130	237	0	Normal	170	N	0	Up	0
9	54	M	ATA	110	208	0	Normal	142	N	0	Up	0
10	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
11	48	F	ATA	120	284	0	Normal	120	N	0	Up	0
12	37	F	NAP	130	211	0	Normal	142	N	0	Up	0
13	58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
14	39	M	ATA	120	204	0	Normal	145	N	0	Up	0
15	49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
16	42	F	NAP	115	211	0	ST	137	N	0	Up	0
17	54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0

Link project:

<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data>

Làm sao đọc vào R?

Đọc dữ liệu vào R

- **Bước 1:** tìm đường link
- **Bước 2:** dùng lệnh/hàm `read.csv` và đặt tên cho dữ liệu (ví dụ: `df`)
- **Bước 3:** kiểm tra qua hàm `head(df)`

```
df <- read.csv("/content/Heart_Data.csv")
# hoặc
library(lessR)
df <- Read("/content/Heart_Data.csv")
head (df)
```

A data.frame: 6 × 12

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
	<int>	<chr>	<chr>	<int>	<int>	<int>	<chr>	<int>	<chr>	<dbl>	<chr>	<int>
1	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0

A data.frame: 918 × 12													
Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease		
<int>	<chr>	<chr>	<int>	<int>	<int>	<chr>	<int>	<chr>	<dbl>	<chr>	<int>		
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0		
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1		
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0		
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1		
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0		
39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0		
45	F	ATA	130	237	0	Normal	170	N	0.0	Up	0		
54	M	ATA	110	208	0	Normal	142	N	0.0	Up	0		

Data frame (dataset) trong trường hợp này là **df**

Biến số (variables): **cột của df**: Age, Sex, ChestPainType, Cholesterol, ...

df này có bao nhiêu variables?

có bao nhiêu samples?

Phân tích mô tả số liệu nghiên cứu

package table 1

Mô tả đặc điểm variables

Overall (N=918)

```
# Tải và gọi package table1
install.packages("table1")
library(table1)
# Thống kê
df <- read.csv("/content/Heart_Data.csv")
table1(~Age + Sex + RestingBP + Cholesterol + FastingBS + HeartDisease,
      data=df)
```

Chỉ trả lời câu hỏi:

“Mẫu nghiên cứu của chúng ta trông như thế nào?”

Age	
Mean (SD)	53.5 (9.43)
Median [Min, Max]	54.0 [28.0, 77.0]
Sex	
F	193 (21.0%)
M	725 (79.0%)
RestingBP	
Mean (SD)	132 (18.5)
Median [Min, Max]	130 [0, 200]
Cholesterol	
Mean (SD)	199 (109)
Median [Min, Max]	223 [0, 603]
FastingBS	
Mean (SD)	0.233 (0.423)
Median [Min, Max]	0 [0, 1.00]
HeartDisease	
Mean (SD)	0.553 (0.497)
Median [Min, Max]	1.00 [0, 1.00]

Mô tả đặc điểm variables tương quan với HeartDisease

```
df <- read.csv("content/Heart_Data.csv")
table1(~Age + Sex + RestingBP + Cholesterol + FastingBS |  
HeartDisease, data=df)
```

Kết luận:

Nhóm mắc bệnh tim có tuổi trung bình ...

Nhóm mắc bệnh tim có [...] cao hơn

Nhóm mắc bệnh tim có [...] cao hơn

	0 (N=410)	1 (N=508)	Overall (N=918)
Age			
Mean (SD)	50.6 (9.44)	55.9 (8.73)	53.5 (9.43)
Median [Min, Max]	51.0 [28.0, 76.0]	57.0 [31.0, 77.0]	54.0 [28.0, 77.0]
Sex			
F	143 (34.9%)	50 (9.8%)	193 (21.0%)
M	267 (65.1%)	458 (90.2%)	725 (79.0%)
RestingBP			
Mean (SD)	130 (16.5)	134 (19.8)	132 (18.5)
Median [Min, Max]	130 [80.0, 190]	132 [0, 200]	130 [0, 200]
Cholesterol			
Mean (SD)	227 (74.6)	176 (126)	199 (109)
Median [Min, Max]	227 [0, 564]	217 [0, 603]	223 [0, 603]
FastingBS			
Mean (SD)	0.107 (0.310)	0.335 (0.472)	0.233 (0.423)
Median [Min, Max]	0 [0, 1.00]	0 [0, 1.00]	0 [0, 1.00]

Xử lý số liệu nghiên cứu (data editing)

Mục đích

Dữ liệu thô lúc nào cũng cần 'biên tập'

- Biến **HeartFailure** có giá trị (0/1) nên cho label: no/yes
- Biến **FastingBS** có giá trị (0/1) nên cho label: no/yes
- Một số biến cần có label
- Tạo ra biến mới và tính toán
- Phân nhóm và coding

```
# HeartFailure chuyển thành factor và gán nhãn 0 = No, 1 = Yes  
df$HeartFailure = factor(df$HeartFailure,  
                          levels = c(0, 1),  
                          labels = c("No", "Yes"))
```

```
# FastingBS chuyển thành factor và gán nhãn 0 = No, 1 = Yes  
df$FastingBS = factor(df$FastingBS,  
                       levels = c(0, 1),  
                       labels = c("No", "Yes"))
```

Prompt ChatGPT

Prompt: Tôi có dataset y sinh với ~900 quan sát.

Các biến gồm: Age (years), Sex (M/F), RestingBP (mmHg), Cholesterol (mm/dl, có giá trị 0), FastingBS (0/1), HeartDisease (0/1).

Tôi muốn tạo Table 1 phân tầng theo HeartDisease.

Hãy cho tôi code R để biên tập dữ liệu cho Table 1 (factor, label, trình bày) trong 1 script

```

# 1. Load packages
library(dplyr)
library(tableone)

# 2. Import data
# Thay đường dẫn và tên file cho phù hợp
# data <- read.csv("your_data.csv")

# Giả sử data đã có sẵn trong môi trường
df <- data

# 3. Data cleaning & recoding
df <- df %>%
  mutate(
    # Outcome
    HeartDisease = factor(
      HeartDisease,
      levels = c(0, 1),
      labels = c("No", "Yes")
    ),
    # Sex
    Sex = factor(
      Sex,
      levels = c("M", "F"),
      labels = c("Male", "Female")
    ),
    # Fasting blood sugar
    FastingBS = factor(
      FastingBS,
      levels = c(0, 1),
      labels = c("≤ 120 mg/dl", "> 120 mg/dl")
    ),
    # Cholesterol: 0 được coi là missing (thường gặp trong dataset tim mạch)
    Cholesterol = na_if(Cholesterol, 0)
  )

```

```

# 4. Variable labels (dùng cho Table 1)
var_labels <- list(
  Age      = "Age (years)",
  Sex      = "Sex",
  RestingBP = "Resting blood pressure (mmHg)",
  Cholesterol = "Serum cholesterol (mg/dl)",
  FastingBS = "Fasting blood sugar",
  HeartDisease = "Heart disease" )

# 5. Variables to include in Table 1
vars <- c(
  "Age",
  "Sex",
  "RestingBP",
  "Cholesterol",
  "FastingBS" )

# 6. Identify categorical variables
cat_vars <- c(
  "Sex",
  "FastingBS" )

# 7. Create Table 1
table1 <- CreateTableOne(
  vars      = vars,
  strata    = "HeartDisease",
  data      = df,
  factorVars = cat_vars,
  addOverall = TRUE )

# 8. Print Table 1 (publication-ready)
print(
  table1,
  label     = TRUE,
  varLabels = TRUE,
  quote     = FALSE,
  noSpaces  = TRUE,
  missing   = TRUE,
  contDigits = 1,
  catDigits = 1,
  pDigits   = 3,
  test      = TRUE )

```



ChatGPT

	Overall n	No 918	Yes 508	p
Age (mean (SD))	53.5 (9.4)	50.6 (9.4)	55.9 (8.7)	<0.001
Sex = Female (%)	193 (21.0)	143 (34.9)	50 (9.8)	<0.001
RestingBP (mean (SD))	132.4 (18.5)	130.2 (16.5)	134.2 (19.8)	0.001
Cholesterol (mean (SD))	244.6 (59.2)	238.8 (55.4)	251.1 (62.5)	0.005
FastingBS = > 120 mg/dl (%)	214 (23.3)	44 (10.7)	170 (33.5)	<0.001
Stratified by HeartDisease				
test Missing				
Age (mean (SD))	0.0			
Sex = Female (%)	0.0			
RestingBP (mean (SD))	0.0			
Cholesterol (mean (SD))	18.7			
FastingBS = > 120 mg/dl (%)	0.0			

	0 (N=410)	1 (N=508)	Overall (N=918)
Age			
Mean (SD)	50.6 (9.44)	55.9 (8.73)	53.5 (9.43)
Median	51.0	57.0	54.0
[Min, Max]	[28.0, 76.0]	[31.0, 77.0]	[28.0, 77.0]
Sex			
F	143 (34.9%)	50 (9.8%)	193 (21.0%)
M	267 (65.1%)	458 (90.2%)	725 (79.0%)
RestingBP			
Mean (SD)	130 (16.5)	134 (19.8)	132 (18.5)
Median	130	132	130
[Min, Max]	[80.0, 190]	[0, 200]	[0, 200]
Cholesterol			
Mean (SD)	227 (74.6)	176 (126)	199 (109)
Median	227	217	223
[Min, Max]	[0, 564]	[0, 603]	[0, 603]
FastingBS			
Mean (SD)	0.107 (0.310)	0.335 (0.472)	0.233 (0.423)
Median	0 [0, 1.00]	0 [0, 1.00]	0 [0, 1.00]
[Min, Max]			

Take home message

- Ngôn ngữ R có thể ứng dụng cho bất cứ mô hình phân tích thống kê nào
- Ưu điểm: miễn phí, nhưng rất 'mạnh' và linh động
- Điểm yếu: cần thời gian để làm quen

Thực hành

- **Vào Github:** https://github.com/luuloi/Scientific_Research_Introduction_Course
- Vào Google Colab bài 5: **Xử lý và mô tả số liệu nghiên cứu bằng R**
- Tạo bản sao colab về drive của bạn
- Tải **Heart_Data.csv**
- Đọc file vào R
- Lặp lại những thao tác R

CÁM ƠN SỰ LẮNG NGHE
CỦA QUÝ ANH/CHỊ