1_Intro

April 16, 2024

1 Statistical Computing

1.1 Phạm vi nghiên cứu

Computational statistics and statistical computing là 2 lĩnh vực riêng trong thống kê:
- **Statistical computing** truyền thống là dựa trên **phương pháp số và thuật toán** như là **tối ưu, sinh số ngẫu nhiên**.
- **Computational statistics** là khám phá dữ liệu, phương pháp Monte Carlo, phân chia dữ liệu (train, valid, test).

→ 2 lĩnh vực có nhiều điểm chung và project này sẽ theo học kiến thức từ cả 2 lĩnh vực trên.

1.2 Monte Carlo

Phương pháp Monte Carlo dựa trên 1 tập đa dạng các phương pháp trong **thống kê kết luận** (statistical inference - dựa trên data mẫu đưa ra kết luận hay dự báo về tổng thể mà data mẫu đại diện) cũng như **phân tích số** khi mà việc mô phỏng được sử dụng: Các dạng **tích phân Monte Carlo**

1.3 Resampling method

Khi thực hiện **bootstrap**, **data mẫu được sinh ra dựa trên phân phối xác suất đã cho trước** để thực hiện việc tính toán xác suất. Như vậy cần việc thu thập thông tin thống kê về phân phối của data mẫu (**độ lệnh bias, sai số chuẩn SE**).

Resampling method (như the **ordinary bootstrap and jackknife**) là phương pháp **không tham số** được sử dụng cho trường hợp: **không tồn tại trực tiếp phân phối của biến ngẫu nhiên or phương pháp mô phỏng**.

→ Phương pháp sinh mẫu giả (chương 3 & 4) sẽ hỗ trợ cho phương pháp Monte Carlo (chương 6 đến 11)

1.4 Markov Chain Monte Carlo (MCMC)

Là phương pháp dựa trên 1 thuật toán có **mục đích sinh data mẫu từ 1 phân phối xác suất cụ thể - chính là phân phối dừng của chuỗi Markov**.

Về 'phân phối dừng của chuỗi Markov':
Một "stationary distribution" của một Markov chain là một phân phối xác suất mà nó đạt được sau một số lượng đủ lớn các bước chuyển trạng thái → tức là phân phối xác suất của các trạng thái khi chuỗi Markov đạt đến trạng thái ổn định.

Về 'Markov chain':
Một chuỗi các sự kiện ngẫu nhiên trong đó xác suất chuyển đến một trạng thái mới chỉ phụ thuộc vào trạng thái hiện tại, không phụ thuộc vào lịch sử của chuỗi.
Trong một Markov chain, mỗi trạng thái có một xác suất chuyển đến các trạng thái khác.

Nếu một Markov chain có "stationary distribution", điều này có nghĩa là phân phối xác suất của các trạng thái không thay đổi theo thời gian và được duy trì ổn định. Trong nhiều ứng dụng, việc xác định "stationary distribution" của một Markov chain là quan trọng để hiểu hành vi của hệ thống và các tính chất xác suất của nó sau một thời gian đủ lớn.

Các phương pháp MCMC được dùng nhiều trong phân tích Bayesian, cũng như vật lý tính toán, tài chính tính toán. **Markov Chain Monte Carlo methods** được đề cập ở chương 11.

1.5 Một số nội dung khác

Ước lượng mật độ (Density estimation) (Chapter 12) - không dùng tham số, dùng để khám phá dữ liệu cho mục đích **phân cụm dữ liệu**.
Phương pháp tính là cần thiết để **trực quan hóa dữ liệu đa chiều** cũng như tiến hành **giảm số chiều**.
Sự gia tăng khối lượng dữ liệu, dữ liệu stream và dữ liệu nhiều chiều được ứng dụng trong y sinh, kỹ thuật.
Chapter 5 giới thiệu phương pháp trực quan hóa dữ liệu đa chiều.
Các chủ đề lựa chọn trong phươn pháp số như root finding and numerical integration ở Chapter 13.
Việc tối ưu hóa với R ở nội dung Chapter 14.

Các chủ đề về lập trình như benchmarking, efficiency and code profiling ở Chapter 15.

There are now many excellent online resources available, in addition to the online R and RStudio documentation, such as galleries of code and graphics, online books, tutorials and blogs.
See the references in the individual chapters for some of these.
The R-bloggers website is worth visiting; it currently combines blog posts from some 750 bloggers at https://www.r-bloggers.com/.

2 The R Environment

R là 1 hệ thống tính toán thống kê và đồ họa.
Bao gồm 1 ngôn ngữ và 1 môi trường chạy (đồ họa, debug, dùng hàm hệ thống, chạy chương trình theo kịch bản).
R dựa trên ngôn ngữ S.

The home page of the R project is http://www.r-project.org/, and the current R distribution and documentation are available on the Comprehensive R Archive Network (CRAN) at http://cran.R-project.org/.

Phần còn lại sau đây sẽ bao gồm kiến thức cơ bản để bắt đầu với R.
- Lựa chọn môi trường phát triển tích hợp (IDE) (hoặc cũng đâu cần dùng IDE)
- Cú pháp cơ bản
- Dùng sự giúp đỡ online
- Data, files, scripts, and packages.
- Vectors, matrices, lists and data frames, an overview of basic graphics functions.

3 Getting Started with R and RStudio

...
Cái này không phải trọng tâm hiện tại (chủ yếu bao gồm các thành phần cửa số của RStudio)

4 Basic Syntax

4.1 Phép toán tử gán

Dùng `<-` hoặc `=`.
Nên dùng `<-` vì nó có thể dùng ở bất kỳ đâu trong khi `=` chỉ cho phép ở mức top level.

4.2 Một số phép toán tử hay dùng

[ ]:
```
# Comment: #
# #this is a comment

# Assignment: <-
x <- log2(2)

# Concatenation operator: c
c(3, 2, 2)

# Elementwise multiplication (nhân từng phần tử của mảng này với phần tử tương␣
 ↪ứng của mảng kia): *
# a <- c(1,2,3)
# b <- c(2,3,4)
a * b # <- c(2,6,12)

# Exponentiation: ^
2^1.5

# x mod y: x %% y
25 %% 3

# Integer division: %/%
25 %/% 3

# Sequence from a to b by h: seq
seq(a, b, h)

# Sequence operator: :
0:20
```

4.3 Một số hàm hay dùng

- Square root: sqrt

- Lấy phần nguyên dưới, trên: floor, ceiling
- Natural logarithm (loga cơ số e): log
- Exponential function ex: exp
- Factorial: factorial
- Random Uniform numbers (phân phối đều): runif
- Random Normal numbers (phân phối chuẩn): rnorm
- Normal distribution (phân phối đều): pnorm, dnorm, qnorm
- Rank, sort: rank, sort
- Variance, covariance: var, cov
- Std. dev., correlation: sd, cor
- Frequency tables: table
- Missing values: NA, is.na

4.4 Một số hàm và toán tử thường dùng với ma trận

[ ]:
```r
# Tạo 1 vector 0
x <- numeric(n)
x <- integer(n)
x <- rep(0,n)

# Zero matrix
matrix(0, n, m)
x <- matrix(0, n, m)
# ith element of vector a
a[i]
a[i] <- 0
# j th column of a matrix A
A[,j]
sum(A[,j])
# ij th entry of matrix A
A[i,j]
x <- A[i, j]

# Matrix multiplication (Nhân ma trận)
%*%
a %*% b
```

```r
# Elementwise multiplication *
a * b

# Matrix transpose
t
t(A)

# Matrix inverse (tìm ma trận khả nghịch)
solve
solve(A)

# Diagonal
diag
diag(A)
```

5 Using the R Online Help System

Dùng ? hoặc ¿‘tên toán tử” hoặc help(...)

6 Distributions and Statistical Tests

Phân phối và kiểm tra thống kê
Chủ yếu dùng tới `stats` package.

[ ]:
```r
help.search("distribution", package="stats")
```

```
R Information

Help files with alias or concept or title matching ‘distribution’ using
fuzzy matching:


stats::Beta             The Beta Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Binomial         The Binomial Distribution
  Concepts: Probability Distributions and Random Numbers
stats::bw.nrd0          Bandwidth Selectors for Kernel Density
                        Estimation
  Concepts: Probability Distributions and Random Numbers
stats::Cauchy           The Cauchy Distribution
  Concepts: Probability Distributions and Random Numbers
stats::chisq.test       Pearson's Chi-squared Test for Count Data
  Concepts: Probability Distributions and Random Numbers
stats::Chisquare        The (non-central) Chi-Squared Distribution
  Concepts: Probability Distributions and Random Numbers
stats::density          Kernel Density Estimation
```

```
  Concepts: Probability Distributions and Random Numbers
stats::distribution     Distributions in the stats package
  Aliases: distribution, distributions, Distributions
  Concepts: Probability Distributions and Random Numbers
stats::ecdf             Empirical Cumulative Distribution Function
stats::Exponential      The Exponential Distribution
  Concepts: Probability Distributions and Random Numbers
stats::FDist            The F Distribution
  Concepts: Probability Distributions and Random Numbers
stats::fivenum          Tukey Five-Number Summaries
  Concepts: Probability Distributions and Random Numbers
stats::GammaDist        The Gamma Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Geometric        The Geometric Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Hypergeometric   The Hypergeometric Distribution
  Concepts: Probability Distributions and Random Numbers
stats::IQR              The Interquartile Range
  Concepts: Probability Distributions and Random Numbers
stats::Logistic         The Logistic Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Lognormal        The Log Normal Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Multinomial      The Multinomial Distribution
  Concepts: Probability Distributions and Random Numbers
stats::NegBinomial      The Negative Binomial Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Normal           The Normal Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Poisson          The Poisson Distribution
  Concepts: Probability Distributions and Random Numbers
stats::ppoints          Ordinates for Probability Plotting
  Concepts: Probability Distributions and Random Numbers
stats::qbirthday        Probability of coincidences
  Concepts: Probability Distributions and Random Numbers
stats::qqnorm           Quantile-Quantile Plots
  Concepts: Probability Distributions and Random Numbers
stats::r2dtable         Random 2-way Tables with Given Marginals
  Concepts: Probability Distributions and Random Numbers
stats::SignRank         Distribution of the Wilcoxon Signed Rank
                        Statistic
  Concepts: Probability Distributions and Random Numbers
stats::Smirnov          Distribution of the Smirnov Statistic
stats::TDist            The Student t Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Tukey            The Studentized Range Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Uniform          The Uniform Distribution
```

```
  Concepts: Probability Distributions and Random Numbers
stats::Weibull          The Weibull Distribution
  Concepts: Probability Distributions and Random Numbers
stats::Wilcoxon         Distribution of the Wilcoxon Rank Sum Statistic
  Concepts: Probability Distributions and Random Numbers


Type '?PKG::FOO' to inspect entries 'PKG::FOO', or 'TYPE?PKG::FOO' for
entries like 'PKG::FOO-TYPE'.
```

[ ]:
```
help.search(".test", package="stats")
```

```
R Information

Help files with alias or concept or title matching '.test' using
regular expression matching:


stats::ansari.test      Ansari-Bradley Test
  Aliases: ansari.test, ansari.test.default, ansari.test.formula
stats::bartlett.test    Bartlett Test of Homogeneity of Variances
  Aliases: bartlett.test, bartlett.test.default, bartlett.test.formula
stats::binom.test       Exact Binomial Test
  Aliases: binom.test
stats::Box.test         Box-Pierce and Ljung-Box Tests
  Aliases: Box.test
stats::chisq.test       Pearson's Chi-squared Test for Count Data
  Aliases: chisq.test
stats::cor.test         Test for Association/Correlation Between Paired
                        Samples
  Aliases: cor.test, cor.test.default, cor.test.formula
stats::fisher.test      Fisher's Exact Test for Count Data
  Aliases: fisher.test
stats::fligner.test     Fligner-Killeen Test of Homogeneity of
                        Variances
  Aliases: fligner.test, fligner.test.default, fligner.test.formula
stats::friedman.test    Friedman Rank Sum Test
  Aliases: friedman.test, friedman.test.default, friedman.test.formula
stats::kruskal.test     Kruskal-Wallis Rank Sum Test
  Aliases: kruskal.test, kruskal.test.default, kruskal.test.formula
stats::ks.test          Kolmogorov-Smirnov Tests
  Aliases: ks.test, ks.test.default, ks.test.formula
stats::mantelhaen.test
                        Cochran-Mantel-Haenszel Chi-Squared Test for
                        Count Data
  Aliases: mantelhaen.test
```

```
stats::mauchly.test     Mauchly's Test of Sphericity
  Aliases: mauchly.test, mauchly.test.SSD, mauchly.test.mlm
stats::mcnemar.test     McNemar's Chi-squared Test for Count Data
  Aliases: mcnemar.test
stats::mood.test        Mood Two-Sample Test of Scale
  Aliases: mood.test, mood.test.default, mood.test.formula
stats::oneway.test      Test for Equal Means in a One-Way Layout
  Aliases: oneway.test
stats::pairwise.prop.test
                        Pairwise comparisons for proportions
  Aliases: pairwise.prop.test
stats::pairwise.t.test
                        Pairwise t tests
  Aliases: pairwise.t.test
stats::pairwise.wilcox.test
                        Pairwise Wilcoxon Rank Sum Tests
  Aliases: pairwise.wilcox.test
stats::poisson.test     Exact Poisson tests
  Aliases: poisson.test
stats::power.anova.test
                        Power Calculations for Balanced One-Way
                        Analysis of Variance Tests
  Aliases: power.anova.test
stats::power.prop.test
                        Power Calculations for Two-Sample Test for
                        Proportions
  Aliases: power.prop.test
stats::power.t.test     Power calculations for one and two sample t
                        tests
  Aliases: power.t.test
stats::PP.test          Phillips-Perron Test for Unit Roots
  Aliases: PP.test
stats::print.htest      Print Methods for Hypothesis Tests and Power
                        Calculation Objects
  Aliases: print.htest, print.power.htest
stats::prop.test        Test of Equal or Given Proportions
  Aliases: prop.test
stats::prop.trend.test
                        Test for trend in proportions
  Aliases: prop.trend.test
stats::quade.test       Quade Test
  Aliases: quade.test, quade.test.default, quade.test.formula
stats::shapiro.test     Shapiro-Wilk Normality Test
  Aliases: shapiro.test
stats::stats-defunct    Defunct Functions in Package 'stats'
  Aliases: mauchley.test
stats::t.test           Student's t-Test
  Aliases: t.test, t.test.default, t.test.formula
```

```
stats::var.test         F Test to Compare Two Variances
  Aliases: var.test, var.test.default, var.test.formula
stats::wilcox.test      Wilcoxon Rank Sum and Signed Rank Tests
  Aliases: wilcox.test, wilcox.test.default, wilcox.test.formula
  Concepts: Mann-Whitney Test


Type '?PKG::FOO' to inspect entries 'PKG::FOO', or 'TYPE?PKG::FOO' for
entries like 'PKG::FOO-TYPE'.
```

7 Functions

[ ]:
```r
sumdice <- function(n, sides = 6) {
    if (sides < 1) {
        return(0)
    }
    k <- sample(1:sides, size = n, replace = TRUE)
    return(k)
}

sumdice(20)
```

1. 4 2. 1 3. 1 4. 1 5. 3 6. 1 7. 6 8. 4 9. 4 10. 2 11. 5 12. 6 13. 5 14. 1 15. 2 16. 1 17. 2 18. 3 19. 6 20. 4

8 Arrays, Data Frames, and Lists

…

9 Formula Specification

…

10 Graphics

…

11 Introduction to ggplot

…

12 Workspace and Files

…

13 The Working Directory

...

14 Reading Data from External Files

...

15 Importing/Exporting .csv Files

...

16 Using Scripts

...

17 Using Packages

...

18 Using R Markdown and knitr

...