# *Bài giảng 1: Giới Thiệu, Dữ liệu và Biến ngẫu nhiên*

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

**Thống kê Tính Toán - Computational Statistics**

# Mục lục


1 Thông tin của học phần

2 Mở đầu

3 Dữ liệu

4 Biến ngẫu nhiên và dữ liệu

5 Cơ bản về thống kê

# Thông tin của học phần

Giảng viên:

TS. Tô Đức Khánh
Bộ môn Xác suất - Thống kê, khoa Toán - Tin học
Trường Đại học Khoa học tự nhiên, Đại học Quốc gia Tp. HCM
e-mail: tdkhanh@hcmus.edu.vn
Phòng F.012, Trường Đại học Khoa học tự nhiên, 227 Nguyễn Văn Cừ, Phường Chợ Quán, Tp. Hồ Chí Minh.

# Thông tin của học phần

Nội dung của môn học trong 11 tuần bao gồm:

1. Nhắc lại một số kiến thức xác suất, thống kê, giới thiệu ngôn ngữ R;
2. Ước lượng mật độ của dữ liệu;
3. Làm trơn xu hướng của dữ liệu;
4. Bootstrap và phương pháp lấy lại mẫu;
5. Mô phỏng và phương pháp Monte Carlo;
6. Tối ưu hóa trong thống kê;
7. Thuật toán EM;
8. Markov Chain Monte Carlo;
9. Tính toán Bayes.

Thực hành trên ngôn ngữ R và kết hợp C/C++ (RCpp).

# Thông tin của học phần

*Cấu trúc điểm*

Học phần có cấu trúc điểm như sau:

- Điểm giữa kỳ: 40%
- Điểm cuối kỳ: 60%

*Hình thức thi*

- Điểm giữa kỳ: 4 bài tập.
- Điểm cuối kỳ: làm đề tài kết thúc học phần.
  *(Danh sách đề tài sẽ được cung cấp vào tuần thứ 11 của môn học.)*

## Yêu cầu kiến thức chuẩn bị

Trước khi tham gia khóa học này, học viên cần nắm được các khái niệm dưới đây:

1. các kiến thức cơ bản về thống kê mô tả: đồ thị tần số, biểu đồ mật độ, biểu đồ hộp (boxplot), giá trị trung bình, phương sai, ...;
2. các kiến thức cơ bản về hàm mật độ xác suất, hội tụ xác suất, phân phối của biến ngẫu nhiên, ...;
3. các kiến thức cơ bản về kiểm định giả thuyết thống kê;
4. các kiến thức về giải tích hàm (khả vi, liên tục, khả tích) hay đại số tuyến tính;
5. ngoài ra, các kiến thức về lập trình cũng cần thiết.

## Tài liệu học tập và tài liệu tham khảo

Slide bài giảng lý thuyết và thực hành được cập nhật tại trang Moodle của môn học.


- Givens, G. H., & Hoeting, J. A. (2013). *Computational Statistics* 2nd. John Wiley & Sons.
- Efron, B., & Hastie, T. (2021). *Computer Age Statistical Inference: Algorithms, Evidence, and Data science*. Cambridge University Press.
- Gentle, J. E. (2009). *Computational Statistics*. New York: Springer.
- Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Springer.

# Ngôn ngữ lập trình tính toán

R là phần mềm miễn phí được dùng phổ biến trong các phân tích dữ liệu thống kê, tính toán và biểu diễn đồ thị. Phần mềm được tải về theo địa chỉ dưới đây:

`https://cran.r-project.org/`

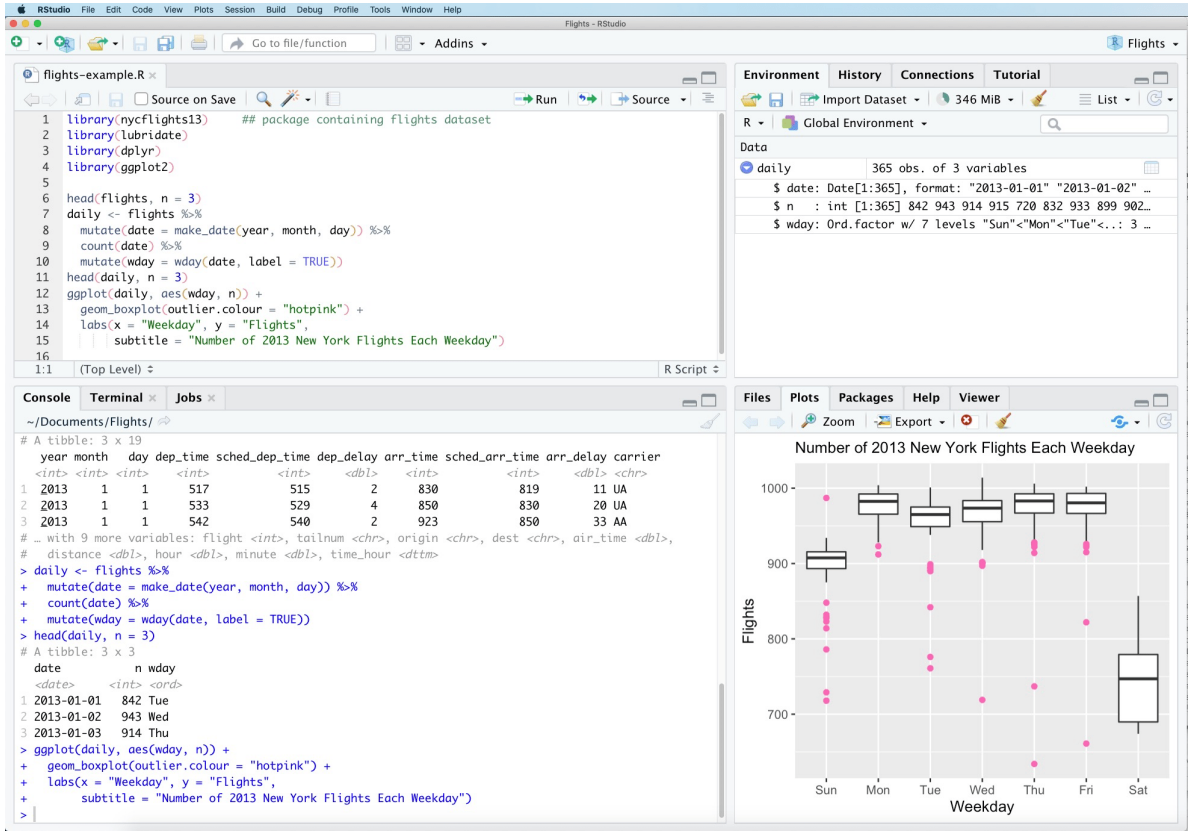
với các phiên bản khác nhau dành cho: Windows, MAC hoặc Linux.

Để cài đặt, ví dụ cho máy hệ điều hành Windows (7/8/10/11):

- lựa chọn "Download R for Windows";
- sau đó, chọn "install R for the first time";
- và "Download R 4.5.2 for Windows";

sau khi hoàn thành quá trình tải, mở tệp "R-4.5.2-win.exe" và cài đặt theo chỉ dẫn.

# Ngôn ngữ lập trình tính toán




Để thuận tiện cho quá trình code, và quản lý việc phân tích, ta sử dụng thêm phần mềm giao diện `RStudio`:

`https://posit.co/download/rstudio-desktop/`

Chú ý rằng, chỉ cài đặt `RStudio` sau khi đã cài đặt `R`.

1 Thông tin của học phần

2 Mở đầu

3 Dữ liệu

4 Biến ngẫu nhiên và dữ liệu

5 Cơ bản về thống kê

## Vai trò của Computing

*Nghịch lý*

Thống kê hiện đại và khoa học dữ liệu đòi hỏi các thuật toán tính toán mới và hiệu quả:

- Dữ liệu lớn phá vỡ nhiều gói phần mềm hiện có;
- Giải các mô hình phức tạp tốn nhiều thời gian;
- Tài nguyên phần cứng không được sử dụng theo cách hiệu quả nhất;
- Tiêu thụ năng lượng cao nhưng hiệu suất thấp.

Tuy nhiên, **computing** từ lâu đã là một lĩnh vực bị đánh giá thấp trong thống kê:

- Các nhà nghiên cứu có xu hướng dành nhiều thời gian hơn cho các đặc tính lý thuyết của mô hình và phương pháp;
- Những người thực hành thích sử dụng trực tiếp các gói phần mềm và ít quan tâm đến các chi tiết tính toán;
- Bản thân điện toán ít có khả năng là đối tượng nghiên cứu chính;
- Một thuật toán tính toán được thiết kế cẩn thận có thể chỉ được coi là một "thủ thuật" chứ không phải là những đóng góp vững chắc.

# Computing không đơn giản

**Ví dụ 1:** cho hai ma trận $A_{n\times m}$ và $B_{m\times n}$, tính $\mathrm{tr}(AB)$.

Trong R, thật dễ dàng để thi hành phép tính này

```
sum(diag(A %*% B))
```

Tuy nhiên,

```
A <- matrix(rnorm(10000 * 2000), 2000)
B <- matrix(rnorm(10000 * 2000), 10000)
system.time(sum(diag(A %*% B)))
```

kết quả là

```
##   user  system elapsed
##   7.56    0.00   19.94
```

Ta có thể làm tốt hơn

```
##   user  system elapsed
##   0.04    0.00    0.10
```

# Computing không đơn giản

**Ví dụ 2:** cho $f(x)$ là một hàm mật độ hỗn hợp

$$f(x) = 0.3\mathcal{N}(-1, 0.1) + 0.7\mathcal{N}(1, 0.01).$$

Viết một hàm để tính giá trị $\log(f(x))$. Kiểm tra kết quả tại $x = -15$ và $x = 15$.

Trên R:

```
logf <- function(x){
  f <- 0.3*dnorm(x, -1, sqrt(0.1)) + 0.7*dnorm(x, 1, 0.1)
  return(log(f))
}
logf(c(-15, 15))
```

cho kết quả:

```
## [1] -Inf -Inf
```

Ta có thể làm tốt hơn để nhận kết quả chính xác

```
## [1]  -980.9716 -1280.9716
```

Làm thế nào để tính được $\dfrac{\partial \log f(x)}{\partial x}$?

# Computing không đơn giản

**Ví dụ 3:** tính toán giá trị tiên đoán của mô hình hồi quy

$$\widehat{y} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Trong R, thật dễ dàng để thi hành công thức này:

```
X %*% solve(t(X) %*% X) %*% t(X) %*% Y
```

Tuy nhiên,

```
set.seed(123)
X <- matrix(rnorm(5000 * 500), 5000)
Y <- rnorm(5000)
system.time(X %*% solve(t(X) %*% X) %*% t(X) %*% Y)
```

kết quả là

```
##   user  system elapsed
##   2.88    0.06    5.72
```

Ta có thể làm tốt hơn

```
##   user  system elapsed
##   0.13    0.00    0.52
```

# Tại sao cần học computing

*Vai trò của computing*

- Nó thúc đẩy nghiên cứu thống kê hiện đại;
- Nhiều vấn đề vẫn chưa được giải quyết tốt và là chủ đề nghiên cứu tốt;
- Nó giúp chúng ta tận dụng tối đa các nguồn tài nguyên tính toán.

*Định nghĩa cơ bản*

- **Statistical Computing - Tính toán thống kê**

  Tập trung vào phát triển và triển khai các công cụ, thuật toán, và phần mềm để thực hiện phân tích thống kê.

- **Computational Statistics - Thống kê tính toán**

  Statistical Computing + các phương pháp thống kê yêu cầu tính toán chuyên sâu

  $\Leftrightarrow$ Sử dụng các kỹ thuật tính toán để giải quyết các vấn đề thống kê và phân tích dữ liệu.

# Tại sao cần học computing

*Phạm vi của Statistical Computing*

Tính toán thống kê là một lĩnh vực đa ngành:

- Thống kê
- Khoa học máy tính
- Giải tích số
- Tối ưu hóa
- Cơ sở dữ liệu
- Lập trình
- Đồ họa
- Kỹ thuật phần mềm
- Nghiên cứu có thể tái tạo

# Tại sao cần học computing

*Phạm vi của Computational Statistics*

Thống kê tính toán được xây dựng dựa trên thống kê toán học, tính toán thống kê và thống kê ứng dụng:

- Mọi thứ của Statistical Computing
- Mô phỏng Monte Carlo
- Kỹ thuật Bootstrap/lấy mẫu lại
- Thuật toán EM
- Ước tính mật độ
- Vận chuyển tối ưu tính toán

## Các vấn đề chính

*Các bài toán trong thống kê*

- Ước lượng: hàm mật độ, mô hình hồi quy, mô hình phân loại, mô hình phân cụm.
- Thống kê suy luận: khoảng tin cậy, thống kê kiểm định
- Mô phỏng: phát sinh dữ liệu ngẫu nhiên, tích phân ngẫu nhiên
- Thống kê Bayes: xấp xỉ phân phối hậu nghiệm, trung bình hậu nghiệm, mô phỏng tiên đoán.

## Các vấn đề chính

Học phần nhắm tới giải quyết các vấn đề chính sau:

1. Ước lượng
   - Ước lượng mật độ;
   - Làm trơn xu hướng của dữ liệu hai chiều;
   - Tối ưu hóa bài toán ước lượng;
   - Thuật toán EM cho các bài toán tối ưu phức tạp;
2. Thống kê suy luận
   - Bootstrap;
   - Permutation test;
   - Jackknife;
3. Mô phỏng
   - Mô phỏng Monte Carlo
   - Tích phân Monte Carlo;
   - Importance sampling;
4. Thống kê Bayes
   - Markov Chain Monte Carlo;
   - Thuật toán Metropolis-Hastings;
   - Gibbs sampling.

Thêm vào đó

- Numerical stability
- Kết hợp C++ code với R

1 Thông tin của học phần

2 Mở đầu

3 Dữ liệu

4 Biến ngẫu nhiên và dữ liệu

5 Cơ bản về thống kê

# Dữ liệu?

*Dữ liệu*

Dữ liệu là một hoặc nhiều thông tin mô tả **tính chất của các đối tượng nghiên cứu**.

Dữ liệu có thể tới từ nhiều nguồn:

- cảm biến (sensor);
- sự kiện;
- văn bản (text);
- hình ảnh (images);
- API;
- âm thanh;
- video.

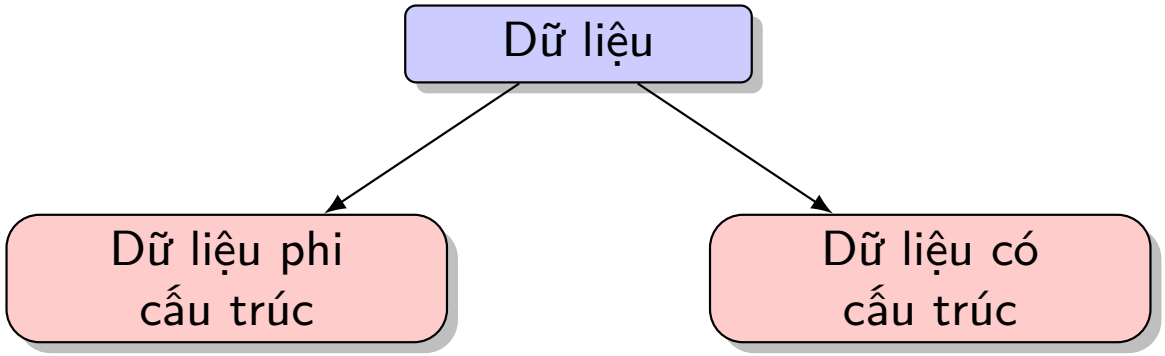
# Dữ liệu?

*Dữ liệu*

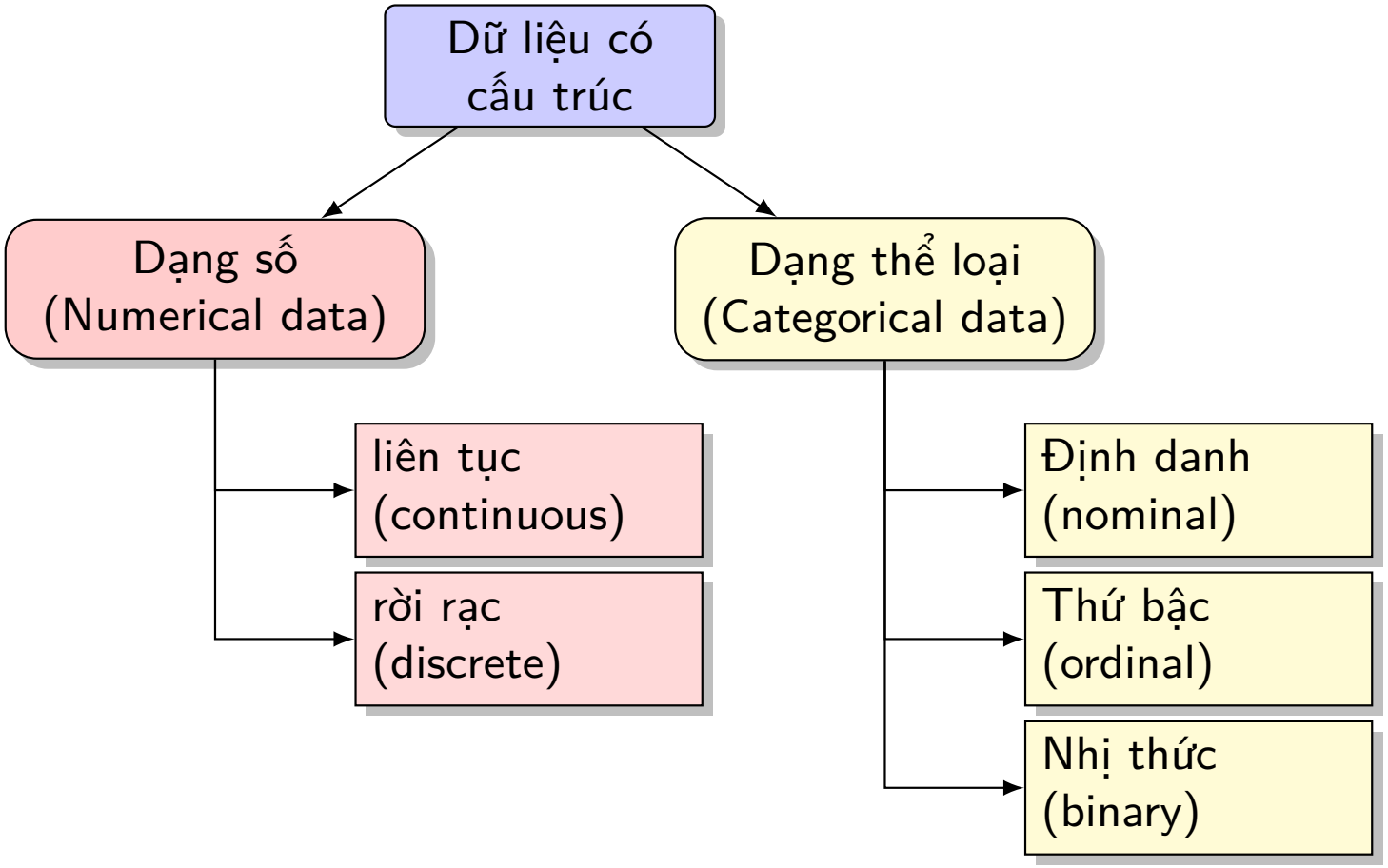
Dữ liệu là một hoặc nhiều thông tin mô tả **tính chất của các đối tượng nghiên cứu**.

Dữ liệu có thể tới từ nhiều nguồn:

- cảm biến (sensor);
- sự kiện;
- văn bản (text);
- hình ảnh (images);
- API;
- âm thanh;
- video.



# Dữ liệu có cấu trúc - Structured data



## *Dữ liệu có cấu trúc - Structured data*

*Dữ liệu dạng số* (*numerical data* hay *quantitative data*) là dữ liệu được cấu trúc với giá trị là số:

- liên tục (*continuous*);
- rời rạc hay số đếm (*discrete*),

ghi lại/biểu diễn số lượng của một đối đối tượng nào đó.

*Dữ liệu dạng phân loại* (*categorical data* hay *qualitative data*) là dữ liệu được cấu trúc với giá trị là các thể loại khác nhau của người hoặc vật. Thông thường ta có hai dạng phổ biến sau:

- biến định danh (*nominal*)
- biến thứ bậc (*ordinal*)

Đặc biệt, đối với dữ liệu định danh, ta có:

- dữ liệu nhị định danh hay dữ liệu nhị phân (*binary data*), khi biến có 2 giá trị;
- dữ liệu đa định danh hay dữ liệu đa thức (*nominal multinomial data* hoặc *multinomial data*).

## Dữ liệu có cấu trúc - Structured data

Ví dụ:

- số liên tục: tốc độ gió, doanh thu, khoảng thời gian, chi phí, ...
- số rời rạc: số lượng lượt truy cập web, số lượng khách hàng trong một ngày, ...
- đa định danh: thể loại màn hành TV: plasma, LCD, LED, ...
- thứ bậc: điểm đánh giá sản phẩm, thăng hài lòng về chất lượng dịch vụ, ...
- nhị phân: có/không, đúng/sai, ...

# Ví dụ dữ liệu cấu trúc

Chi phí quảng cáo trên các nền tảng:

- truyền hình (`TV`)
- đài phát thanh (`radio`)
- nhật báo in (`newspaper`)

và doanh thu bán sản phẩm của công ty đều là các dữ liệu dạng số liên tục.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | TV | radio | newspaper | sales |
| 2 | 1 | 230.1 | 37.8 | 69.2 | 22.1 |
| 3 | 2 | 44.5 | 39.3 | 45.1 | 10.4 |
| 4 | 3 | 17.2 | 45.9 | 69.3 | 9.3 |
| 5 | 4 | 151.5 | 41.3 | 58.5 | 18.5 |
| 6 | 5 | 180.8 | 10.8 | 58.4 | 12.9 |
| 7 | 6 | 8.7 | 48.9 | 75 | 7.2 |
| 8 | 7 | 57.5 | 32.8 | 23.5 | 11.8 |
| 9 | 8 | 120.2 | 19.6 | 11.6 | 13.2 |
| 10 | 9 | 8.6 | 2.1 | 1 | 4.8 |
| 11 | 10 | 199.8 | 2.6 | 21.2 | 10.6 |
| 12 | 11 | 66.1 | 5.8 | 24.2 | 8.6 |

# *Dummy variable - Biến giả*

*Dummy variable - Biến giả*

**Biến giả** (Dummy variable) là biến số được mã hoá từ biến thể loại (categorical variable) thành giá trị số, thường là 0 hoặc 1, để sử dụng trong mô hình hồi quy hoặc phân tích thống kê.

**Mục đích:**

- Cho phép đưa biến định tính vào mô hình hồi quy.
- Đại diện cho sự có mặt/không có mặt của một đặc tính.

**Ví dụ:**

- Biến "Giới tính": {Nam, Nữ} $\rightarrow$ 1 biến giả $D = \{0, 1\}$, Nam $= 0$, Nữ $= 1$.
- Biến "Màu sắc": {Đỏ, Xanh, Vàng} $\rightarrow$ 2 biến giả

$$D_1 = \begin{cases} 1 & \text{nếu Đỏ,} \\ 0 & \text{nếu ngược lại,} \end{cases} \qquad D_2 = \begin{cases} 1 & \text{nếu Xanh,} \\ 0 & \text{nếu ngược lại,} \end{cases}$$

Vàng được chọn làm baseline, tức là mức tham chiếu.

**Chú ý:** Cần tránh dummy variable trap (đa cộng tuyến hoàn hảo) $\rightarrow$ với $k$ danh mục, chỉ dùng $k - 1$ biến giả.

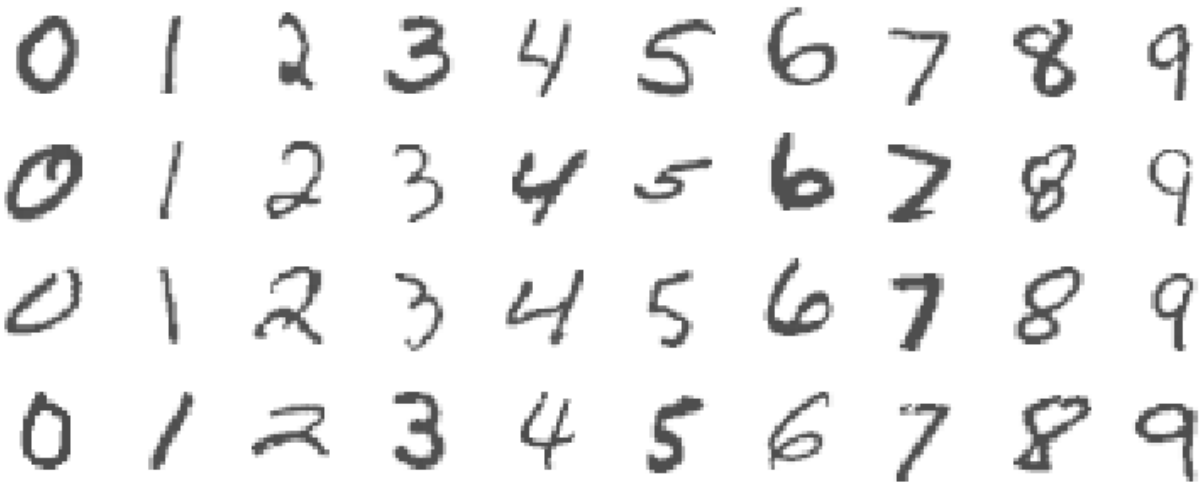
# Dữ liệu phi cấu trúc - Unstructured data

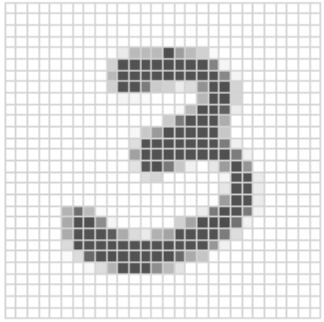
**Dữ liệu phi cấu trúc** là những dữ liệu được thu thập không có dạng số hoặc dạng thể loại:

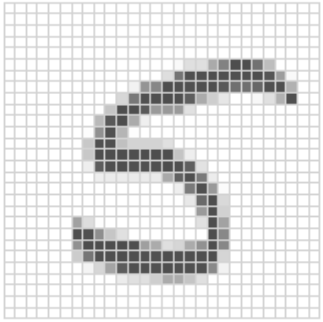
- hình ảnh - images: là tập hợp các pixel với mỗi pixel chứa thông tin màu RGB (đỏ - red, xanh lá - green, xanh dương - blue);
- văn bản - text: là chuỗi các từ và ký tự không phải từ, thường được sắp xếp theo các phần, phần phụ, v.v;
- dòng nhấp chuột - clickstreams: là chuỗi hành động do người dùng tương tác với một ứng dụng hoặc trang web.

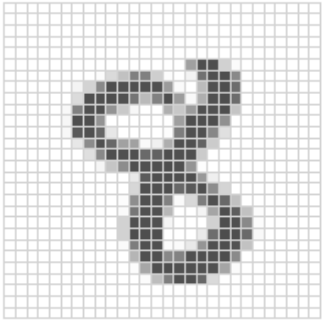
Để áp dụng các mô hình của statistical learning, dữ liệu thô phi cấu trúc phải được xử lý và thao tác thành dạng có cấu trúc.

# Ví dụ dữ liệu phi cấu trúc







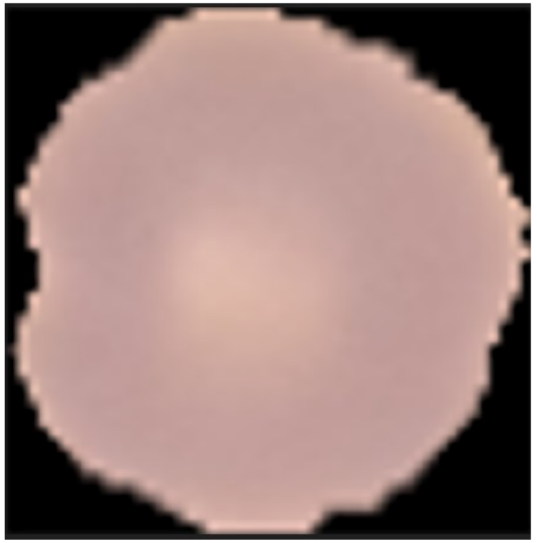


# Ví dụ dữ liệu phi cấu trúc

Dữ liệu về hình chụp phiết máu mỏng của hai người:

- nhiễm ký sinh trùng (a);
- không bị nhiễm ký sinh trùng (b).

(a) Parasitized



(b) Uninfected

# Rectangular Data

*Rectangular Data*

**Rectangular Data** có hình dạng giống như hình chữ nhật, tức là một ma trận hai chiều với các hàng (row) biểu thị các đối tượng (trường hợp) và các cột (column) biểu thị các đặc tính (feature) hay biến (variable).

| Category | currency | sellerRating | Duration | endDay | ClosePrice | OpenPrice | Competitive? |
|---|---|---|---|---|---|---|---|
| Music/Movie/Game | US | 3249 | 5 | Mon | 0.01 | 0.01 | 0 |
| Music/Movie/Game | US | 3249 | 5 | Mon | 0.01 | 0.01 | 0 |
| Automotive | US | 3115 | 7 | Tue | 0.01 | 0.01 | 0 |
| Automotive | US | 3115 | 7 | Tue | 0.01 | 0.01 | 0 |
| Automotive | US | 3115 | 7 | Tue | 0.01 | 0.01 | 0 |
| Automotive | US | 3115 | 7 | Tue | 0.01 | 0.01 | 0 |
| Automotive | US | 3115 | 7 | Tue | 0.01 | 0.01 | 1 |
| Automotive | US | 3115 | 7 | Tue | 0.01 | 0.01 | 1 |

# Rectangular Data

Các thành phần của rectangular data:

- feature: một cột trong bảng dữ liệu thường ám chỉ tới đặc tính (feature) của đối tượng,

  đồng nghĩa với attribute, input, predictor, variable;
- outcome: là một đặc tính của đối tượng, được sử dụng như là kết quả của sự tiên đoán, hay phân loại,

  đồng nghĩa với dependent variable, response, target, output;
- records: một dòng của bảng dữ liệu thường ám chỉ tới một ghi chép của một đối tượng;

  đồng nghĩa với case, example, instance, observation, pattern, sample.
- data frame: bảng dữ liệu tổng hợp nhiều cột và dòng, biểu diễn ghi chép các đặc tính khác nhau của nhiều đối tượng;

  một bảng dữ liệu có thể chứa hỗn hợp các dạng dữ liệu có cấu trúc.

## Rectangular Data

Tuy nhiên, dữ liệu mà ta làm việc, không phải lúc nào cũng ở sẵn dạng rectangular data.

- dữ liệu phi cấu trúc: phải được xử lý và thao tác để có thể biểu diễn dưới dạng tập hợp các tính năng;
- dữ liệu trong cơ sở dữ liệu phải được trích xuất và đưa vào một bảng duy nhất cho hầu hết các nhiệm vụ phân tích và lập mô hình dữ liệu.

## Non-rectangular Data

Một số dạng dữ liệu sau không phải rectangular data:

- dữ liệu dạng danh sách: trong đó, mỗi danh sách là một rectangular data (cross-sectional data, panel data);
- dữ liệu chuỗi thời gian (time series data): ghi lại các phép đo liên tiếp của cùng một biến (feature);
- dữ liệu không gian (spatial data): được sử dụng trong phân tích bản đồ và vị trí;
- dữ liệu đồ thị hoặc mạng lưới (graph, network data): được sử dụng để thể hiện các mối quan hệ vật lý, xã hội và trừu tượng.

1 Thông tin của học phần

2 Mở đầu

3 Dữ liệu

4 Biến ngẫu nhiên và dữ liệu

5 Cơ bản về thống kê

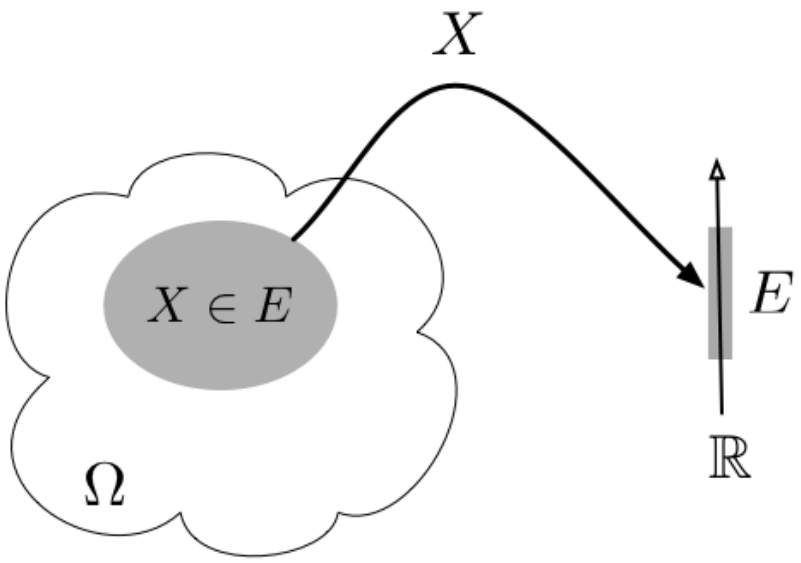
# Biến ngẫu nhiên

*Định nghĩa biến ngẫu nhiên - công thức toán học*

Xét phép thử ngẫu nhiên, với không gian mẫu $\Omega$. Gọi $\omega$ là một biến cố sơ cấp (hay kết quả) trong không gian mẫu $\Omega$. Khi đó, hàm số $X$ được xác định bởi:

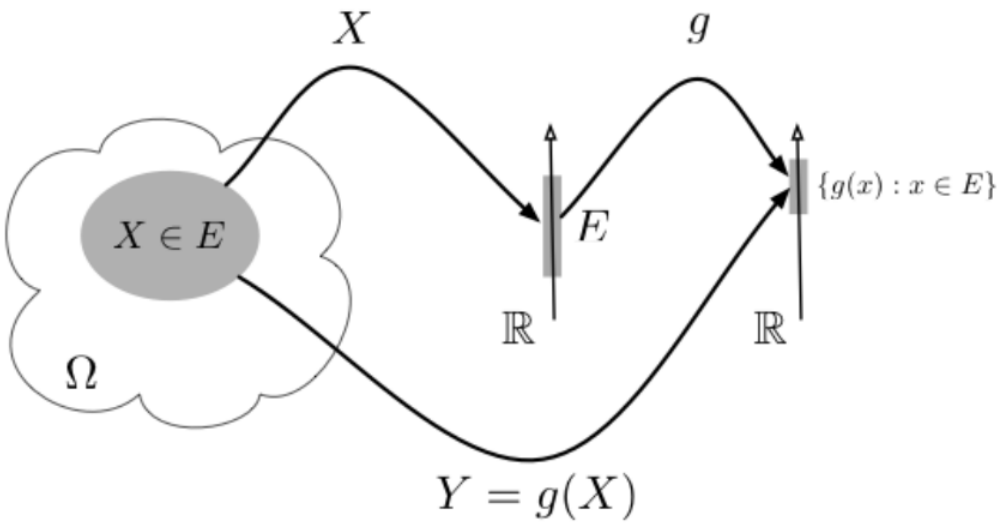
$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

được gọi là một biến ngẫu nhiên.



# Biến ngẫu nhiên

Một phép biến đổi được áp dụng trên 1 biến ngẫu nhiên được gọi là hàm ngẫu nhiên (random function)



# Biến ngẫu nhiên

Xét về mặt giá trị số, ta cơ bản chia biến ngẫu nhiên thành hai loại:

- biến ngẫu nhiên **rời rạc**;
- biến ngẫu nhiên **liên tục**.

*Biến ngẫu nhiên rời rạc*

Biến ngẫu nhiên rời rạc là biến ngẫu nhiên có dạng số rời rạc nhau (số tự nhiên, số nguyên) có phạm vi hữu hạn (hoặc vô hạn) đếm được.

Ví dụ: số lần truy cập trang web, số xe đạp được thuê trong 1 ngày.

Ví dụ cho giá trị: 0, 1, 2, 3, 4, ...

*Biến ngẫu nhiên liên tục*

Biến ngẫu nhiên liên tục là một biến ngẫu nhiên có vô số kết quả có thể xảy ra, được biểu thị bằng một khoảng trên trục số thực hoặc toàn bộ trục số thực.

Ví dụ: chiều dài, áp suất, nhiệt độ, thời gian, điện áp, trọng lượng.

Ví dụ cho giá trị: 0.1, 0.123, 2.36, 3.85, 4.235, ...

# Đặc trưng của biến ngẫu nhiên

Biến ngẫu nhiên được đặc trưng bởi các đại lượng

- phân phối xác suất (hàm xác suất, hàm phân phối tích lũy)
- kỳ vọng (expected value)
- phương sai (variance)

*Hàm trọng lượng xác suất*

Xét biến ngẫu nhiên rời rạc $X$, với các giá trị có thể $x_1$, $x_2$, $x_3$, ..., $x_k$, một hàm trọng lượng xác suất $f$ của $X$ là một hàm thỏa các tính chất sau:

- $f(x_i) = P(X = x_i)$;
- $0 \leq f(x_i) \leq 1$;
- $\sum_{i=1}^{k} f(x_i) = f(x_1) + f(x_2) + \ldots + f(x_k) = 1$.

# Đặc trưng của biến ngẫu nhiên

*Hàm mật độ xác suất*

Hàm miêu tả mật độ tập trung của một biến ngẫu nhiên liên $X$ trong không gian giá trị $S$, được gọi là *hàm mật độ xác suất*, ký hiệu là $f(x)$, với các tính chất sau:

- $f(x)$ là khả tích;
- $f(x) \geq 0$, với mọi giá trị của $x$;
- $\int_S f(x)\,\mathrm{d}x = 1$;
- $P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$, tức là diện tích phía dưới đường cong xác định bởi hàm $f(x)$ trong khoảng $[a, b]$ của $X$.

# Đặc trưng của biến ngẫu nhiên

*Hàm phân phối tích lũy rời rạc*

Cho biến ngẫu nhiên rời rạc $X$, với các giá trị có thể $x_1, x_2, \ldots, x_n$ và hàm trọng lượng xác suất $f(x)$. Hàm phân phối xác suất tích lũy, (*cumulative distribution function*), $F(x)$ của $X$ là

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

Thỏa những điều kiện sau:

- $0 \leq F(x) \leq 1$;
- nếu $x \leq y$ thì $F(x) \leq F(y)$ .

# Đặc trưng của biến ngẫu nhiên

*Hàm phân phối tích lũy liên tục*

Xét biến ngẫu nhiên liên tục $X$ có không gian giá trị $\mathbb{R}$ và hàm mật độ $f(x)$. Hàm phân phối tích lũy của $X$ được định nghĩa:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)\,\mathrm{d}s.$$

Ta có một số tính chất của $F(x)$ như sau:

- $F(x)$ là hàm liên tục, khả vi;
- đạo hàm của $F(x)$ là $f(x)$, tức là $\dfrac{\mathrm{d}F(x)}{\mathrm{d}x} = f(x)$.

## *Đặc trưng của biến ngẫu nhiên*

*Kỳ vọng và phương sai của biến ngẫu nhiên rời rạc*

Xét biến ngẫu nhiên rời rạc $X$ với $S = \{x_1, x_2, \ldots, x_k\}$, và hàm trọng lượng xác suất $f(x)$. **Kỳ vọng** (hay **trung bình**) của $X$ được ký hiệu là $\mathbb{E}(X)$ hay $\mu_X$, xác định bởi:

$$\mathbb{E}(X) = \sum_{i=1}^{k} x_i f(x_i).$$

**Phương sai** của $X$ được ký hiệu là $\mathbb{V}\mathrm{ar}(X)$ hay $\sigma_X^2$, xác định bởi:

$$\mathbb{V}\mathrm{ar}(X) = \sum_{i=1}^{k} (x_i - \mathbb{E}(X))^2 f(x_i).$$

hay tương đương với

$$\mathbb{V}\mathrm{ar}(X) = \sum_{i=1}^{k} x_i^2 f(x_i) - \mathbb{E}(X)^2.$$

Căn bậc hai của phương sai, $\sqrt{\mathbb{V}\mathrm{ar}(X)}$, được gọi là **độ lệch chuẩn** của $X$, ký hiệu là $\sigma_X$.

## Đặc trưng của biến ngẫu nhiên

*Kỳ vọng và phương sai của biến ngẫu nhiên liên tục*

Xét $X$ là một biến ngẫu nhiên liên tục trên $\mathbb{R}$, có hàm mật độ xác suất $f(x)$. Khi đó, kỳ vọng (hay trung bình) của $X$, được ký hiệu là $\mathbb{E}(X)$ hay $\mu_X$:

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x\, f(x)\, \mathrm{d}x$$

Phương sai của $X$ được ký hiệu là $\mathbb{V}\mathrm{ar}(X)$ hay $\sigma_X^2$:

$$\mathbb{V}\mathrm{ar}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2\, f(x)\, \mathrm{d}x = \int_{-\infty}^{+\infty} x^2\, f(x)\, \mathrm{d}x - \mathbb{E}(X)^2.$$

Căn bậc 2 của phương sai, $\sqrt{\mathbb{V}\mathrm{ar}(X)}$ được gọi là **độ lệch chuẩn** của $X$, ký hiệu là $\sigma_X$.

## Liên hệ giữa biến ngẫu nhiên và dữ liệu có cấu trúc

Nhắc lại, dữ liệu có cấu trúc có dạng bảng:

- 1 cột = 1 đặc tính của tất cả đối tượng, tức là biến;
- 1 dòng = 1 bộ giá trị quan sát được của các biến của 1 đối tượng.

Ta có các nhận xét như sau:

- Quan sát 1 đặc tính của 1 đối tượng $\Leftrightarrow$ 1 phép thử ngẫu nhiên.
- Kết quả của 1 phép thử ngẫu nhiên $\Rightarrow$ 1 biến ngẫu nhiên.

Do đó, 1 đặc tính của 1 đối tượng $\Leftrightarrow$ 1 biến ngẫu nhiên.

$\Rightarrow$ mô hình hóa đặc tính thành 1 biến ngẫu nhiên, nhằm:

- hiểu về sự phân bố giá trị (khả năng kết quả xuất hiện);
- đưa được đặc tính vào mô hình thống kê/máy học.

## *Liên hệ giữa biến ngẫu nhiên và dữ liệu có cấu trúc*

*Ví dụ 1:*

Xét dữ liệu của một trang thương mại điện tử, trong đó, ghi lại thông tin và trạng thái mua/không mua hàng của từng khách hàng.

Trạng thái mua/không mua hàng của 1 khách hàng $\Leftrightarrow$ 1 biến ngẫu nhiên Bernoulli, với xác suất mua hàng là $p$ với mọi khách hàng, và

$$\text{Trạng thái} = \begin{cases} 0, & \text{nếu không mua,} \\ 1, & \text{nếu mua.} \end{cases}$$

*Ví dụ 2:*

Xét dữ liệu của một trang thương mại điện tử, trong đó, ghi lại thông tin và số lượng sản phẩm đã mua của từng khách hàng.

Số lượng sản phẩm đã mua của 1 khách hàng $\Leftrightarrow$ 1 biến ngẫu nhiên Poisson, với trung bình sản phẩm là $\lambda$ với mọi khách hàng.

*Ví dụ 3:*

Xét dữ liệu của một công ty tài chính, trong đó, ghi lại thông tin và thu nhập trung bình của từng khách hàng.

Thu nhập trung bình của 1 khách hàng $\Leftrightarrow$ 1 biến ngẫu nhiên Gamma.

# Sơ đồ minh hoạ: biến ngẫu nhiên tới dữ liệu

**Biến ngẫu nhiên**
$X$

$\xrightarrow{\text{định nghĩa}}$

**Phân phối**
(PMF / PDF)
$P(X = x)$ hoặc $f_X(x)$

$\xrightarrow{\text{quan sát / mẫu}}$

**Dữ liệu (bảng)**

| ID | Giá trị của biến $X$ |
|---|---|
| 1 | 170 |
| 2 | 165 |
| 3 | 180 |
| 4 | 172 |
| 5 | 169 |

Toán học: không chắc chắn
(ví dụ: $X$ = chiều cao)

Mô tả xác suất:
ví dụ $X \sim \mathcal{N}(\mu, \sigma^2)$

Quan sát thực tế (samples)
dùng cho ước lượng, mô hình hoá

**Ghi chú:** Tập dữ liệu (bảng) thường có nhiều cột (mỗi cột là một biến ngẫu nhiên) và nhiều hàng (mỗi hàng là một quan sát độc lập giả định).

1 Thông tin của học phần

2 Mở đầu

3 Dữ liệu

4 Biến ngẫu nhiên và dữ liệu

5 Cơ bản về thống kê

# Khái niệm cơ bản

*Tham số*

**Tham số** (**Parameter**) là một số hoặc một vector (thường cố định và không biết), đóng vai trò:

- đại lượng định hình phân phối của biến ngẫu nhiên;
- đặc trưng của biến ngẫu nhiên;
- hoặc phản ánh tính chất của tổng thể (hoặc mô hình).

Ta thường ký hiệu tham số là

- $\theta \in \mathbb{R}$ trong trường hợp một chiều (1 số thực duy nhất);
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ trường hợp $p$ chiều (1 vector thực).

**Ví dụ:**

- Trung bình của $X$ là 1 tham số, ký hiệu là $\mu$.
- Phương sai của $X$ là 1 tham số, ký hiệu là $\sigma^2$.
- Trong mô hình hồi quy tuyến tính

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

$\beta_0$ và $\beta_1$ là các tham số đóng vai trò là các hệ số (coefficient) của mô hình.

# Khái niệm cơ bản

*Mẫu ngẫu nhiên*

Gọi $F(\cdot;\boldsymbol{\theta})$ là một phân phối xác suất xác định theo tham số $\boldsymbol{\theta}$. Một dãy biến ngẫu nhiên $X_1,\ldots,X_n$ được gọi là **một mẫu ngẫu nhiên** (**random sample**, hay **sample**) kích thước $n$ từ phân phối $F(\cdot)$ nếu và chỉ nếu:

- $X_1,\ldots,X_n$ là độc lập nhau (independent);
- $X_1,\ldots,X_n$ có cùng phân phối $F$ (identically distributed).

Khi đó, ta ký hiệu ngắn gọn:

$$X_1,\ldots,X_n \overset{\text{i.i.d}}{\sim} F(\cdot;\boldsymbol{\theta}),$$

trong đó, i.i.d là viết tắt của "**independent and identically distributed**".

**Ý nghĩa:**

- ngẫu nhiên hàm ý mỗi $X_i$ là một biến ngẫu nhiên;
- sự độc lập có nghĩa là $X_i$ không mang thông tin của $X_j$, với $i \neq j$;
- sự cùng phân phối tức là tất cả các quan sát được sinh ra từ cùng một phân phối (hay cùng một cơ chế sinh dữ liệu - sẽ được lấy mẫu từ cùng một quần thể).

## Khái niệm cơ bản

Ta cần phân biệt rõ:

*Mẫu ngẫu nhiên:* $X_1, \ldots, X_n$ tương ứng là đối tượng lý thuyết;

*Dữ liệu quan sát:* $x_1, \ldots, x_n$ là $n$ giá trị quan sát được của các $X_1 \ldots, X_n$ sau khi thực hiện lấy mẫu một cách ngẫu nhiên.

Thống kê (xây dựng phương pháp luận) làm việc với mẫu ngẫu nhiên, trong khi tính toán làm việc với dữ liệu đã quan sát.

**Ví dụ:** Mẫu ngẫu nhiên tuân theo phân phối chuẩn

$$X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

trong đó, $\mu$ và $\sigma^2$ là hai tham số cho trung bình và phương sai.

**Ví dụ:** Mẫu không ngẫu nhiên - mô hình chuỗi thời gian

$$X_t = \rho X_{t-1} + \varepsilon_t,$$

các $X_t$ là phụ thuộc nhau.

# Statistic vs. Estimator

*Thống kê - Statistic*

Cho mẫu ngẫu nhiên

$$X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} F(\cdot; \boldsymbol{\theta}).$$

Một thống kê (statistic) là một hàm thực của mẫu ngẫu nhiên, không phụ thuộc vào tham số chưa biết $\boldsymbol{\theta}$, ta ký hiệu

$$T = T(X_1, \ldots, X_n).$$

Chú ý:

- Statistic là **đại lượng ngẫu nhiên**;
- Thống kê $T$ có phân phối và phân phối này phụ thuộc vào $\boldsymbol{\theta}$ và phân phối của mẫu ngẫu nhiên, nhưng bản thân công thức của $T$ thì không.

Khi áp dụng $T$ lên trên dữ liệu quan sát, ta thu được giá trị thống kê (observed statistic):

$$t_{\text{obs}} = T(x_1, \ldots, x_n).$$

## Statistic vs. Estimator

**Ví dụ:**

- trung bình mẫu

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

- phương sai mẫu

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- thống kê kiểm định $H_0 : \mu = \mu_0$

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

Nếu $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$ thì

- $\overline{X} \sim \mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$;
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ (phân phối Chi-square bậc $n-1$);
- $T \sim t_{n-1}$ (phân phối $t$-student bậc $n-1$).

## Statistic vs. Estimator

*Ước lượng - Estimator*

Một ước lượng (estimator) là một thống kê được dùng để ước lượng một tham số chưa biết $\theta$. Ta ký hiệu:

$$\widehat{\theta} = T(X_1, \ldots, X_n).$$

Giá trị ước lượng (estimate) là kết quả thu được của ước lượng sau quá trình tính toán với dữ liệu quan sát.

**Chú ý:** Mọi ước lượng đều là thống kê, nhưng không phải thống kê nào cũng là ước lượng.

**Ví dụ:** Xét mẫu ngẫu nhiên

$$X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Thống kê $T = \max\{X_1 \ldots, X_n\}$, nó không phải là ước lượng, do nó không được thiết kế nhằm ước lượng tham số $\mu, \sigma^2$.

# Làm gì với Statistic và Estimator

*Làm gì với Statistic*

Thống kê (Statistic) là đại lượng toán học, do đó,

- khảo sát các tính chất toán học;
- khảo sát tính đủ hoặc đầy đủ của thống kê - Sufficiency (Completeness), tức là mang đủ hoặc không dư thừa thông tin cho tính chất của tham số.

$\Rightarrow$ Các công việc mang tính lý thuyết thống kê.

*Làm gì với Estimator*

Ước lượng (Estimator) là đại lượng thống kê có mục tiêu suy luận cho tổng thể dựa trên dữ liệu, do đó, ta cần

- cách thực hiện tính toán (computation method);
- khảo sát tính vững (consistency);
- khảo sát tính không chệch (unbiased);
- khảo sát quy luật phân phối (sampling distribution);
- đánh giá sai số của ước lượng (mean squared error - MSE).

$\Rightarrow$ Các công việc mang tính lý thuyết thống kê và thực hiện trên máy tính với thuật toán và dữ liệu.

# Tài liệu tham khảo


- Wasserman, L. (2004). *All of Statistics: a Concise Course in Statistical Inference*. Springer Science & Business Media.
- Agresti, A., & Kateri, M. (2021). *Foundations of Statistics for Data Scientists: with R and Python*. Chapman and Hall/CRC.
- Allan, G. (2009). *An Intermediate Course in Probability*, 2nd, Springer Verlag.