# Xác Suất Thống Kê

Võ Chí Công

2022/8/25

# Table of contents

# Lời nói đầu

"Xác suất" và "thống kê" là hai môn học gây cho tôi nhiều khó khăn nhất. Môn "xác suất" tôi đã học những kiến thức cơ bản vài lần, tính ra là ở cấp 3, trong đại học, và gần đây là học trực tuyến. Môn "thống kê" tôi chưa học được cho ra bài bản lần nào, mấy tháng đầu năm 2022 có thử sức học trực tuyến nhưng đầu tư thời gian không đủ nên thi rớt thảm hại.

Lần này nhất quyết học lại môn thống kê một cách nghiêm túc hơn, tôi tóm tắt lại kiến thức xác suất thống kê bằng tiếng Việt, mặc dù tài liệu học hầu hết là tiếng Anh, tiếng Nhật. Hy vọng tiếng mẹ đẻ sẽ giúp tôi hiểu rõ hơn các vấn đề, và trau dồi vốn từ vựng để chia sẻ kiến thức với các đồng nghiệp và bạn bè người Việt.

Động cơ của việc học xác suất thống kê của tôi là để hiểu rõ hơn các lý thuyết căn bản trong ngành học máy và trí tuệ nhân tạo và áp dụng vào thực tiễn một cách đúng đắn, an toàn và công bằng hơn.

Mục tiêu cụ thể trước mắt tôi đặt ra là học hiểu và lấy được chứng chỉ hoàn thành khoá học Fundamentals of Statistics của Philippe Rigollet (2022), giáo sư đại học MIT dạy trên nền tảng trực tuyến edX. Tài liệu tham khảo là quyển "All of Statistics" của Wasserman (2004), cũng chính là tài liệu tham khảo của khoá học nêu trên.

Môn xác suất nghiên cứu cách suy luận ra các đặc tính của tập dữ liệu sẽ được tạo ra từ một nguyên lý, quy trình sản sinh dữ liệu. Ngược lại, môn thống kê nghiên cứu cách dự đoán đặc tính của một quy trình sản sinh dữ liệu từ tập dữ liệu về hiện tượng đã phát sinh và được quan sát. Xem minh hoạ ở hình 1.

Phân tích, khai thác dữ liệu, học máy và khoa học dữ liệu là những tên gọi khác của thống kê, tùy theo bối cảnh và trào lưu. Một số ứng dụng cụ thể của thống kê là tính toán hồi quy, mật độ, phân loại và giả lập.

Tài liệu này không đi sâu vào các chứng minh chi tiết, nhưng sẽ cố gắng ghi rõ các công thức và định nghĩa. Thuật ngữ chuyên môn trong tài liệu này chắc chắn có nhiều chỗ không chuẩn chỉnh do vốn tiếng Việt và kiến thức hạn chế của tác giả. Xin vui lòng góp ý tại GitHub issues.

Tài liệu này được viết bằng các công cụ là Quarto và VSCode. Truy cập trực tuyến xstk và có thể tải PDF.

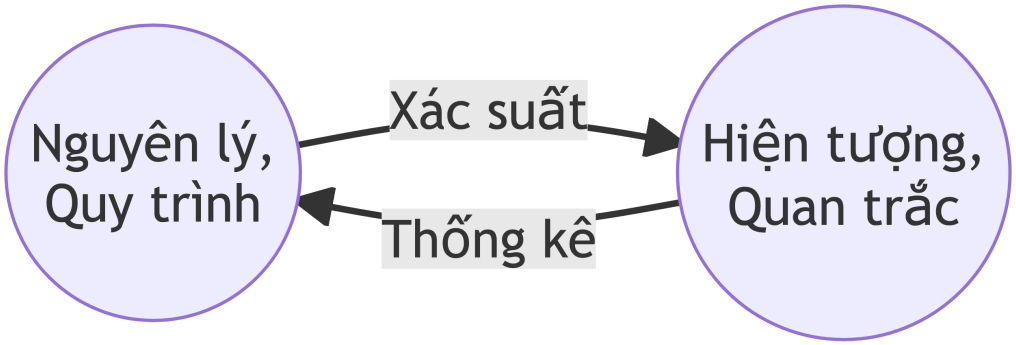


Figure 1: Xác suất và thống kê.

# 1 MITx 18.6501x

"Fundamentals of Statistics" (MITx 18.6501x ) là khóa học của Philippe Rigollet (2022) đại học MIT dạy trên edX.

- Important dates
- Calendar (private)

## 1.1 Unit 0

### 1.1.1 Kỳ hạn nộp bài

- Exercises and homework: Wednesdays 11:59AM UTC (Wed. 20:59 JST)
- Exams (48 hours): Tuesdays 11:59AM UTC(Tue. 20:59 JST)

### 1.1.2 Thời gian cần thiết

Mỗi tuần khoảng hơn 12 tiếng

- 5-7 hours on exercises, including 3 hours of lecture clips
- 1-2 hours watching recitations
- 5-7 hours for weekly problem sets

### 1.1.3 Cách tính điểm

Điểm thi đậu chứng chỉ là 60% tổng số điểm tối đa.

- 20% for the lecture exercises (divided equally among the 20 out of 23 lectures)
- 20% for the homeworks (divided equally among 10 (out of 12) homeworks)
- 18% for the first midterm exam (timed)
- 18% for the second midterm exam (timed)
- 24% for the final exam (timed)

# 2 Xác suất

## 2.1 Sự kiện

**Không gian** $\Omega$ là tập hợp chứa tất cả những **hiện tượng** $\omega$ có thể xảy ra từ một **thí nghiệm**. Các tập con của $\Omega$ là các **sự kiện**.

Ví dụ xem xét thí nghiệm tung một đồng xu đúng hai lần, quan sát đồng xu rớt xuống nằm ngửa ($H$) hay sấp ($T$), ta có $\Omega = \{HH, HT, TH, TT\}$ bao gồm 4 kết quả có thể xảy ra. Sự kiện lần tung đầu tiên ra mặt ngửa của đồng xu là tập hợp $\{HH, HT\}$.

Cho một sự kiện $A \subseteq \Omega$, ta nói $A$ **xảy ra**, hoặc $A$ là **đúng**, nếu có một hiện tượng $\omega \in A$ **xảy ra**. Sự kiện **ngược lại** với $A$ là $A^c := \Omega - A := \{\omega \in \Omega : \omega \notin A\}$, tức là "không xảy ra $A$". Theo định nghĩa, rõ ràng $\Omega$ luôn luôn đúng, còn sự kiện rỗng $\emptyset \equiv \Omega^c$ luôn luôn sai. Cho thêm sự kiện $B$, ta có $A \cup B$ là sự kiện "$A$ **hoặc** $B$ ít nhất một việc xảy ra", còn $AB := A \cap B$ là sự kiện "$A$ **và** $B$ đồng thời xảy ra".

Chuỗi sự kiện $A_1, A_2, ...$ được gọi là **phân ly** nếu $A_i A_j \equiv \emptyset$ với mọi $i \neq j$. Khi đó nếu $A_1 \cup A_2 \cup \cdots \equiv C$ thì ta nói $A_1, A_2, ...$ là một cách **phân hoạch** sự kiện $C$.

## 2.2 Xác suất

Nếu một xạ ảnh $\mathbb{P}$ từ không gian các sự kiện $A \subseteq \Omega$ lên tập hợp số thực $\mathbb{R}$ thỏa mãn các điều kiện:

1. $\mathbb{P}(A) \geq 0 \quad \forall A$
2. $\mathbb{P}(\Omega) = 1$
3. Nếu chuỗi $A_1, A_2, ...$ phân hoạch $C$ thì $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots = \mathbb{P}(C)$

thì ta gọi $\mathbb{P}$ là một **phân phối xác suất** hoặc **độ đo xác suất**.

Có hai cách cắt nghĩa khái niệm xác suất là tần suất và niềm tin. Theo cách hiểu tần suất thì $\mathbb{P}(A)$ chính là tỷ lệ số lần sự kiện $A$ xảy ra nếu ta thực hiện thí nghiệm vô số lần. Còn theo cách hiểu niềm tin thì $\mathbb{P}(A)$ là thước đo mức độ mà một quan sát viên tin tưởng rằng hiện tượng $A$ sẽ xảy ra.

## 2.3 Phân phối

Có một số phân phối xác suất thông dụng.

**Definition 2.1** (Phân phối Bernoulli)**.** $X \sim \text{Bernoulli}(p)$ với $p \in (0,1)$ nếu:

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0).$$

**Definition 2.2** (Phân phối Binomial)**.** $X \sim \text{Binomial}(n, p)$ với $n \in \mathbb{Z}_+, p \in (0,1)$ nếu

$$X := \sum_{i=1}^{n} X_i$$

và iid $X_i \sim \text{Bernoulli}(p)$, $\forall i = 1, 2, ...$

# 3 Tóm tắt

Xác suất thống kê là nền tảng giúp ích cho việc phân tích dữ liệu, tổ chức thực hành thí nghiệm, chạy các mô hình giả lập, giải các bài toán tìm nghiệm tối ưu, nghiên cứu và ứng dụng học máy.

# Trích dẫn

Philippe Rigollet, Tyler Maunu, Jan Christian Huetter. 2022. "Fundamentals of Statistics." MITx. https://www.edx.org/course/fundamentals-of-statistics.

Wasserman, Larry. 2004. *All of Statistics : A Concise Course in Statistical Inference*. New York: Springer. https://archive.org/details/springer_10.1007-978-0-387-21736-9.