Bài 1 - Ôn tập Xác suất Cơ bản và Giới thiệu Thống kê Tính toán (Review of Basic Probability and Introduction to Computational Statistics)

Thống kê máy tính và ứng dụng (CLC)

Vũ Quốc Hoàng (vqhoang@fit.hcmus.edu.vn)
FIT - HCMUS

Ngày 17 tháng 1 năm 2022

Nội dung

Bài toán sinh nhật

Bài toán sinh nhật (birthday problem). Tính xác suất p của biến cố có ít nhất 2 người cùng sinh nhật (cùng ngày và tháng sinh) trong nhóm k người?

Giả sử: ngày sinh của mỗi người là một ngày ngẫu nhiên trong một năm gồm 365 ngày và "không liên quan nhau".

Ôn tập xác suất
Không gian mẫu và biến cố

- **Lý thuyết xác suất** (probability theory) là ngành Toán học giúp định lượng, tính toán và suy diễn trên các hiện tượng **ngẫu nhiên** (random) và/hoặc **không chắc chắn** (uncertain).
- **Thí nghiệm ngẫu nhiên** (random experiment) là các quá trình/hoạt động/thử nghiệm/công việc/thao tác không biết chắc chắn **kết quả** (outcome) nhưng xác định được tập tất cả các kết quả có thể.
- Tập tất cả các kết quả có thể của một thí nghiệm được gọi là **không gian mẫu** (sample space) của thí nghiệm, thường được kí hiệu là S hay Ω (omega).
- Nếu việc xảy ra hay không của một tình huống E được xác định hoàn toàn khi biết kết quả của thí nghiệm T thì E được gọi là **biến cố** (event) **liên quan** đến T. Biến cố được xác định bởi các **kết quả thuận lợi** cho nó

$$E = \{\omega \in \Omega : \omega \text{ làm cho } E \text{ xảy ra}\} \subset \Omega.$$

Ôn tập xác suất
Không gian mẫu và biến cố (tt)

"Lý thuyết biến cố" được hình thức hóa bằng "lý thuyết tập hợp". Xét thí nghiệm T với không gian mẫu Ω và các biến cố $E, F \subset \Omega$

- $\omega \in \Omega$: **biến cố sơ cấp** (elementary event),
- Ω: **biến cố chắc chắn** (certain event),
- $\varnothing$: **biến cố không thể** (impossible event),
- $E^c = \Omega \setminus E$: biến cố **đối** (complement) của E, biến cố "E không xảy ra",
- $E \cup F$: biến cố E **hoặc** (or) F, biến cố "E xảy ra hoặc F xảy ra",
- $E \cap F$: biến cố E **và** (and) F, biến cố "E xảy ra và F xảy ra",
- $E \setminus F$: biến cố E **không** (not) F, biến cố "E xảy ra nhưng F không xảy ra",
- $E \subset F$: E **kéo theo** (imply) F, E xảy ra thì F xảy ra,
- $E = F$: E **là** (is) F, E và F cùng xảy ra hoặc cùng không xảy ra,
- $E \cap F = \varnothing$: E, F **rời nhau** (disjoint) hay **xung khắc** (mutually exclusive), E và F không thể đồng thời xảy ra.

Ôn tập xác suất
Xác suất

Xét thí nghiệm T với không gian mẫu Ω, một hàm P gắn mỗi biến cố $E \subset \Omega$ với số thực $P(E)$ được gọi là một **độ đo xác suất** (probability measure) trên Ω nếu P thỏa mãn 3 tiên đề

1. Với mọi biến cố $E \subset \Omega, 0 \leq P(E) \leq 1$.
2. Với mọi dãy biến cố $E_1, E_2, ...$ đôi một xung khắc ($E_i \cap E_j = \varnothing, \forall i \neq j$):

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i),$$

   tức là $P(E_1 \cup E_2 \cup ...) = P(E_1) + P(E_2) + ...$
3. $P(\Omega) = 1$.

$P(E)$ được gọi là **xác suất** (probability) của E và là số đo khả năng xảy ra của biến cố E khi không biết kết quả của thí nghiệm T.

Ôn tập xác suất
Xác suất (tt)

Các tính chất cơ bản của xác suất (hệ quả từ 3 tiên đề)

1. $P(E^c) = 1 - P(E)$
2. $P(\varnothing) = 0$
3. Nếu $E_1 \subset E_2$ thì $P(E_1) \leq P(E_2)$ và $P(E_2 \setminus E_1) = P(E_2) - P(E_1)$
4. Nếu $E_1 \cap E_2 = \varnothing$ thì $P(E_1 \cup E_2) = P(E_1) + P(E_2)$
5. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$ (**addition law of probability**)
6. $P(E_1 \cup E_2 \cup E_3) =$
   $P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3)$
7. $P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i)$ (**union bound**)
8. $P\left(\bigcap_{i=1}^{\infty} E_i\right) \geq 1 - \sum_{i=1}^{\infty} P(E_i^c)$ (**Bonferroni inequality**)

Ôn tập xác suất
Mô hình xác suất đơn giản

Khi không gian mẫu hữu hạn, $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, độ đo xác suất được xác định bởi xác suất của các biến cố sơ cấp $p_i = P(\omega_i)$

- $p_i \geq 0, i = 1, ..., n$,
- $\sum_{i=1}^{n} p_i = 1$,
- $P(E) = \sum_{w_i \in E} p_i$ với mọi biến cố $E \subset \Omega$.

Khi không gian mẫu hữu hạn và các **kết quả đồng khả năng** (equiprobable outcomes), ta có **mô hình xác suất đơn giản** (simple/classical probability model)

- $p_i = \frac{1}{n}, i = 1, ..., n$,
- $P(E) = \frac{|E|}{|\Omega|}$ với mọi biến cố $E \subset \Omega$, ($|X|$ là số lượng phần tử của tập X)
- Xác suất là tỉ lệ và việc tính xác suất được đưa về việc **đếm** (counting).

Ôn tập xác suất
Phương pháp xấp xỉ xác suất bằng mô phỏng

Để xấp xỉ xác suất của một biến cố E liên quan đến thí nghiệm T, ta có thể dùng phương pháp thống kê như sau

- Thực hiện lặp lại N lần (độc lập) thí nghiệm T và đếm số lần biến cố E xảy ra, m. Khi đó $f(E) = \frac{m}{N}$ được gọi là tần suất của E.
- Khi N đủ lớn, ta có $f(E) \approx P(E)$.
- Việc thực hiện lặp lại nhiều lần thí nghiệm T có thể được **mô phỏng** (simulation) trên máy tính.

Bài toán sinh nhật
Tính toán chính xác

Không mất tính tổng quát, ta có thể gọi tập tất cả các ngày trong năm là

$$\mathcal{Y} = \{1, 2, ..., 365\}.$$

Không gian mẫu $\Omega = \{(x_1, x_2, ..., x_k) : x_i \in \mathcal{Y}, i = 1, ..., k\} = \mathcal{Y}^k$ có $|\Omega| = 365^k$.
Đặt các biến cố:

- A: "có ít nhất 2 người cùng sinh nhật",
- B: "không có người nào cùng sinh nhật".

Như vậy, $A = B^c$ và $B = \{\text{chỉnh hợp chọn } k \text{ của } \mathcal{Y}\}$ với $|B| = P_{365}^k$.
Dùng mô hình xác suất đơn giản ta có

$$p = P(A) = 1 - P(B) = 1 - \frac{P_{365}^k}{365^k} = 1 - \frac{365!}{(365-k)!365^k}.$$

Bài toán sinh nhật
Mô phỏng

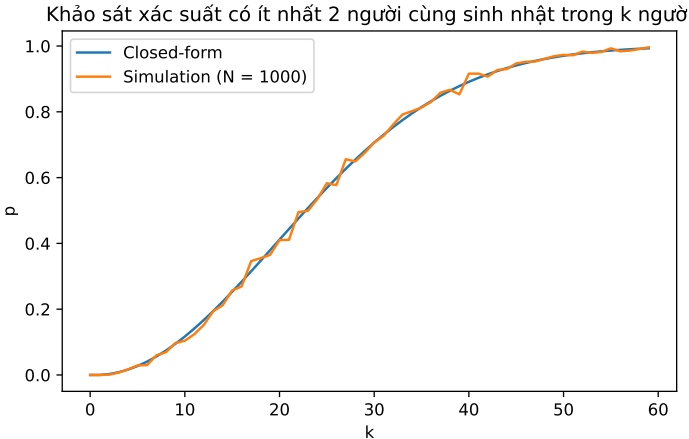
```
def birthday(k):
    return [random.randint(1, 365) for _ in range (k)]

def at_least_2(outcome):
    return len(set(outcome)) < len(outcome)

def relative_frequency(k, N, event):
    m = sum(event(birthday(k)) for _ in range (N))
    return m/N

relative_frequency(50 , 1000, at_least_2)
#0.973
```

Bài toán sinh nhật
Kết quả

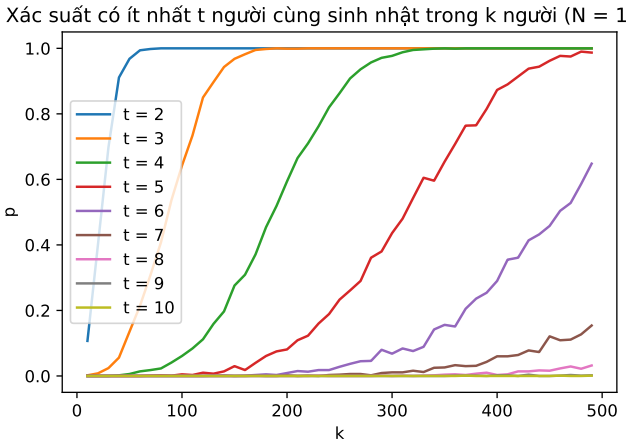


Bài toán sinh nhật
Mở rộng

Mở rộng bài toán. Tính xác suất p của biến cố có ít nhất t người cùng sinh nhật (cùng ngày và tháng sinh) trong nhóm k người?

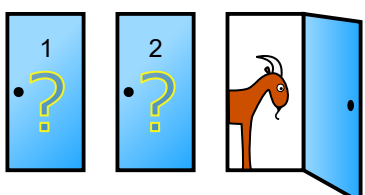
- Tính toán lý thuyết: được kết quả chính xác nhưng khó và không khả thi trong nhiều trường hợp.
- Mô phỏng máy tính: dễ dàng (chỉnh sửa mã) nhưng tốn thời gian (đợi mô phỏng), tài nguyên máy và chỉ được kết quả xấp xỉ.
  (Xem mã Python trong Notebook)

Bài toán sinh nhật
Mở rộng (tt)



Bài toán Monty Hall

Monty Hall problem. Có 3 cửa #1, #2, #3 chứa 1 chiếc xe và 2 con dê. Người chơi chọn 1 cửa (chẳng hạn #1). Người dẫn chương trình biết cửa nào có gì. Người dẫn chọn và mở cửa có dê trong 2 cửa còn lại (chẳng hạn #3). Người dẫn hỏi người chơi có muốn đổi lựa chọn không (vẫn giữ #1 hay chọn #2). Người chơi nên giữ hay đổi (để khả năng được xe cao hơn)?



(https://en.wikipedia.org/wiki/Monty_Hall_problem)

Ôn tập xác suất có điều kiện
Xác suất có điều kiện

Cần điều chỉnh, cập nhật xác suất (khả năng xảy ra) của các biến cố liên quan đến thí nghiệm T khi có thêm thông tin về T

- Thông tin về T được thể hiện bằng việc biết (các) biến cố nào đó đã xảy ra.

Xác suất của biến cố A khi biết biến cố B đã xảy ra được gọi là **xác suất có điều kiện** (conditional probability) của A khi biết B xảy ra, kí hiệu là $P(A|B)$ và được tính bằng định nghĩa

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ (với } P(B) > 0).$$

- $A \cap B$ là "A trong B",
- Chia cho $P(B)$ giúp chuẩn hóa xác suất,
- $P(.|B)$ có thể hiểu là xác suất "tính trong không gian mẫu mới" B và là một độ đo xác suất hợp lệ.

Ôn tập xác suất có điều kiện
Công thức nhân xác suất

Công thức nhân xác suất (multiplication rule)

$$P(A \cap B) = P(B)P(A|B) \text{ (khi } P(B) > 0),$$

$$P(A \cap B) = P(A)P(B|A) \text{ (khi } P(A) > 0).$$

Trong nhiều trường hợp, xác suất có điều kiện $P(A|B)$ dễ tính hơn xác suất $P(A \cap B)$.

Công thức nhân tổng quát. Giả sử có n biến cố $A_1, ..., A_n$ với $P(A_1 \cap ... \cap A_n) > 0$, ta có

$$P(A_1 \cap ... \cap A_n) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1, A_2) \times ... \times P(A_n|A_1, A_2, ..., A_{n-1}).$$

Ôn tập xác suất có điều kiện
Công thức xác suất toàn phần

$B_1, B_2, ..., B_n$ được gọi là một **họ đầy đủ** các biến cố (hay một **phân hoạch**) của Ω nếu

1. $B_i \cap B_j = \varnothing, \forall i \neq j$,
2. $\Omega = B_1 \cup B_2 \cup ... \cup B_n$.

Công thức xác suất toàn phần (law of total probability). Giả sử có phân hoạch $B_1, B_2, ..., B_n$ với $P(B_i) > 0$ $(i = 1, ..., n)$, khi đó

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

Đặc biệt

$$P(A) = P(B)P(A|B) + P(B^c)P(A|B^c).$$

Ôn tập xác suất có điều kiện
Định lý Bayes

Định lý Bayes (Bayes' theorem, Bayes's rule). Giả sử có phân hoạch $B_1, B_2, ..., B_n$ với $P(B_i) > 0$ $(i = 1, ..., n)$ và biến cố A với $P(A) > 0$, khi đó, với mọi $i = 1, ..., n$ ta có

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

- $P(B_i)$: **xác suất tiên nghiệm** (prior probability) của B_i,
- $P(B_i|A)$: **xác suất hậu nghiệm** (posterior probability) của B_i khi biết A,
- $P(A|B_i)$: **xác suất hợp lý** (likelihood) của A theo B_i,
- Lưu ý, $P(A)$ không phụ thuộc vào B_i nên ta có

$$P(B_i|A) \propto P(B_i)P(A|B_i) \text{ (kí hiệu } \propto \text{ là "tỉ lệ với").}$$

Ôn tập xác suất có điều kiện
Các biến cố độc lập

Hai biến cố $\{A, B\}$ được gọi là **độc lập** (independent, statistically independent, stochastically independent) với nhau nếu

$$P(A \cap B) = P(A) \times P(B).$$

Một cách tương đương: $P(A|B) = P(A)$ $(P(B) > 0)$ hay $P(B|A) = P(B)$ $(P(A) > 0)$.

Mệnh đề. Nếu $\{A, B\}$ độc lập thì các cặp biến cố $\{A^c, B\}, \{A, B^c\}, \{A^c, B^c\}$ cũng độc lập.

Ba biến cố $\{A, B, C\}$ được gọi là độc lập (với nhau) nếu từng đôi $\{A, B\}, \{A, C\}, \{B, C\}$ độc lập và

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C).$$

Ôn tập xác suất có điều kiện
Mô hình xác suất "lặp lại thí nghiệm độc lập"

Họ các biến cố $\{A_1, A_2, ...\}$ được gọi là độc lập nếu với mọi tập con khác rỗng và hữu hạn $\{B_1, B_2, ..., B_k\}$ của họ ta có

$$P\left(\bigcap_{i=1}^{k} B_i\right) = \prod_{i=1}^{k} P(B_i).$$

- Chiều ngược: dùng để cho thấy (hay kiểm tra, chứng minh) tính độc lập,
- Chiều xuôi: dùng giả thuyết về tính độc lập để tính toán xác suất đơn giản.

Mô hình xác suất "lặp lại thí nghiệm độc lập": thực hiện lặp lại thí nghiệm T nhiều lần một cách độc lập, gọi A_i là biến cố "liên quan đến lần thực hiện thứ i" thì

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Ôn tập xác suất có điều kiện
Độc lập có điều kiện

Cho biến cố C với $P(C) > 0$, họ các biến cố $\{A_1, A_2, ...\}$ được gọi là **độc lập có điều kiện** (conditionally independent) khi biết C nếu với mọi tập con khác rỗng và hữu hạn $\{B_1, B_2, ..., B_k\}$ của họ ta có

$$P\left(\bigcap_{i=1}^{k} B_i | C\right) = \prod_{i=1}^{k} P(B_i|C).$$

Hai biến cố $\{A, B\}$ được gọi là độc lập có điều kiện khi biết C nếu

$$P(A \cap B|C) = P(A|C) \times P(B|C).$$

Một cách tương đương:

$$P(A|B, C) = P(A|C) \ (P(B|C) > 0) \text{ hay } P(B|A, C) = P(B|C) \ (P(A|C) > 0).$$

Ôn tập xác suất có điều kiện
Phương pháp xấp xỉ xác suất bằng mô phỏng

Để xấp xỉ xác suất có điều kiện của một biến cố B khi biết biến cố A đã xảy ra trong thí nghiệm T, ta có thể dùng phương pháp thống kê như sau

- Thực hiện lặp lại N lần (độc lập) thí nghiệm T, đếm số lần biến cố A xảy ra, m, và trong các lần A xảy ra thì đếm số lần B cũng xảy ra, p. Khi đó $f(B|A) = \frac{p}{m}$ được gọi là tần suất của B trên A.
- Khi N đủ lớn, ta có $f(B|A) \approx P(B|A)$.
- Việc thực hiện lặp lại nhiều lần thí nghiệm T có thể được **mô phỏng** (simulation) trên máy tính.

Bài toán Monty Hall
Tính toán chính xác

Do "tính đối xứng" nên ta có thể giả sử kịch bản như mô tả (người chơi chọn cửa #1, người dẫn mở cửa #3).

Đặt A_i là biến cố "xe được đặt ở cửa #i" ($1 \leq i \leq 3$) và B_j là biến cố "người dẫn mở cửa #j" ($1 \leq j \leq 3$). Từ bài toán và kịch bản đã cho (người chơi chọn cửa #1), ta có

- $P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$.
- $P(B_3|A_1) = \frac{1}{2}$ (người chơi đã chọn cửa #1, xe cũng được đặt ở cửa #1 nên người dẫn có thể mở 1 trong 2 cửa #2, #3).
- $P(B_3|A_2) = 1$ (người chơi đã chọn cửa #1, xe được đặt ở cửa #2 nên người dẫn chỉ có thể mở cửa #3).
- $P(B_3|A_3) = 0$ (xe được đặt ở cửa #3 nên người dẫn không được mở cửa #3).

Bài toán Monty Hall
Tính toán chính xác (tt)

Dùng công thức Bayes, ta có xác suất người chơi được xe khi không đổi cửa là

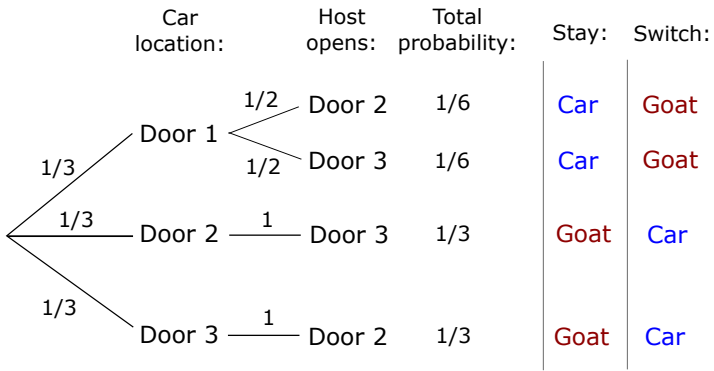
$$\begin{aligned} P(A_1|B_3) &= \frac{P(A_1)P(B_3|A_1)}{P(A_1)P(B_3|A_1) + P(A_2)P(B_3|A_2) + P(A_3)P(B_3|A_3)} \\ &= \frac{\frac{1}{3}\frac{1}{2}}{\frac{1}{3}\frac{1}{2} + \frac{1}{3}1 + \frac{1}{3}0} = \frac{1}{3} \end{aligned}$$

và xác suất người chơi được xe khi đổi cửa là

$$\begin{aligned} P(A_2|B_3) &= \frac{P(A_2)P(B_3|A_2)}{P(A_1)P(B_3|A_1) + P(A_2)P(B_3|A_2) + P(A_3)P(B_3|A_3)} \\ &= \frac{\frac{1}{3}1}{\frac{1}{3}\frac{1}{2} + \frac{1}{3}1 + \frac{1}{3}0} = \frac{2}{3}. \end{aligned}$$

Vậy người chơi nên chọn đổi cửa (xác suất được xe cao gấp đôi so với không đổi cửa).

Bài toán Monty Hall
Sơ đồ cây (tree diagram)



(https://en.wikipedia.org/wiki/Monty_Hall_problem#Conditional_probability_by_direct_calculation)

Bài toán Monty Hall
Mô phỏng

```
def Monty_Hall(doors={"#1", "#2", "#3"}):
    car_door = random.choice(list(doors))
    choice_door = random.choice(list(doors))
    open_door = random.choice(list(doors - {choice_door,
       car_door}))
    op_door = random.choice(list(doors - {choice_door,
       open_door}))

    return car_door == choice_door, car_door == op_door
```

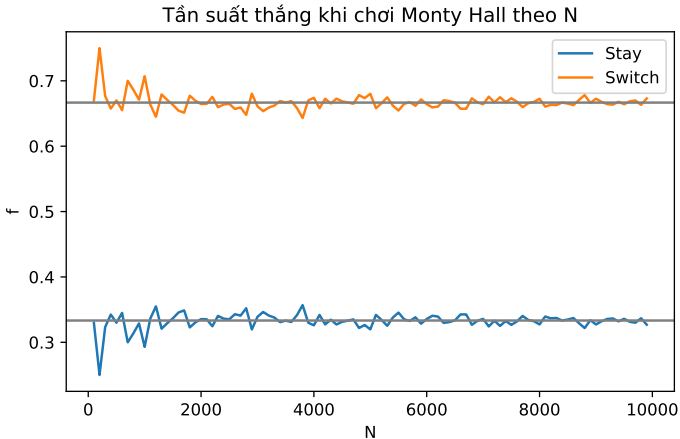
Ví dụ minh họa - Bài toán Monty Hall
Mô phỏng (tt)

```
N = 10000
results = [Monty_Hall() for _ in range(N)]

stay_freq = sum([v for v, _ in results])/N
switch_freq = sum([v for _, v in results])/N

print(stay_freq, switch_freq)
#0.3317 0.6683
```

Ví dụ minh họa - Bài toán Monty Hall
Mô phỏng (tt)



Tài liệu tham khảo

Chapter 1-2. Morris H. DeGroot, Mark J. Schervish. *Probability and Statistics*. Addison-Wesley, 2012.

Chapter 1-2. H. Pishro-Nik. *"Introduction to probability, statistics, and random processes"*, available at `https://www.probabilitycourse.com`. Kappa Research LLC, 2014.