Buổi 2. Các khái niệm xác suất cơ bản

Ho Huu Binh

2025-10-19

1 Một số khái niệm cơ bản về xác suất

Xác suất đặt nền tảng cho suy luận thống kê. Trong ghi chí này, ta sẽ được cung cấp một cái nhìn tổng quan ngắn gọn về các khái niệm xác suất cần thiết để hiểu các chủ đề được trình bày trong những chương tiếp theo.

1.1 Giới thiệu

Hiểu biết về các khái niệm xác suất cơ bản trong thống kê là rất quan trọng, bởi vì xác suất tạo nên nền tảng cho việc phân tích sự bất định và đưa ra các quyết định có cơ sở.

1.2 Hai quan điểm về xác suất: Khách quan và Chủ quan

Xác suất có thể được phân thành ba loại chính: xác suất cổ điển, xác suất tương đối (thực nghiệm),và xác suất chủ quan. Mỗi loại có đặc điểm và ứng dụng riêng biệt.

Xác suất cổ điển (hay Xác suất lý thuyết)

- Dựa trên giả định rằng *tất cả các kết quả đều có khả năng xảy ra như nhau.*
- Được sử dụng trong các tình huống mà xác suất có thể được xác định *mà không cần thí nghiệm.*
- **Công thức:** Gọi A là biến cố quan tâm.

$$P(A) = \frac{\text{Số kết quả mong muốn}}{\text{Tổng số kết quả có thể xảy ra}}$$

Ví dụ: Xác suất để gieo được mặt 3 khi tung một con xúc xắc công bằng sáu mặt là $1/6$, vì có sáu kết quả có khả năng xảy ra như nhau.

Xác suất tương đối (hay Xác suất thực nghiệm)

- Dựa trên *dữ liệu quan sát được* từ các thí nghiệm hoặc hiện tượng trong thế giới thực.
- Được *tính toán dựa trên tần suất* xuất hiện của một biến cố trong mẫu quan sát.
- **Công thức:**

$$P(A) = \frac{\text{Số lần biến cố A xảy ra}}{\text{Tổng số lần thử nghiệm}}$$

Ví dụ: Nếu tung một đồng xu 100 lần và xuất hiện mặt ngửa 52 lần, thì xác suất thực nghiệm của việc xuất hiện mặt ngửa là $52/100 = 0.52$.

Xác suất chủ quan

- Dựa trên *phán đoán cá nhân, trực giác hoặc kinh nghiệm*, thay vì các phép tính chính thức.
- Thường được sử dụng trong việc *ra quyết định khi dữ liệu quá khứ không có sẵn.*

Ví dụ: Một người hâm mộ bóng đá nhiệt thành nghĩ rằng đội **Manchester United** có **70% khả năng** chiến thắng trong trận đấu bóng tiếp theo của họ.

Phương pháp Bayes

- Các phương pháp Bayes trong thống kê cung cấp *một khuôn khổ để cập nhật niềm tin* về một tham số hoặc giả thuyết khi *dữ liệu mới xuất hiện.*
- Không giống như các phương pháp tần suất (frequentist), vốn dựa trên các xác suất cố định, *thống kê Bayes coi xác suất là mức độ tin tưởng* có thể được *điều chỉnh và hoàn thiện dần theo thời gian.*

Ví dụ: Một người hâm mộ bóng đá nghĩ rằng đội **Manchester United** có **70% khả năng** chiến thắng trong trận đấu bóng tiếp theo. Khi *hiệp một* kết thúc, người hâm mộ này sẽ *cập nhật lại xác suất* chiến thắng của đội MU trên diễn biến trận đấu.

1.3 Các tính chất sơ cấp của xác suất

Cách tiếp cận tiên đề đối với xác suất, được *Andrey Kolmogorov* chính thức hóa (formalize) vào năm *1933*, đã cung cấp một *nền tảng toán học chặt chẽ* cho lý thuyết xác suất. Phương pháp này *định nghĩa xác suất* như một *hàm* thỏa mãn *ba tiên đề cơ bản*, đảm bảo tính nhất quán giữa các mô hình xác suất khác nhau.

Ba tiên đề của Kolmogorov

- **Tính không âm:**

$$P(A) \geq 0$$

Xác suất của bất kỳ biến cố A nào đó đều phải *không âm*

- **Chuẩn hóa (Quy tắc tổng xác suất):**

$$P(\Omega) = 1$$

Trong đó Ω là *không gian mẫu* – tập hợp tất cả các kết quả có thể xảy ra. Quy tắc này có nghĩa là *tổng xác suất của toàn bộ không gian mẫu bằng 1*, tức là chắc chắn có một kết quả nào đó xảy ra.

- **Tính cộng (đối với các biến cố xung khắc lẫn nhau):**

$$P(A \cup B) = P(A) + P(B), \quad \text{nếu } A \cap B = \emptyset$$

Nếu hai biến cố A và B *xung khắc* (không thể xảy ra đồng thời), thì xác suất để A hoặc B xảy ra bằng *tổng xác suất riêng của từng biến cố.*

1.4 Tính xác suất của một biến cố

Xác suất có điều kiện

Xác suất có điều kiện là xác suất để một biến cố xảy ra *khi đã biết rằng một biến cố khác đã xảy ra.* Nó giúp *định lượng mức độ thay đổi của khả năng xảy ra* của một biến cố khi có *thông tin bổ sung.*

Công thức xác suất có điều kiện:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Trong đó:

- $P(A|B)$: xác suất để (A) xảy ra *khi đã biết rằng* B xảy ra.
- $P(A \cap B)$: xác suất để *cả hai biến cố* A *và* B cùng xảy ra.
- $P(B)$: là *xác suất để biến cố B xảy ra* (giả sử $P(B) \neq 0$).

Nếu *hai biến cố độc lập*, thì việc biết rằng **B đã xảy ra** không làm thay đổi xác suất xảy ra của A. Ngược lại, nếu *hai biến cố phụ thuộc*, thì việc biết rằng **B đã xảy ra** có ảnh hưởng xác suất xảy ra của A.

Xác suất đồng thời

Xác suất đồng thời là *khả năng hai biến cố cùng xảy ra* tại một thời điểm, và được biểu diễn bằng *ký hiệu giao nhau* ($\cap$) trong ký pháp xác suất. Phép giao ($\cap$) biểu thị "và" (AND) - tức là cả hai biến cố A và B cùng xảy ra. Xác suất đồng thời của A và B được viết là:

$$P(A \cap B)$$

Trong một số trường hợp, dấu phẩy cũng được dùng thay cho ký hiệu giao, chẳng hạn $ P(A, B)$.

Các loại biến cố, quy tắc nhân và quy tắc cộng

Biến cố độc lập (Independent Events): Nếu hai biến cố *không ảnh hưởng lẫn nhau*, thì xác suất đồng thời của chúng *bằng tích* của các xác suất riêng lẻ.

Biến cố phụ thuộc (Dependent Events): Nếu một biến cố *ảnh hưởng đến biến cố kia*, thì xác suất đồng thời được *tính có điều kiện* (tức là sử dụng xác suất có điều kiện).

Quy tắc nhân (The Multiplication Rule):

- Quy tắc nhân trong xác suất giúp *xác định khả năng hai biến cố cùng xảy ra.* Cách áp dụng phụ thuộc vào việc hai biến cố là *độc lập* hay *phụ thuộc*.
- Nếu hai biến cố A và B *độc lập*, tức là không ảnh hưởng lẫn nhau, thì xác suất đồng thời của chúng là:

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

Ví dụ: Tung xúc xắc và tung đồng xu

- Xác suất tung được mặt 3: $P(3) = \frac{1}{6}$
- Xác suất tung được mặt ngửa: $P(\text{Heads}) = \frac{1}{2}$
- Xác suất đồng thời:

$$P(3 \cap \text{Heads}) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

Ta có thể thực hiện mô phỏng trong R như sau:

```r
## Simulation experiment: coin tossing
n <- 100000

die_coin <- data.frame(
  die = sample(1:6, n, replace = TRUE),
  coin = sample(c('H', 'T'), n, replace = TRUE)
)

whr <- with(die_coin, die == 3 & coin == 'H')
```

```r
proportions(table(whr))
```

```
## whr
##   FALSE    TRUE
## 0.9173 0.0827
```

- Nếu biến cố B phụ thuộc vào biến cố A, thì xác suất để *cả hai cùng xảy ra* được tính theo công thức:

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

Ví dụ: Rút hai lá bài đỏ từ một bộ bài mà không hoàn lại

- Xác suất rút được một lá đỏ ở lần đầu:

$$P(R_1) = \frac{26}{52}$$

- Xác suất rút được một lá đỏ ở lần thứ hai, khi biết rằng lá đầu tiên đã là đỏ:

$$P(R_2|R_1) = \frac{25}{51}$$

- Xác suất đồng thời (cả hai lá đều đỏ):

$$P(R_1 \cap R_2) = \frac{26}{52} \times \frac{25}{51} = \frac{650}{2652} \approx 0.245$$

Ta thực hiện mô phỏng trong R như sau:

```r
## Building deck
deck <- expand.grid(
  rank = c('A', 2:10, 'J', 'Q', 'K'),
  suit = c('Club', 'Diamond', 'Heart', 'Spade')
)

n <- 100000

## Draw two cards with no replacement
two_cards <- sapply(1:n, FUN = function(x) sample(1:52, 2, replace = FALSE))

## Kiểm tra xem cả hai lá có đều là đỏ (Diamond hoặc Heart) không
whr <- deck[two_cards[1, ], 'suit'] %in% c('Diamond', 'Heart') &
       deck[two_cards[2, ], 'suit'] %in% c('Diamond', 'Heart')

proportions(table(whr))
```

```
## whr
##   FALSE    TRUE
## 0.75587 0.24413
```

Ta có kết quả dự kiến:

```
FALSE   TRUE
0.75463 0.24537
```

Quy tắc cộng (The Addition Rule):

Quy tắc cộng trong xác suất giúp xác định khả năng xảy ra của ít nhất một trong hai biến cố. Cách tính phụ thuộc vào việc hai biến cố có xung khắc nhau hay không.

- **Đối với các sự kiện loại trừ lẫn nhau** (các sự kiện không thể xảy ra cùng lúc):
  $P(A \cup B) = P(A) + P(B)$.

Ví dụ: Nếu tung một con xúc xắc, xác suất tung được số 2 hoặc số 5 là $P(2) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$.

- **Đối với các sự kiện không loại trừ lẫn nhau** (các sự kiện có thể xảy ra cùng lúc):
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Ví dụ: Nếu rút một lá bài từ bộ bài, xác suất rút được lá bài là quân Vua hoặc chất Cơ là
$P(\text{quân Vua}) + P(\text{chất Cơ}) - P(\text{quân Vua chất Cơ}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$.

```r
## Simulation die roll and card deck
n <- 100
# đây là một vector

die_rolls <- sample(1:6, size=n, replace=TRUE)

whr <- die_rolls %in% c(2, 5)

proportions(table(whr))
```

```
## whr
## FALSE  TRUE
##  0.75  0.25
```

```r
# or
table(whr) / sum(table(whr))
```

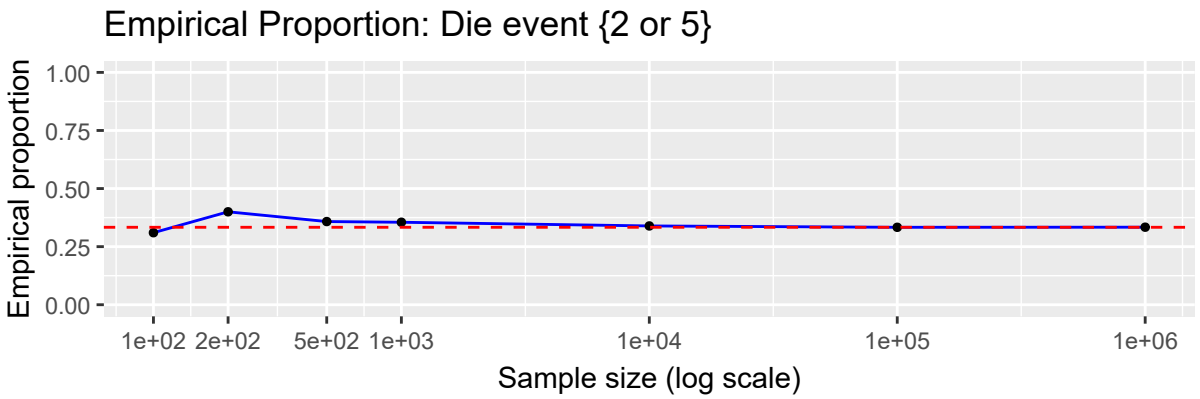
```
## whr
## FALSE  TRUE
##  0.75  0.25
```

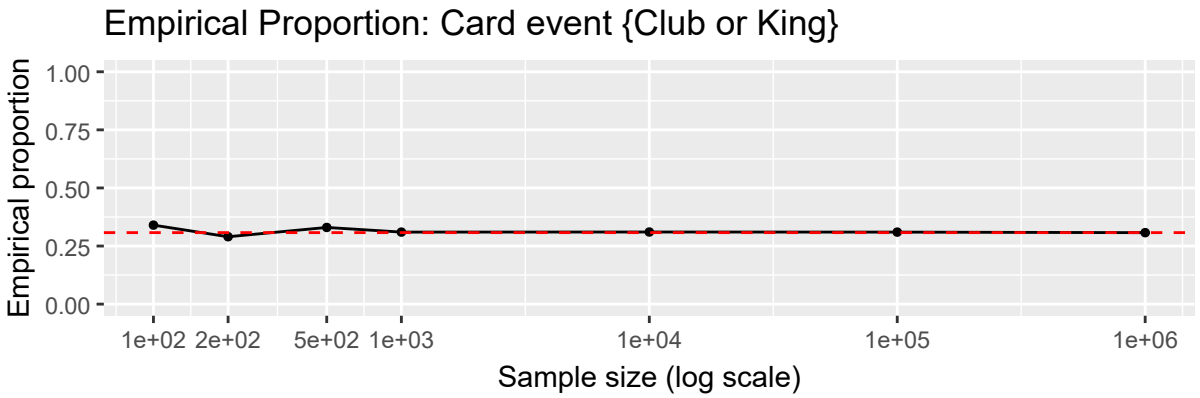
```r
# or
mean(whr) # empirical p = fraction of TRUE
```

```
## [1] 0.25
```

```r
compute_prop <- function(n) {
  # Die rolls
  die_rolls <- sample(1:6, size = n, replace = TRUE)
  whr1 <- die_rolls %in% c(2, 5)
  emp_prop1 <- mean(whr1)

  # Card draws (with replacement)
  deck <- expand.grid(
    rank = c('A', 2:10, 'J', 'Q', 'K'),
    suit = c('Club', 'Diamond', 'Heart', 'Spade'),
    stringsAsFactors = FALSE
  )

  one_card <- sample(1:52, size = n, replace = TRUE)

  suit_is_club <- deck[one_card, 'suit'] == 'Club'
  rank_is_king <- deck[one_card, 'rank'] == 'K'
```

```r
  whr2 <- suit_is_club | rank_is_king
  emp_prop2 <- mean(whr2)

  return(c(die = emp_prop1, card = emp_prop2))
}

sample_size <- c(100, 200, 500, 1000, 10000, 100000, 1e6)

res_mat <- sapply(sample_size, compute_prop)

df_die  <- data.frame(n = sample_size, prop = as.numeric(res_mat["die", ]))
df_card <- data.frame(n = sample_size, prop = as.numeric(res_mat["card", ]))

# Theoretical probs
p_die_true  <- 2/6          # = 1/3
p_card_true <- 4/13         # P(Club | King) = 1/4 + 1/13 - 1/52 = 4/13

p1 <- ggplot(df_die, aes(x=n, y=prop)) +
  geom_line(color='blue', linewidth=0.5) +
  geom_point(size=1) +
  geom_hline(yintercept=1/3, linetype='dashed', color='red') +
  scale_x_continuous(trans = "log10", breaks = sample_size) +
  ylim(0, 1) +
  labs(
    title = "Empirical Proportion: Die event {2 or 5}",
    x = "Sample size (log scale)",
    y = "Empirical proportion"
  )

p2 <- ggplot(df_card, aes(x = n, y = prop)) +
  geom_line(linewidth = 0.5) +
  geom_point(size = 1) +
  geom_hline(yintercept = p_card_true, linetype = 'dashed', color = 'red') +
  scale_x_continuous(trans = "log10", breaks = sample_size) +
  ylim(0, 1) +
  labs(
    title = "Empirical Proportion: Card event {Club or King}",
    x = "Sample size (log scale)",
    y = "Empirical proportion"
  )

p1 / p2
```





Sự kiện độc lập

Hai sự kiện A và B là độc lập nếu sự xảy ra của một sự kiện không ảnh hưởng đến xác suất xảy ra của sự kiện kia. Về mặt toán học, điều này được biểu diễn là:
$P(A \cap B) = P(A) \times P(B)$.

Điều này có nghĩa là xác suất cả hai sự kiện xảy ra cùng lúc đơn giản là tích của các xác suất riêng lẻ của chúng.

- **Xác suất có điều kiện** đo lường khả năng xảy ra của một sự kiện khi biết rằng một sự kiện khác đã xảy ra. Nó được định nghĩa là: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- Tuy nhiên, nếu A và B độc lập, thì: $P(A|B) = P(A)$. Điều này có nghĩa là biết B đã xảy ra không thay đổi xác suất xảy ra của A.

Sự kiện bù trừ

Sự kiện bù trừ là các cặp sự kiện mà một sự kiện xảy ra khi và chỉ khi sự kiện kia không xảy ra. Nếu một sự kiện A xảy ra, thì sự kiện bù trừ A^c hoặc A' hoặc $\bar{A}$ đại diện cho kịch bản mà A không xảy ra.

- **Loại trừ lẫn nhau**: A và A^c không thể xảy ra cùng lúc.
- **Toàn kiệt**: Cùng nhau, A và A^c bao quát tất cả các kết quả có thể.
- **Quy tắc xác suất**: Tổng xác suất của chúng luôn là 1: $P(A) + P(A^c) = 1$.

Phân hoạch một sự kiện thành các sự kiện con loại trừ lẫn nhau với sự kiện bù trừ và giao của các sự kiện

- **Phân hoạch** $P(B)$ **với** A:

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

- **Phân hoạch** $P(B)$ **với** A **và** C:

$$P(B) = P((B \cap A) \cup (B \cap A^c \cap C) \cup (B \cap A^c \cap C^c)) \quad = P(B \cap A) + P(B \cap A^c \cap C) + P(B \cap A^c \cap C^c).$$

Ví dụ: Bài toán sinh nhật

Đôi khi, việc tính $P(A)$ trực tiếp khó khăn, nên dễ hơn khi tính $P(A^c)$ trước, sau đó sử dụng:

$$P(A) = 1 - P(A^c).$$

- Xác suất để trong một tập hợp gồm n người được chọn ngẫu nhiên, ít nhất hai người có cùng ngày sinh.
  Dễ hơn khi tính phần bù trừ, tức xác suất mà không ai có cùng ngày sinh!
- A là sự kiện ít nhất hai người có cùng ngày sinh.
- $P(A) = 1 - P(A^c)$. Dễ tính $P(A^c)$, xác suất mà không ai chia sẻ cùng ngày sinh.

$$P(A^c) = \left(\frac{365}{365}\frac{364}{365} \ldots \frac{365-(n-1)}{365}\right)$$

```
n <- 23

prop_zero <- prod(365:(365 - (n-1))) / (365^n)

1-prop_zero
```

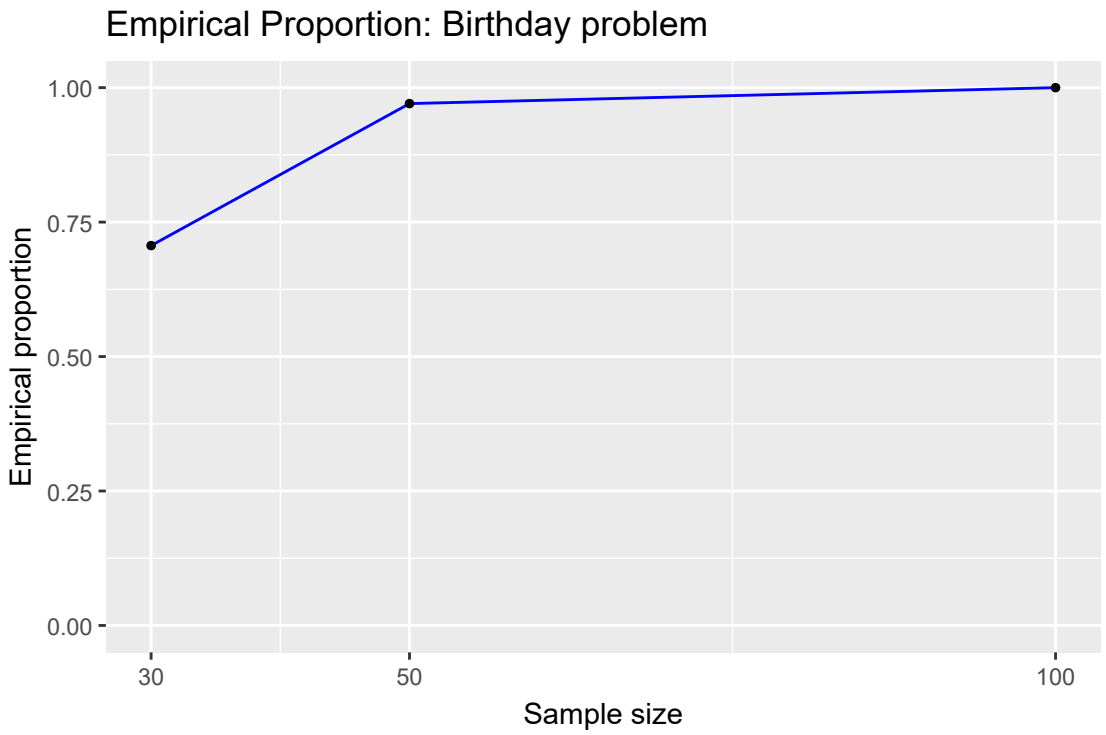
```
## [1] 0.5072972
```

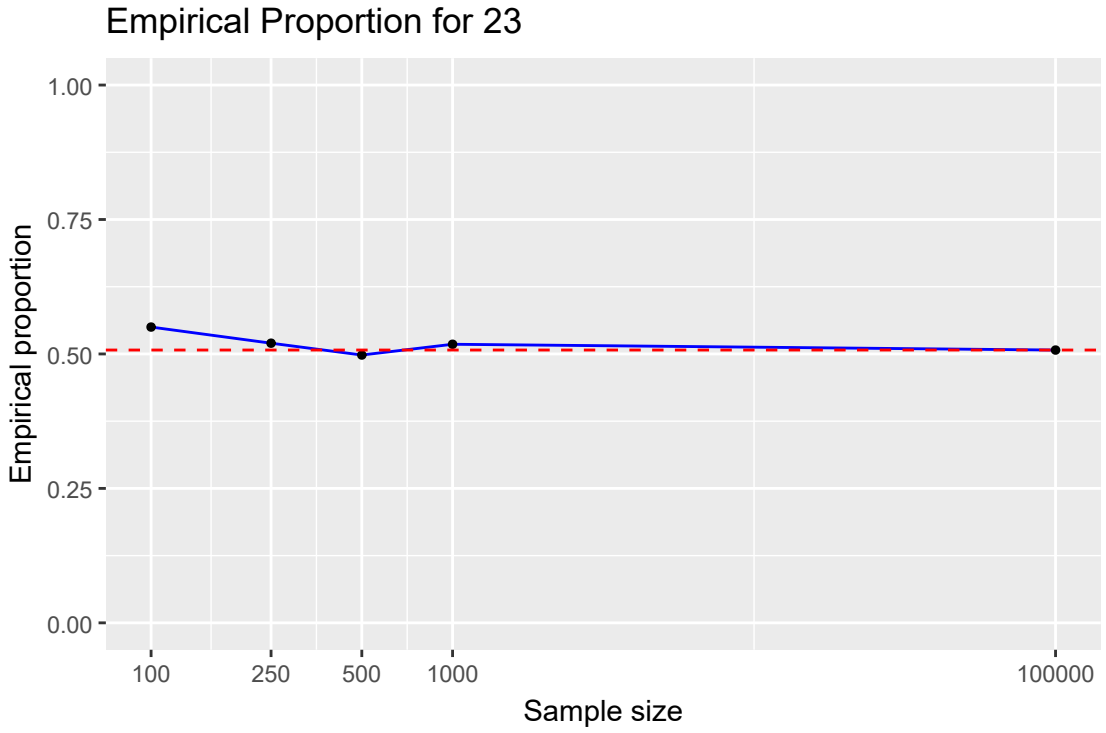
Bài tập 1: Dự đoán xem với n **tăng thì xác suất này sẽ thế nào?**

Hãy tính thử:

(a) Khi n tăng tới 100 thì xác suất cần tìm thế nào?

(b) Hãy thử mô phỏng các sinh nhật ngẫu nhiên cho một nhóm n người với n bất kỳ và tính xác suất thực nghiệm, so sánh với xác suất chính xác.





Xác suất biên

Xác suất biên đề cập đến xác suất xảy ra của một sự kiện đơn lẻ, mà không xem xét bất kỳ sự kiện liên quan nào khác. Nó được suy ra từ phân phối xác suất liên hợp bằng cách tổng (hoặc tích phân) trên tất cả các giá trị có thể của biến kia.

- Đối với hai sự kiện A và B, xác suất biên của A được tìm bằng cách tổng trên tất cả các giá trị có thể của B:

$$P(A) = \sum_B P(A, B).$$

- Đối với biến liên tục, điều này trở thành tích phân:

$$P(A) = \int P(A, B)\, dB.$$

Ví dụ: Xúc xắc đỏ và xanh

Giả sử có xúc xắc đỏ R và xúc xắc xanh G. Tìm $P(R = 2)$.

$$P(R = 2) = \sum_{i=1}^{6} P(R = 2, G = i)$$

$$\begin{aligned}
&= P(R = 2, G = 1) + P(R = 2, G = 2) + P(R = 2, G = 3) \\
&\quad + P(R = 2, G = 4) + P(R = 2, G = 5) + P(R = 2, G = 6) \\
&= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\
&= \frac{6}{36} = \frac{1}{6}.
\end{aligned}$$

```
library(MASS)
joint_prob <- matrix(1/36, nrow=6, ncol=6)

fractions(joint_prob)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] 1/36 1/36 1/36 1/36 1/36 1/36
## [2,] 1/36 1/36 1/36 1/36 1/36 1/36
## [3,] 1/36 1/36 1/36 1/36 1/36 1/36
## [4,] 1/36 1/36 1/36 1/36 1/36 1/36
## [5,] 1/36 1/36 1/36 1/36 1/36 1/36
## [6,] 1/36 1/36 1/36 1/36 1/36 1/36
```

```
dimnames(joint_prob) <- list(paste('R', 1:6, sep='_'),
                             paste('G', 1:6, sep='_'))

fractions(addmargins(joint_prob))
```

```
##      G_1  G_2  G_3  G_4  G_5  G_6  Sum
## R_1 1/36 1/36 1/36 1/36 1/36 1/36  1/6
## R_2 1/36 1/36 1/36 1/36 1/36 1/36  1/6
## R_3 1/36 1/36 1/36 1/36 1/36 1/36  1/6
## R_4 1/36 1/36 1/36 1/36 1/36 1/36  1/6
## R_5 1/36 1/36 1/36 1/36 1/36 1/36  1/6
## R_6 1/36 1/36 1/36 1/36 1/36 1/36  1/6
## Sum  1/6  1/6  1/6  1/6  1/6  1/6    1
```

```
fractions(sum(joint_prob['R_2', ]))
```

```
## [1] 1/6
```

1.5 Phân phối xác suất

Xác suất của biến rời rạc

Phân phối xác suất là xương sống của phân tích thống kê. Việc nghiên cứu chúng rất quan trọng vì chúng cung cấp một cách có cấu trúc để mô tả sự không chắc chắn và biến thiên trong dữ liệu.

Biến ngẫu nhiên rời rạc

Một biến ngẫu nhiên rời rạc là một loại biến ngẫu nhiên có thể nhận một số lượng đếm được các giá trị phân biệt. Các giá trị này thường là số nguyên và phát sinh từ các quá trình mà kết quả là phân biệt và riêng biệt.

Về mặt hình thức, một biến ngẫu nhiên rời rạc X là một hàm ánh xạ các kết quả của không gian mẫu sang số thực, trong đó mỗi giá trị có thể có một xác suất liên kết. $P(X = x) = p(x) = f(x)$.

Các ví dụ phổ biến bao gồm:

- Số mặt sấp trong ba lần tung một đồng xu.
- Số học sinh vắng mặt trong một lớp học vào một ngày nhất định.
- Số lần một con xúc xắc ra mặt sáu trong mười lần tung.

Phân phối xác suất của một biến ngẫu nhiên rời rạc là một phương pháp trình bày tất cả các giá trị có thể mà biến có thể nhận, cùng với các xác suất tương ứng của chúng. Theo ba tiên đề của Kolmogorov

1. $0 \leq P(X = x) \leq 1$ cho mọi x.
2. $$\sum_{\text{tất cả } x} P(X = x) = \sum_{\text{tất cả } x} p(x) = 1.$$
3. $P(X = x_i \text{ hoặc } X = x_j) = P(X = x_i) + P(X = x_j)$ với $i \neq j$.

Phân phối xác suất có thể được biểu diễn dưới nhiều dạng khác nhau, bao gồm:

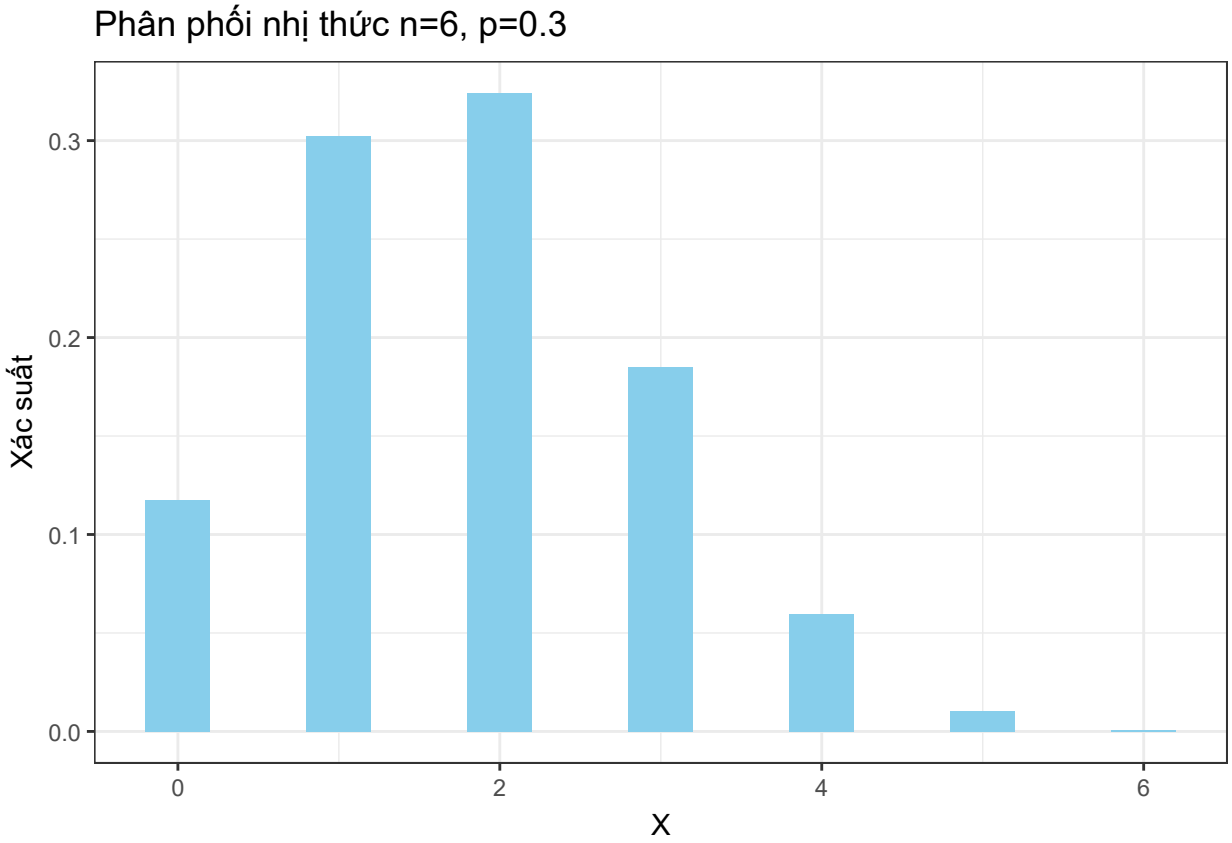
- **Bảng**: Liệt kê từng giá trị có thể và xác suất của nó một cách rõ ràng.

```
# Phân phối nhị thức
n <- 6
x <- 0:n
p.dist <- rbind(x, dbinom(x, n, p = 0.3))
dimnames(p.dist) <- list(c("X", "p(X)"), x)
p.dist
```

```
##              0        1        2       3        4        5        6
## X     0.000000 1.000000 2.000000 3.00000 4.000000 5.000000 6.000000
## p(X) 0.117649 0.302526 0.324135 0.18522 0.059535 0.010206 0.000729
```

- **Đồ thị**: Sử dụng biểu đồ histogram hoặc biểu đồ cột để trực quan hóa các xác suất.

```
library(ggplot2)
n <- 6
x <- 0:n
data <- data.frame(x = x, prob = dbinom(x, n, p = 0.3))
ggplot(data, aes(x = x, y = prob)) +
  geom_col(width = 0.4, fill = "skyblue") +
  labs(title = "Phân phối nhị thức n=6, p=0.3",
       x = "X", y = "Xác suất") +
  theme_bw()
```



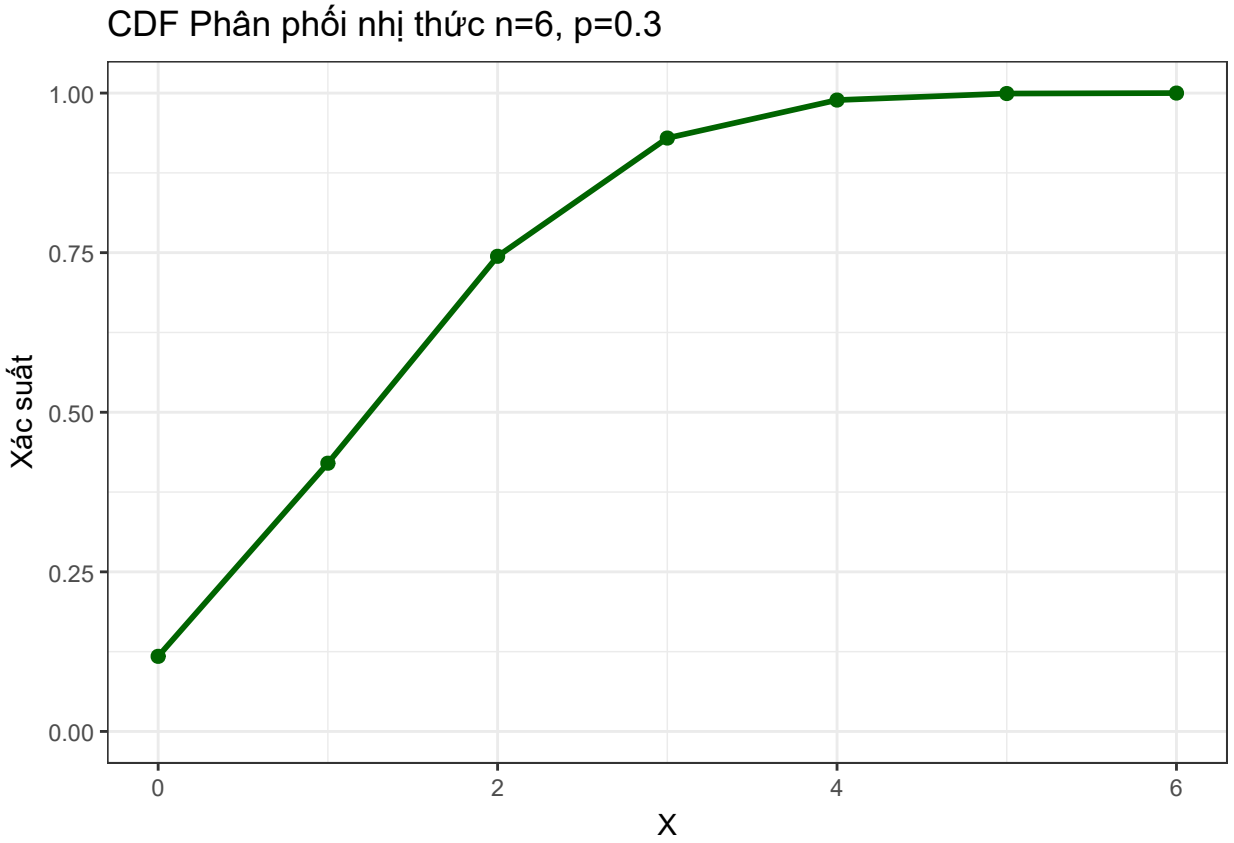
- **Công thức**: Biểu diễn phân phối xác suất dưới dạng hàm.

$$p(X = x) = p(x) = \binom{6}{x} (0.3)^x (0.7)^{6-x}$$

- **Các phương pháp khác**: Chẳng hạn như hàm phân phối tích lũy (CDF).

$$P(X \leq x) = F(x)$$

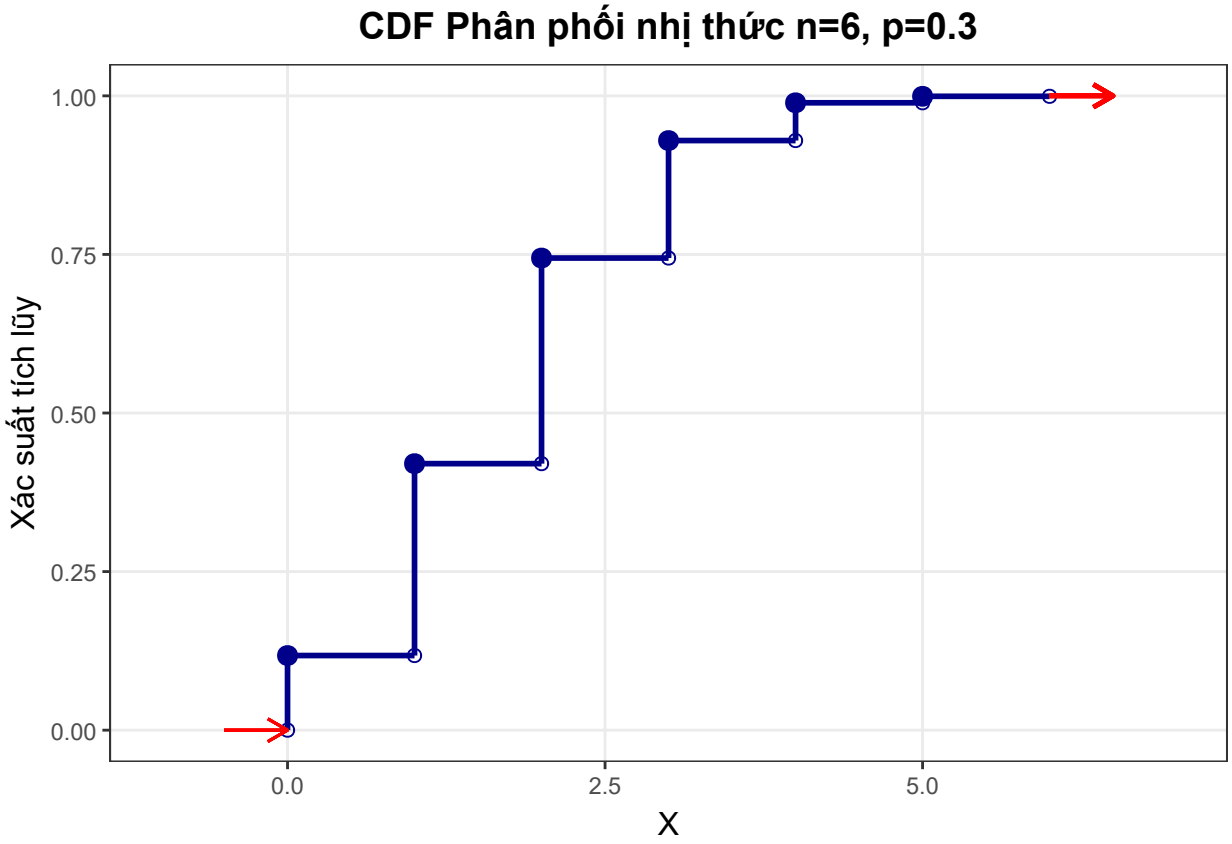
```r
library(ggplot2)
n <- 6
x <- 0:n
data_cdf <- data.frame(x = x, cdf = pbinom(x, n, p = 0.3))
ggplot(data_cdf, aes(x = x, y = cdf)) +
  geom_line(color = "darkgreen", linewidth = 1) +
  geom_point(color = "darkgreen", size = 2) +
  labs(title = "CDF Phân phối nhị thức n=6, p=0.3",
       x = "X", y = "Xác suất") +
  ylim(c(0, 1)) +
  theme_bw()
```



Tuy nhiên ta nên minh họa hàm phân phối tích lũy (CDF) cho biến rời rạc được vẽ dưới dạng hàm bậc thang, nơi giá trị nhảy tại các điểm xác suất tích lũy.Ta tự định nghĩa hàm sau để vẽ hình đẹp hơn:

```r
library(ggplot2)
library(dplyr)

disc_cdf <- function(x,
                     fx,
                     x_title="X",
                     y_title="Xác suất tích lũy",
                     main_title="CDF Phân phối nhị thức n=6, p=0.3"){
  Fx <- cumsum(fx)
  cdf_data <- data.frame(x = x, Fx = Fx) |>
  mutate(
    # Điểm bắt đầu và kết thúc cho đường ngang
    x_start = x,
    x_end = lead(x, default = max(x) + 1),
    y_start = Fx,
    y_end = lead(Fx, default = 1),
    # Điểm dọc nối (nếu cần, nhưng CDF rời rạc thường chỉ ngang rồi nhảy)
    x_vertical = x,
    y_vertical_from = lag(Fx, default = 0),
    y_vertical_to = Fx) |>
  filter(!is.na(x_end))  # loại hàng cuối nếu cần

  final_data <- data.frame(x = x, pdf_pmf=fx, cdf=Fx)
```

```r
  # Vẽ CDF bậc thang
  plot <- ggplot(cdf_data, aes(x = x)) +
    # Đường ngang cho bậc thang
    geom_segment(aes(x = x_start, xend = x_end, y = y_start, yend = y_start),
                 color = "darkblue", size = 1) +
    # Đường dọc cho bước nhảy
    geom_segment(aes(x = x, xend = x, y = y_vertical_from, yend = y_vertical_to),
                 color = "darkblue", size = 1, linetype = "solid") +
    # Điểm đánh dấu tại các bước
    geom_point(aes(y = Fx), size = 3, color = "darkblue", shape = 19) +
    # Điểm mở tại đầu mỗi bậc
    geom_point(aes(y = lag(Fx, default = 0)), size = 2,
               color = "darkblue", shape = 1) +
    # Mũi tên chỉ hướng tăng ở cuối
    geom_segment(aes(x = max(x),
                     xend = max(x) + 0.5,
                     y = 1,
                     yend = 1),
                 arrow = arrow(length = unit(0.3, "cm")),
                 color = "red", size = 0.8) +
    # Mũi tên chỉ hướng bắt đầu ở đầu
    geom_segment(aes(x = min(x) - 0.5,
                     xend = min(x),
                     y = 0,
                     yend = 0),
                 arrow = arrow(length = unit(0.3, "cm")),
                 color = "red",
                 ssize = 0.8) +
    labs(
      title = main_title,
      x = x_title,
      y = y_title
    ) +
    xlim(min(x) - 1, max(x) + 1) +
    ylim(0, 1) +
    theme_bw() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
      axis.title = element_text(size = 12),
      panel.grid.minor = element_blank()
    )
  return(list(data=final_data, p=plot))
}

n <- 6
x <- 0:n
fx <- dbinom(x, n, p = 0.3)
Fx <- cumsum(fx)

res <- disc_cdf(x, fx)

res$p
```



1.6 Phân phối nhị thức

Quá trình Bernoulli

- Quá trình Bernoulli là một chuỗi các thử nghiệm Bernoulli độc lập.
- Mỗi thử nghiệm chỉ có hai kết quả có thể: thành công (thường ký hiệu là 1) hoặc thất bại (ký hiệu là 0).
- Mỗi thử nghiệm có xác suất thành công cố định là p, và thất bại là $(1-p)$, giữ nguyên qua tất cả các thử nghiệm.

Biến ngẫu nhiên nhị thức

- Một biến ngẫu nhiên nhị thức, X, đếm số lượng thành công trong một số lượng cố định n thử nghiệm Bernoulli độc lập.

Hàm khối lượng xác suất (PMF)

Hàm khối lượng xác suất (PMF) của biến ngẫu nhiên nhị thức như sau:

$$f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

trong đó: - $\binom{n}{k}$ là hệ số nhị thức, biểu thị số cách chọn k thành công từ n thử nghiệm.

- p^k là xác suất của k thành công.
- $(1-p)^{n-k}$ là xác suất của $(n-k)$ thất bại.

Bản chất

Về bản chất, biến ngẫu nhiên nhị thức tóm tắt kết quả của quá trình Bernoulli qua (n) thử nghiệm.

Nếu mỗi thử nghiệm theo phân phối Bernoulli, thì tổng của các thử nghiệm Bernoulli độc lập sẽ tạo thành biến ngẫu nhiên nhị thức.

Ví dụ: Trò chơi Yahtzee

Yahtzee là trò chơi xúc xắc cổ điển, nơi người chơi tung năm con xúc xắc để tạo ra các tổ hợp điểm cao qua 13 vòng. Xác suất để có "four-of-a-kind" (bốn con giống nhau) trong một lần tung là gì?

Trước tiên, tập trung vào việc tung được bốn số 1 trong một lần tung. Gọi $X_1 = \#$ số 1 trong một lần tung năm con xúc xắc.

Một cách khác để nghĩ về nó:

$$P(X_1 = 4) = \binom{5}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^1 = 0.003215021$$

Các chuỗi như 11110, 11101, 11011, 10111, 01111, mỗi chuỗi có xác suất $\left(\frac{1}{6}\right)^4 \cdot \frac{5}{6}$.

Tiếp theo, vì mỗi trong các sự kiện sau $X_1 = 4$, $X_2 = 4$, $X_3 = 4$, $X_4 = 4$, $X_5 = 4$, $X_6 = 4$ là loại trừ lẫn nhau và có cùng xác suất xảy ra, chúng ta cộng các xác suất. Như vậy, xác suất tung được four-of-a-kind trong một lần tung là:

$$P(X_1 = 4 \cup X_2 = 4 \cup X_3 = 4 \cup X_4 = 4 \cup X_5 = 4 \cup X_6 = 4)$$

$$= P(X_1 = 4) + P(X_2 = 4) + P(X_3 = 4) + P(X_4 = 4) + P(X_5 = 4) + P(X_6 = 4)$$

$$= 6 \times P(X_i = 4)$$

$$= 6 \times \binom{5}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^1$$

$$= 6 \times 0.003215021 = 0.01929012$$

Hàm phân phối tích lũy (CDF) cho phân phối nhị thức

Giả sử $X \sim \text{Binom}(n, p)$.

- Hàm CDF:

$$F(k) = P(X \leq k) = P(X = 0) + P(X = 1) + \cdots + P(X = k).$$

- Và:

$$F(k) - F(k-1) = P(X \leq k) - P(X \leq k-1) = P(X = k).$$

Ví dụ: Lấy bài có thay thế

Lấy năm lá bài ngẫu nhiên, một lần một lá, với thay thế, từ bộ bài. Chúng ta quan tâm đến xác suất lấy được 1 hoặc nhiều lá Át. Gọi X là số lá Át được lấy.

$$P(X \geq 1) = \sum_{x=1}^{5} f(x \mid n = 5, p = 4/52)$$

$$= \sum_{x=1}^{5} \binom{5}{x} \left(\frac{4}{52}\right)^{x} \left(\frac{48}{52}\right)^{5-x}$$

Một cách khác (sử dụng phần bù):

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{5}{0} \left(\frac{4}{52}\right)^{0} \left(\frac{48}{52}\right)^{5-0}$$

$$= 1 - \left(\frac{48}{52}\right)^{5} = 0.3298$$

Ta có đoạn mã R để tính xác suất trên như sau:

```r
1 - dbinom(0, 5, 4/52)
```

```
## [1] 0.3298231
```

```r
# [1] 0.3298231

# Tổng từng trường hợp
sum(dbinom(1:5, 5, 4/52))
```

```
## [1] 0.3298231
```

```r
# [1] 0.3298231

# Sử dụng hàm CDF
pbinom(0, 5, 4/52, lower.tail = FALSE)
```

```
## [1] 0.3298231
```

```r
# [1] 0.3298231
```

1.7 Tính toán kỳ vọng và phương sai từ mô phỏng

Kỳ vọng ($\mathbb{E}[X]$) là giá trị trung bình dài hạn của biến ngẫu nhiên X, còn phương sai (ký hiệu $\text{Var}(X)$) đo lường mức độ biến thiên của X quanh kỳ vọng. Chúng ta sẽ sử dụng mô phỏng để ước lượng chúng một cách thực nghiệm, sau đó so sánh với công thức lý thuyết. Điều này giúp hiểu rõ cách mô phỏng hội tụ về giá trị thực khi số lần lặp tăng.

- Kỳ vọng giống như "trung tâm khối lượng" của phân phối, phương sai giống như "mức độ lan tỏa". Từ mô phỏng, chúng ta tính trung bình và độ lệch chuẩn bình phương của các mẫu giả lập.

Ví dụ: Xúc xắc

Đối với $X \sim \text{Uniform}\{1, 2, \ldots, 6\}$, $E[X] = \frac{1+6}{2} = 3.5$, $\text{Var}(X) = \frac{6^2-1}{12} = \frac{35}{12} \approx 2.9167$. Vì

$$E[X] = \sum xP(X = x) = \frac{1}{6}\sum_{x=1}^{6} x$$

và sử dụng tính chất

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

.

Cách tính từ mô phỏng

1. **Mô phỏng dữ liệu**: Tạo vector mẫu lớn (e.g., 10,000 lần tung) bằng `sample()`.
2. **Tính trung bình thực nghiệm**: `mean(samples)`.
3. **Tính phương sai thực nghiệm**: `var(samples)` (mẫu) hoặc `var(samples) * (n-1)/n` (dân số).
4. **So sánh với lý thuyết**: Vẽ hình để xem hội tụ khi cỡ mẫu tăng.

Ví dụ: Mô phỏng xúc xắc

Giả sử X là mặt xúc xắc (1-6, mỗi mặt xác suất $1/6$). Chúng ta mô phỏng để ước lượng $E[X] \approx 3.5$ và $\text{Var}(X) \approx 2.9167$. Ta có đoạn mã R tham khảo như sau:

```r
library(ggplot2)
library(dplyr)

set.seed(123)
n_sim <- 100
die_rolls <- sample(1:6, size = n_sim, replace = TRUE)

# Empirical stats
emp_mean <- mean(die_rolls)
emp_var <- var(die_rolls)

# Theoretical stats
theo_mean <- (1 + 6) / 2  # 3.5
theo_var <- (6^2 - 1) / 12  # 2.9167

# Print result
cat("Empirical Mean:", round(emp_mean, 4), "\n")
```

```
## Empirical Mean: 3.53
```

```r
cat("Theoretical Mean:", theo_mean, "\n")
```

```
## Theoretical Mean: 3.5
```

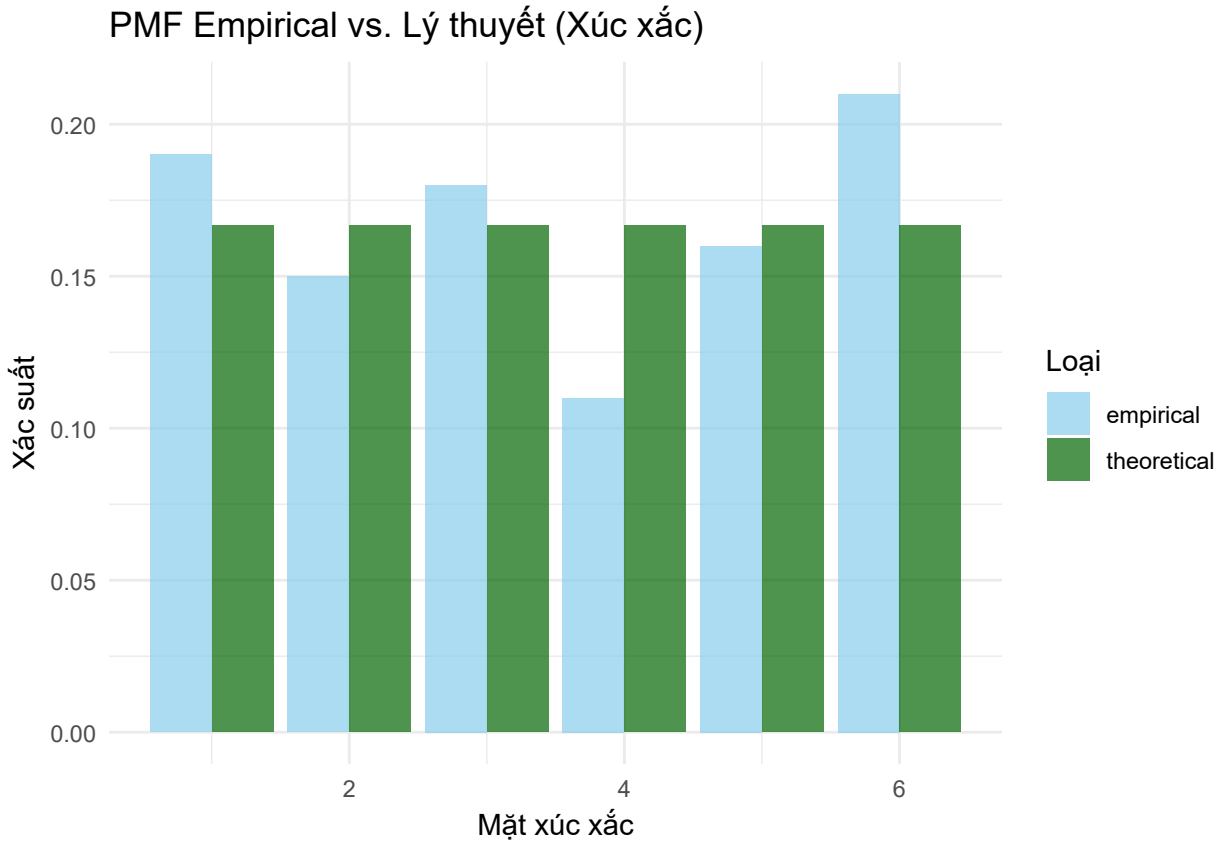
```r
cat("Empirical Var:", round(emp_var, 4), "\n")
```

```
## Empirical Var: 3.3021
```

```r
cat("Theoretical Var:", round(theo_var, 4), "\n")
```

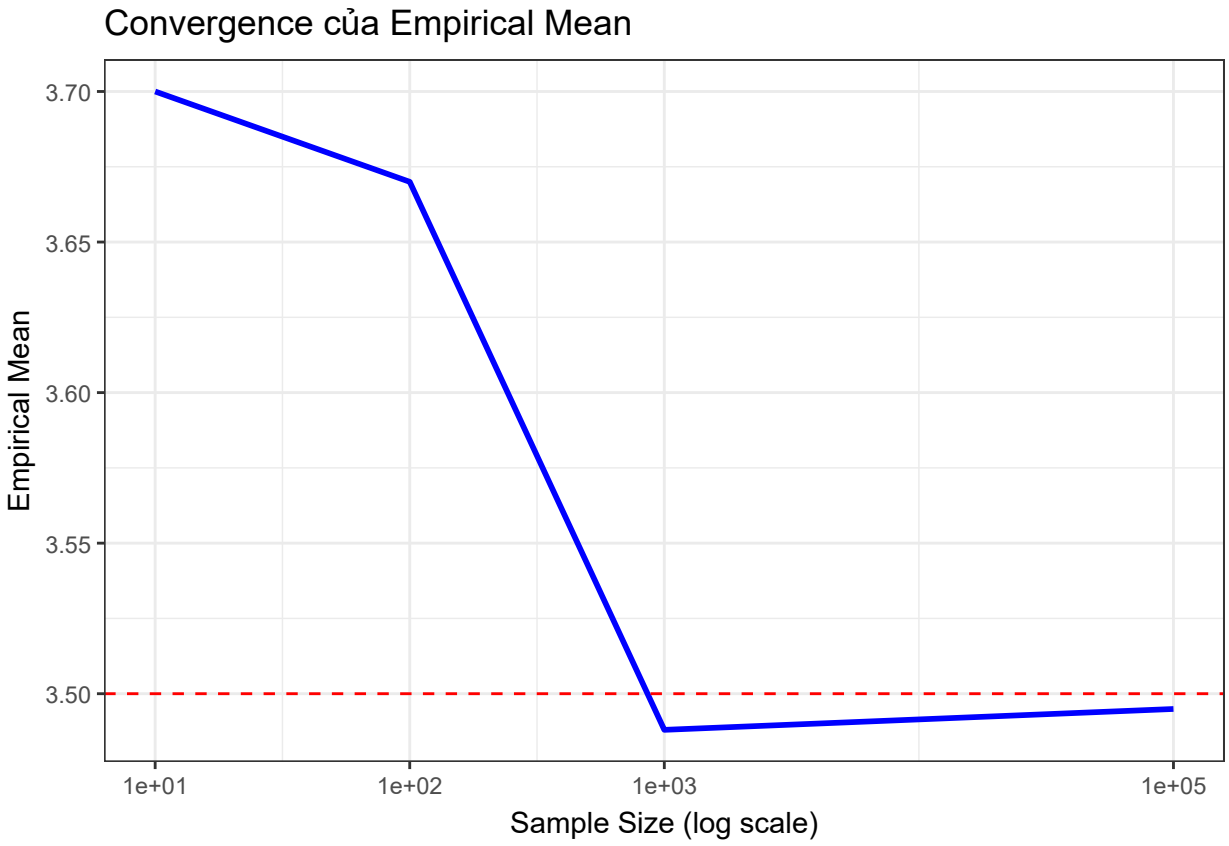
```
## Theoretical Var: 2.9167
```

```r
#  PMF empirical vs theoretical
data_plot <- data.frame(
  x = 1:6,
  empirical = sapply(1:6, function(k) mean(die_rolls == k)),
  theoretical = rep(1/6, 6)
) |>
  pivot_longer(cols = c(empirical, theoretical),
  names_to = "type", values_to = "prob")

ggplot(data_plot, aes(x = x, y = prob, fill = type)) +
  geom_col(position = "dodge", alpha = 0.7) +
  labs(title = "PMF Empirical vs. Lý thuyết (Xúc xắc)",
       x = "Mặt xúc xắc", y = "Xác suất",
       fill = "Loại") +
  theme_minimal() +
  scale_fill_manual(values = c("empirical" = "skyblue", "theoretical" = "darkgreen"))
```




```r
# Convergence: Mean theo sample size
sample_sizes <- c(10, 100, 1000, 100000)
convergence_data <- data.frame(
  size = sample_sizes,
  emp_mean = sapply(sample_sizes, function(ns) mean(sample(1:6, ns, replace = TRUE)))
)

ggplot(convergence_data, aes(x = size, y = emp_mean)) +
```

```r
  geom_line(color = "blue", size = 1) +
  geom_hline(yintercept = theo_mean, linetype = "dashed", color = "red") +
  scale_x_continuous(trans="log10", breaks=convergence_data$size) +
  labs(title = "Convergence của Empirical Mean",
       x = "Sample Size (log scale)", y = "Empirical Mean") +
  theme_bw()
```



Bài tập 2. Trung bình và phương sai của phân phối nhị thức

Đối với $X \sim \text{Binom}(n = 10, p = 0.3)$, $E[X] = np = 3$, $\text{Var}(X) = np(1-p) = 2.1$. Hãy thực hiện mô phỏng với $n_{\text{sim}} = 10000$ và kiêm tra trung bình và phương sai thực nghiệm

```
## Binomial Empirical Mean: 3.0048
## Binomial Theoretical Mean: 3
## Binomial Empirical Var: 2.0922
## Binomial Theoretical Var: 2.1
```

Bài tập 3: Kiểm tra sự hội tụ của trung bình và phương sai thực nghiệm khi n_{sim} tăng

1. Chạy mã R xúc xắc với n_sim = 100, 1000, 10000. Quan sát hội tụ của trung bình/phương sai.
2. Thay đổi p=0.5 cho nhị thức, tính và vẽ PMF empirical. So sánh với `dbinom()`.
3. Tại sao trung bình thực nghiệm hội tụ về giá trị lý thuyết khi n_sim tăng? Có thể gợi ra tên của một định lý nào đó đảm bảo điều này không?

Ta tham khảo hình vẽ mẫu sau:

```
## NULL
```