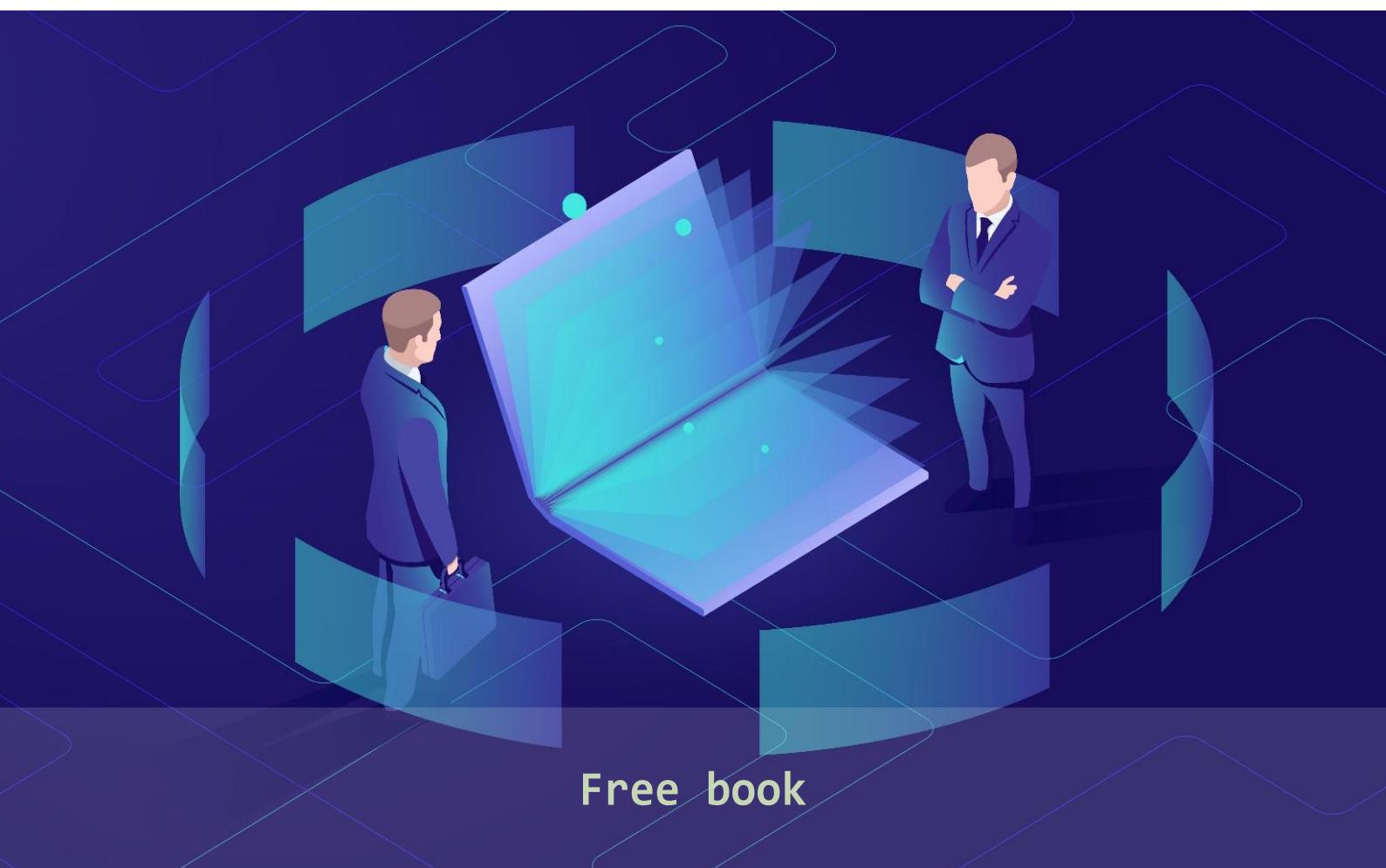


Faculty of Information Technology

DATA ANALYSIS



Len Bui

PHÂN TÍCH DỮ LIỆU

Bùi Tiến Lên

2023

Danh mục thuật ngữ và ký hiệu

ký hiệu	ý nghĩa
$a, b, c, N \dots$	số nguyên, số thực, giá trị
$\mathbf{w}, \mathbf{v}, \mathbf{x}, \mathbf{y} \dots$	vector cột
$\mathbf{X}, \mathbf{Y} \dots$	ma trận
\mathbb{R}	tập hợp số thực
\mathbb{Z}	tập hợp số nguyên
\mathbb{N}	tập hợp số tự nhiên
\mathbb{R}^D	tập hợp số thực nhiều chiều
$\mathcal{D}, \mathcal{X}, \mathcal{Y}, \dots$	tập hợp
$X, Y \dots$	biến ngẫu nhiên
$\mathbf{X}, \mathbf{Y} \dots$	biến ngẫu nhiên nhiều chiều
$x, y \dots$	giá trị
$\mathbf{x}, \mathbf{y} \dots$	giá trị
p, pr, P, Pr	xác suất
\mathbf{w}^\top	chuyển vị ma trận
\mathbf{XY}	nhân ma trận
\mathbf{X}^{-1}	nghịch đảo ma trận
$\mathbf{X} \odot \mathbf{Y}$	nhân từng phân tử

Mục lục

I. XÁC SUẤT	1
1. Giải tích tổ hợp	2
1.1. Cơ sở phép đếm	2
1.1.1. Những nguyên lý đếm cơ bản	2
1.1.2. Nguyên lý bù trừ	4
1.2. Nguyên lý Dirichlet	6
1.2.1. Giới thiệu	6
1.2.2. Nguyên lý Dirichlet tổng quát	7
1.2.3. Một số ứng dụng của nguyên lý Dirichlet	7
1.3. Chính hợp và tổ hợp	8
1.3.1. Chính hợp có lắp	8
1.3.2. Tổ hợp có lắp	9
1.3.3. Hoán vị của tập hợp có các phần tử giống nhau	9
1.3.4. Sự phân bố các đồ vật vào trong hộp	10
1.4. Sinh các hoán vị và tổ hợp	11
1.4.1. Sinh các hoán vị	11
1.4.2. Sinh các tổ hợp	12
Tóm tắt chương	13
📝 Bài tập	14
💻 Lập trình	18
2. Xác suất	19
2.1. Khái niệm xác suất	19
2.1.1. Phép thử và sự kiện	19
2.1.2. Định nghĩa xác suất	21
2.2. Các phép toán xác suất	24
2.2.1. Tổng xác suất	24
2.2.2. Xác suất có điều kiện	25
2.2.3. Tích xác suất	25
2.2.4. Công thức Bayes	26
2.2.5. Công thức Bernoulli	27

2.3.	Nguyên lý xác suất nhỏ, nguyên lý xác suất lớn	28
2.3.1.	Nguyên lý xác suất nhỏ	28
2.3.2.	Nguyên lý xác suất lớn	28
Bài tập		28
3.	Biến ngẫu nhiên và hàm phân phối	37
3.1.	Biến ngẫu nhiên	37
3.2.	Phân phối xác suất	38
3.2.1.	Hàm khối xác suất của biến rời rạc	38
3.2.2.	Hàm phân phối xác suất biến liên tục	38
3.2.3.	Hàm mật độ xác suất của biến liên tục	40
3.3.	Các đặc trưng	43
3.3.1.	Đồ thị hàm phân bố xác suất	43
3.3.2.	Kỳ vọng	43
3.3.3.	Phương sai	44
3.3.4.	Độ lệch chuẩn	45
3.3.5.	Trung vị	45
3.3.6.	Yếu vị	45
3.3.7.	Moment	46
3.3.8.	Entropy	46
3.3.9.	Cross entropy và KL divergence	46
3.4.	Biến ngẫu nhiên nhiều chiều	47
3.4.1.	Phân phối xác suất	47
3.4.2.	Các đặc trưng	51
3.4.3.	Đặc trưng có điều kiện	52
3.5.	Hàm và biến ngẫu nhiên	54
3.5.1.	Hàm của biến ngẫu nhiên	54
3.5.2.	Hàm lồi và biến ngẫu nhiên	55
3.6.	Luật số lớn và định lý giới hạn trung tâm	55
3.6.1.	Luật số lớn	55
3.6.2.	Định lý giới hạn trung tâm	57
Bài tập		58
4.	Một số phân phối phổ biến	66
4.1.	Biến ngẫu nhiên rời rạc	66
4.1.1.	Phân phối đều (discrete uniform distribution)	66
4.1.2.	Phân phối Bernoulli (Bernoulli distribution)	67
4.1.3.	Phân phối loại (categorical distribution)	68
4.1.4.	Phân phối nhị thức (Binomial distribution)	68
4.1.5.	Phân phối đa thức (multinomial distribution)	69

Mục lục

4.1.6. Phân phối Poisson (Poisson distribution)	70
4.1.7. Phân phối hình học (geometric distribution)	71
4.1.8. Phân phối nhị thức âm (negative binominal distribution)	72
4.2. Biến ngẫu nhiên liên tục	72
4.2.1. Phân phối đều (continuous uniform distribution)	72
4.2.2. Phân phối chuẩn (normal distribution)	73
4.2.3. Phân phối mũ (exponential distribution)	77
4.2.4. Phân phối Student	77
4.2.5. Phân phối Chi squared	78
4.2.6. Phân phối Fisher	79
4.2.7. Phân phối Beta	80
4.2.8. Phân phối Gamma	81
4.2.9. Phân phối Logistic	82
4.2.10. Phân phối Cauchy	83
☒ Bài tập	84

II. PHÂN TÍCH DỮ LIỆU 92

5. Mẫu thống kê và ước lượng tham số 94	
5.1. Tổng thể	94
5.2. Mẫu thống kê	94
5.2.1. Một số cách chọn mẫu cơ bản	95
5.2.2. Mẫu ngẫu nhiên	95
5.2.3. Đại lượng thống kê	96
5.3. Phân phối của đại lượng thống kê của mẫu thống kê	100
5.3.1. Phân phối của trung bình mẫu	100
5.3.2. Phân phối của phương sai mẫu	101
5.4. Ước lượng điểm cho tham số	101
5.4.1. Tham số	101
5.4.2. Các tiêu chuẩn lựa chọn phương pháp ước lượng	102
5.4.3. ML	102
5.4.4. MAP	104
5.4.5. EM	107
5.5. Ước lượng khoảng tin cậy cho tham số	111
5.5.1. Khái niệm ước lượng khoảng tin cậy	111
5.5.2. Phương pháp	111
5.6. Ước lượng kích thước mẫu	113
☒ Bài tập	114

6. Kiểm định thống kê	120
6.1. Thủ tục kiểm định giả thuyết thống kê	120
6.1.1. Sai lầm loại I và II	121
6.2. Kiểm định tham số	121
6.2.1. Kiểm định một tham số, một tổng thể, một mẫu	121
6.2.2. Kiểm định hai tham số, hai tổng thể, hai mẫu	124
6.3. Kiểm định phi tham số	128
6.3.1. Kiểm định tính độc lập của hai dấu hiệu định tính	128
6.3.2. Kiểm định tính phân phối chuẩn - Jacque-Berra	128
6.3.3. Kiểm định phân phối - Kolmogorov-Smirnov	128
☒ Bài tập	129
7. Phân tích hồi quy tuyến tính đơn	136
7.1. Hồi quy tuyến tính đơn	136
7.2. Ước lượng qui luật bằng OLS	138
7.3. Tính chất của hàm hồi quy mẫu	141
7.4. Vấn đề về thay đổi đơn vị của biến	143
7.5. Vấn đề tương quan	143
7.5.1. Nguyên nhân của tự tương quan	143
7.5.2. Phát hiện hiện tượng tự tương quan	144
7.5.3. Các biện pháp khắc phục	144
7.6. Vấn đề phương sai thay đổi	145
7.6.1. Giải pháp khắc phục WLS	146
7.7. Kiểm định mô hình ước lượng	147
7.8. Ứng dụng mô hình hồi qui	148
☒ Bài tập	149
8. Phân tích hồi quy bội	158
8.1. Biểu diễn đại số của mô hình hồi quy	158
8.2. Ví dụ hồi quy	161
8.3. Kiểm định mô hình	162
8.3.1. Phân tích phần dư	162
8.3.2. Phân tích phương sai	163
8.3.3. Kiểm định đa cộng tuyến	164
8.4. Lựa chọn mô hình	167
8.4.1. Phương pháp stepwise regression	167
8.5. Hồi quy với biến chuẩn hóa	168
8.6. Hồi quy với biến định tính	169
8.6.1. Biến độc lập là biến định tính	169
8.6.2. Biến độc lập là biến định tính + biến định lượng	171

Mục lục

8.6.3. Ảnh hưởng tương tác	172
8.7. Hồi quy với hàm cơ sở	172
8.8. Nghịch lý Simpson	173
☒ Bài tập	174
9. Phân tích hồi quy logistic	183
9.1. Mô hình tổng quát và Ước lượng tham số	184
9.2. Ví dụ hồi quy	185
9.3. Đánh giá mô hình	187
9.3.1. Đường cong ROC	187
9.4. Mô hình hồi quy Probit	188
☒ Bài tập	189
10. Phân tích phương sai (ANOVA)	190
10.1. Phân tích phương sai một yếu tố (one-way ANOVA)	190
10.1.1. Mô hình phân tích phương sai	191
10.2. Phân tích phương sai hai yếu tố (two-way ANOVA)	193
10.2.1. Mô hình phân tích phương sai	193
☒ Bài tập	196
11. Phân tích dữ liệu chuỗi thời gian	201
11.1. Một số đặc trưng của chuỗi thời gian	203
11.2. Các kỹ thuật tính toán xử lý dữ liệu	204
11.2.1. Làm khớp đường cong với chuỗi dữ liệu	204
11.2.2. Hàm biến đổi chuỗi dữ liệu	205
11.2.3. Lọc (làm tròn) chuỗi dữ liệu	205
11.2.4. Sai phân chuỗi dữ liệu	206
11.2.5. Hàm tự tương quan	207
11.3. Các thành phần của chuỗi thời gian	208
11.3.1. Phân rã chuỗi dữ liệu	209
11.4. Quá trình ngẫu nhiên	210
11.4.1. Bước ngẫu nhiên	211
11.4.2. Quá trình trung bình trượt	212
11.4.3. Quá trình tự hồi quy	213
11.5. Chuỗi dữ liệu dừng	216
11.5.1. Kiểm định tính dừng	216
11.6. Cú pháp Backshift	217
11.7. Mô hình ARMA, ARIMA, SARIMA	217
11.7.1. Mô hình ARMA	217
11.7.2. Mô hình ARIMA	218

Mục lục

11.7.3. Mô hình SARIMA	222
11.8. Phương pháp Box-Jenkins	222
11.8.1. Tiêu chí lựa chọn mô hình	223
11.8.2. Ví dụ hồi quy mô hình và áp dụng	224
11.9. Hồi quy với lỗi ARIMA	225
☒ Bài tập	225
12. Phân tích dữ liệu bảng	226
12.1. Các loại dữ liệu	226
12.2. Các ưu điểm của dữ liệu bảng	227
12.3. Mô hình VAR	227
12.4. Mô hình hồi quy	227
12.5. Mô hình ảnh hưởng cố định (Fixed effects model)	228
12.6. Mô hình ảnh hưởng ngẫu nhiên (Random effects model)	229
☒ Bài tập	229
13. Phương pháp mô phỏng Monte Carlo	230
13.1. Một số khái niệm	230
13.2. Mô hình hóa hệ thống	231
13.2.1. Tại sao cần mô hình hóa hệ thống	231
13.2.2. Các loại mô hình hóa hệ thống	231
13.3. Phương pháp mô phỏng	232
13.4. Phương pháp mô phỏng Monte Carlo	233
13.4.1. Các loại số ngẫu nhiên	234
13.4.2. Phương pháp tạo số giả ngẫu nhiên và lấy mẫu	235
13.4.3. Phương pháp Monte Carlo	237
☒ Bài tập	239
Tài liệu tham khảo	241

Phần I.

XÁC SUẤT

Lý thuyết tổ hợp là một phần quan trọng của toán học rời rạc chuyên nghiên cứu sự phân bố các phần tử vào các tập hợp. Thông thường các phần tử này là hữu hạn và việc phân bố chúng phải thoả mãn những điều kiện nhất định nào đó, tùy theo yêu cầu của bài toán cần nghiên cứu. Mỗi cách phân bố như vậy gọi là một cấu hình tổ hợp. Chủ đề này đã được nghiên cứu từ thế kỷ 17, khi những câu hỏi về tổ hợp được nêu ra trong những công trình nghiên cứu các trò chơi may rủi. Liệt kê, đếm các đối tượng có những tính chất nào đó là một phần quan trọng của lý thuyết tổ hợp. Chúng ta cần phải đếm các đối tượng để giải nhiều bài toán khác nhau. Các kỹ thuật đếm được dùng rất nhiều trong tin học đặc biệt là trong phân tích thuật toán và lý thuyết xác suất.

1.1 Cơ sở phép đếm

1.1.1. Những nguyên lý đếm cơ bản

Quy tắc cộng

Phát biểu. Giả sử có k công việc T_1, T_2, \dots, T_k . Các việc này có thể làm tương ứng bằng n_1, n_2, \dots, n_k cách và giả sử không có hai việc nào có thể làm đồng thời. Khi đó số cách làm một trong k việc đó là $n_1 + n_2 + \dots + n_k$.

Ví dụ 1.1. Một sinh viên có thể chọn bài thực hành máy tính từ một trong ba danh sách tương ứng có 23, 15 và 19 bài. Vì vậy, theo quy tắc cộng có $23 + 15 + 19 = 57$ cách chọn bài thực hành.

Ví dụ 1.2. Giá trị của biến m bằng bao nhiêu sau khi đoạn chương trình Python sau được thực hiện?

```
m = 0
for i1 in range(n1):
    m += 1
for i2 in range(n2):
    m += 1
...
```

1. Giải tích tổ hợp

```
for ik in range(nk):  
    m += 1
```

Lời giải. Giá trị khởi tạo của m bằng 0. Khối lệnh này gồm k vòng lặp khác nhau. Sau mỗi bước lặp của từng vòng lặp giá trị của k được tăng lên một đơn vị. Gọi T_i là việc thi hành vòng lặp thứ i . Có thể làm T_i bằng n_i cách vì vòng lặp thứ i có n_i bước lặp. Do các vòng lặp không thể thực hiện đồng thời nên theo quy tắc cộng, giá trị cuối cùng của m bằng số cách thực hiện một trong số các nhiệm vụ T_i , tức là $m = n_1 + n_2 + \dots + n_k$. ■

Ngôn ngữ tập hợp Nếu A_1, A_2, \dots, A_k là các tập hợp đôi một rời nhau, khi đó số phần tử của hợp các tập hợp này bằng tổng số các phần tử của các tập thành phần. Giả sử T_i là việc chọn một phần tử từ tập A_i với $i = 1, 2, \dots, k$. Có $|A_i|$ cách làm và không có hai việc nào có thể được làm cùng một lúc. Số cách chọn một phần tử của hợp các tập hợp này, một mặt bằng số phần tử của nó, mặt khác theo quy tắc cộng nó bằng $|A_1| + |A_2| + \dots + |A_k|$. Do đó ta có: $|A_1 \cup A_2 \cup \dots \cup A_k| = |A_1| + |A_2| + \dots + |A_k|$.

Quy tắc nhân

Phát biểu. Giả sử một nhiệm vụ nào đó được tách ra thành k việc T_1, T_2, \dots, T_k . Nếu việc T_i có thể làm bằng n_i cách sau khi các việc T_1, T_2, \dots, T_{i-1} đã được làm, khi đó có $n_1 \cdot n_2 \dots n_k$ cách thi hành nhiệm vụ đã cho.

Ví dụ 1.3. Người ta có thể ghi nhãn cho những chiếc ghế trong một giảng đường bằng một chữ cái và một số nguyên dương không vượt quá 100. Bằng cách như vậy, nhiều nhất có bao nhiêu chiếc ghế có thể được ghi nhãn khác nhau?

Lời giải. Thủ tục ghi nhãn cho một chiếc ghế gồm hai việc, gán một trong 26 chữ cái và sau đó gán một trong 100 số nguyên dương. Quy tắc nhân chỉ ra rằng có $26 \cdot 100 = 2600$ cách khác nhau để gán nhãn cho một chiếc ghế. Như vậy nhiều nhất ta có thể gán nhãn cho 2600 chiếc ghế. ■

Ví dụ 1.4. Có bao nhiêu chuỗi nhị phân có độ dài n .

Lời giải. Mỗi bit trong n bit của chuỗi nhị phân có thể chọn bằng hai cách vì mỗi bit hoặc bằng 0 hoặc bằng 1. Bởi vậy theo quy tắc nhân có tổng cộng 2^n chuỗi nhị phân khác nhau có độ dài bằng n . ■

Ví dụ 1.5. Có thể tạo được bao nhiêu ánh xạ từ tập A có m phần tử vào tập B có n phần tử?

Lời giải. Theo định nghĩa, một ánh xạ xác định trên A có giá trị trên B là một phép tương ứng mỗi phần tử của A với một phần tử nào đó của B . Rõ ràng sau khi đã chọn được ảnh của $i - 1$ phần tử đầu, để chọn ảnh của phần tử thứ i của A ta có n cách. Vì vậy theo quy tắc nhân, ta có $n \cdot n \dots n = n^m$ ánh xạ xác định trên A nhận giá trị trên B . ■

1. Giải tích tổ hợp

Ví dụ 1.6. Có bao nhiêu đơn ánh xác định trên tập A có m phần tử và nhận giá trị trên tập B có n phần tử?

Lời giải. Nếu $m > n$ thì với mọi ánh xạ, ít nhất có hai phần tử của A có cùng một ảnh, điều đó có nghĩa là không có đơn ánh từ A đến B . Vậy giờ giả sử $m \leq n$ và gọi các phần tử của A là a_1, a_2, \dots, a_m . Rõ ràng có n cách chọn ảnh cho phần tử a_1 . Vì ánh xạ là đơn ánh nên ảnh của phần tử a_2 phải khác ảnh của a_1 nên chỉ có $n - 1$ cách chọn ảnh cho phần tử a_2 . Nói chung, để chọn ảnh của a_k ta có $n - k + 1$ cách. Theo quy tắc nhân, ta có

$$n(n-1)(n-2)\dots(n-m+1) = \frac{n!}{(n-m)!}$$

đơn ánh từ tập A đến tập B . ■

Ví dụ 1.7. Giá trị của biến m bằng bao nhiêu sau khi đoạn chương trình Python sau được thực hiện?

```
m = 0
for i1 in range(n1):
    for i2 in range(n2):
        ...
        for ik in range(nk):
            m += 1
```

Lời giải. Giá trị khởi tạo của m bằng 0. Ta có k vòng lặp được lồng nhau. Gọi T_i là việc thi hành vòng lặp thứ i . Khi đó số lần đi qua vòng lặp bằng số cách làm các việc T_1, T_2, \dots, T_k . Số cách thực hiện việc T_j là n_j với ($j = 1, 2, \dots, k$), vì vòng lặp thứ j được duyệt với mỗi giá trị nguyên i_j nằm giữa 1 và n_j . Theo quy tắc nhân vòng lặp lồng nhau này được duyệt qua $n_1.n_2\dots n_k$ lần. Vì vậy giá trị cuối cùng của m là $n_1.n_2\dots n_k$. ■

Ngôn ngữ tập hợp Nếu A_1, A_2, \dots, A_k là các tập hữu hạn, khi đó số phần tử của tích Descartes của các tập này bằng tích của số các phần tử của mọi tập thành phần. Ta biết rằng việc chọn một phần tử của tích Descartes $A_1 \times A_2 \times \dots \times A_k$ được tiến hành bằng cách chọn lần lượt một phần tử của A_1 , một phần tử của A_2, \dots , một phần tử của A_k . Theo quy tắc nhân ta có: $|A_1 \times A_2 \times \dots \times A_k| = |A_1| \cdot |A_2| \dots |A_k|$.

1.1.2. Nguyên lý bù trừ

Khi hai công việc có thể được làm đồng thời, ta không thể dùng quy tắc cộng để tính số cách thực hiện nhiệm vụ gồm cả hai việc. Để tính đúng số cách thực hiện nhiệm vụ này ta cộng số cách làm mỗi một trong hai việc rồi trừ đi số cách làm đồng thời cả hai việc.

1. Giải tích tổ hợp

Ngôn ngữ tập hợp.

- Cho A_1, A_2 là hai tập hữu hạn, khi đó

$$|A_1 \cup A_2| = |A_1| + |A_2| - |A_1 \cap A_2|$$

- Từ đó với ba tập hợp hữu hạn A_1, A_2, A_3 , ta có:

$$|A_1 \cup A_2 \cup A_3| = |A_1| + |A_2| + |A_3| - |A_1 \cap A_2| - |A_2 \cap A_3| - |A_3 \cap A_1| + |A_1 \cap A_2 \cap A_3|$$

- Bằng quy nạp, với k tập hữu hạn A_1, A_2, \dots, A_k ta có:

$$|A_1 \cup A_2 \cup \dots \cup A_k| = N_1 - N_2 + N_3 - \dots + (-1)^{k-1} N_k$$

trong đó $N_m (m = 1, \dots, k)$ là tổng phần tử của tất cả các giao m tập lấy từ k tập đã cho, nghĩa là

$$N_m = \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq k} A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_m}$$

- Bây giờ ta đồng nhất tập $A_m (1 \leq m \leq k)$ với tính chất A_m cho trên tập vũ trụ hữu hạn U nào đó và đếm xem có bao nhiêu phần tử của U sao cho không thỏa mãn bất kỳ một tính chất A_m nào. Gọi \bar{N} là số cần đếm, N là số phần tử của U . Ta có:

$$N = N - |A_1 \cup A_2 \cup \dots \cup A_k| = N - N_1 + N_2 - N_3 + \dots + (-1)^k N_k$$

trong đó N_m là tổng các phần tử của U thỏa mãn m tính chất lấy từ k tính chất đã cho. Công thức này được gọi là nguyên lý bù trừ. Nó cho phép tính \bar{N} qua các N_m .

Ví dụ 1.8. Có n lá thư và n phong bì ghi sẵn địa chỉ. Bỏ ngẫu nhiên các lá thư vào các phong bì. Hỏi xác suất để xảy ra không một lá thư nào đúng địa chỉ.

Lời giải. Mỗi phong bì có n cách bỏ thư vào, nên có tất cả $n!$ cách bỏ thư. Vấn đề còn lại là đếm số cách bỏ thư sao cho không lá thư nào đúng địa chỉ. Gọi U là tập hợp các cách bỏ thư và A_m là tính chất lá thư thứ m bỏ đúng địa chỉ. Khi đó theo công thức về nguyên lý bù trừ ta có:

$$\bar{N} = n! - N_1 + N_2 - \dots + (-1)^n N_n$$

trong đó $N_m (1 \leq m \leq n)$ là số tất cả các cách bỏ thư sao cho có m lá thư đúng địa chỉ. Nhận xét rằng, N_m là tổng theo mọi cách lấy m lá thư từ n lá, với mỗi cách lấy m lá thư, có $(n-m)!$ cách bỏ để m lá thư này đúng địa chỉ, ta nhận được:

$$N_m = \binom{n}{m} (n-m)! = \frac{n!}{m!}$$

1. Giải tích tổ hợp

và

$$\bar{N} = n! \left(1 - \frac{1}{1!} + \frac{1}{2!} - \dots + (-1)^n \frac{1}{n!} \right)$$

Từ đó xác suất cần tìm là

$$P = 1 - \frac{1}{1!} + \frac{1}{2!} - \dots + (-1)^n \frac{1}{n!} \rightarrow \frac{1}{e}$$

khi n khá lớn. ■

n	2	3	4	5	6	7	8	9	10	11
\bar{N}	1	2	9	44	265	1854	14833	133496	1334961	14684570

1.2

Nguyên lý Dirichlet

1.2.1. Giới thiệu

Giả sử có một đàn chim bồ câu bay vào chuồng. Nếu số chim nhiều hơn số ngăn chuồng thì ít nhất trong một ngăn có nhiều hơn một con chim. Nguyên lý này dĩ nhiên là có thể áp dụng cho các đối tượng không phải là chim bồ câu và chuồng chim.

Định lý 1.1. *Nếu có $k+1$ (hoặc nhiều hơn) đồ vật được đặt vào trong k hộp thì tồn tại một hộp có ít nhất hai đồ vật.*

Chứng minh. Giả sử không có hộp nào trong k hộp chứa nhiều hơn một đồ vật. Khi đó tổng số vật được chứa trong các hộp nhiều nhất là bằng k . Điều này trái giả thiết là có ít nhất $k+1$ vật. ■

Nguyên lý này thường được gọi là nguyên lý Dirichlet, mang tên nhà toán học người Đức ở thế kỷ 19. Ông thường xuyên sử dụng nguyên lý này trong công việc của mình.

Ví dụ 1.9. Trong bất kỳ một nhóm 367 người thế nào cũng có ít nhất hai người có ngày sinh nhật giống nhau.

Lời giải. Bởi vì chỉ có tất cả 366 ngày sinh nhật khác nhau. ■

Ví dụ 1.10. Trong kỳ thi học sinh giỏi, điểm bài thi được đánh giá bởi một số nguyên trong khoảng từ 0 đến 100. Hỏi rằng ít nhất có bao nhiêu học sinh dự thi để cho chắc chắn tìm được hai học sinh có kết quả thi như nhau?

Lời giải. Theo nguyên lý Dirichlet, số học sinh cần tìm là 102, vì ta có 101 kết quả thi khác nhau. ■

Ví dụ 1.11. Trong số những người có mặt trên trái đất, phải tìm được hai người có hàm răng giống nhau.

1. Giải tích tổ hợp

Lời giải. Nếu xem mỗi hàm răng gồm 32 cái như là một chuỗi nhị phân có chiều dài 32, trong đó răng còn ứng với bit 1 và răng mất ứng với bit 0, thì có tất cả $2^{32} = 4294967296$ hàm răng khác nhau. Trong khi đó số người trên hành tinh này là vượt quá 7 tỉ, nên theo nguyên lý Dirichlet ta có điều cần tìm. ■

1.2.2. Nguyên lý Dirichlet tổng quát

Định lý 1.2. *Nếu có N đồ vật được đặt vào trong k hộp thì sẽ tồn tại một hộp chứa ít nhất $\lceil \frac{N}{k} \rceil$ đồ vật.*

Chứng minh. Giả sử mọi hộp đều chứa ít hơn $\lceil \frac{N}{k} \rceil$ vật. Khi đó tổng số đồ vật là $\leq k(\lceil \frac{N}{k} \rceil - 1) < k \cdot \frac{N}{k} = N$. Điều này mâu thuẫn với giả thiết là có N đồ vật cần xếp. ■

Ví dụ 1.12. Xét một số ví dụ đơn giản sau

- Trong 100 người, có ít nhất 9 người sinh cùng một tháng. Xếp những người sinh cùng tháng vào một nhóm. Có 12 tháng tất cả. Vậy theo nguyên lý Dirichlet, tồn tại một nhóm có ít nhất $\lceil \frac{100}{12} \rceil = 9$ người.
- Có năm loại học bổng khác nhau. Hỏi rằng phải có ít nhất bao nhiêu sinh viên để chắc chắn rằng có ít ra là 6 người cùng nhận học bổng như nhau. Gọi N là số sinh viên, khi đó $\lceil \frac{N}{5} \rceil = 6$ khi và chỉ khi $5 < \frac{N}{5} \leq 6$ hay $25 < N \leq 30$. Vậy số N cần tìm là 26.
- Số mã vùng cần thiết nhỏ nhất phải là bao nhiêu để đảm bảo 25 triệu máy điện thoại trong nước có số điện thoại khác nhau, mỗi số có 9 chữ số (giả sử số điện thoại có dạng 0XX - 8XXXXXX với X nhận các giá trị từ 0 đến 9). Có $10^7 = 10.000.000$ số điện thoại khác nhau có dạng 0XX - 8XXXXXX. Vì vậy theo nguyên lý Dirichlet tổng quát, trong số 25 triệu máy điện thoại ít nhất có $\lceil \frac{25.000.000}{10.000.000} \rceil = 3$ có cùng một số. Để đảm bảo mỗi máy có một số cần có ít nhất 3 mã vùng.

1.2.3. Một số ứng dụng của nguyên lý Dirichlet

Trong nhiều ứng dụng thú vị của nguyên lý Dirichlet, khái niệm đồ vật và hộp cần phải được lựa chọn một cách khôn khéo. Trong phần này có vài thí dụ như vậy.

Ví dụ 1.13. Trong một phòng họp có n người, bao giờ cũng tìm được 2 người có số người quen trong số những người dự họp là như nhau. Số người quen của mỗi người trong phòng họp nhận các giá trị từ 0 đến $n - 1$.

Lời giải. Rõ ràng trong phòng không thể đồng thời có người có số người quen là 0 (tức là không quen ai) và có người có số người quen là $n - 1$ (tức là quen tất cả). Vì vậy theo số lượng người quen, ta chỉ có thể phân n người ra thành $n - 1$ nhóm. Vậy theo nguyên lý Dirichlet tồn tại một nhóm có ít nhất 2 người, tức là luôn tìm được ít nhất 2 người có số người quen là như nhau. ■

1. Giải tích tổ hợp

Ví dụ 1.14. Trong một tháng gồm 30 ngày, một đội bóng chuyền thi đấu mỗi ngày ít nhất 1 trận nhưng chơi không quá 45 trận. Chứng minh rằng tìm được một giai đoạn gồm một số ngày liên tục nào đó trong tháng sao cho trong giai đoạn đó đội chơi đúng 14 trận.

Lời giải. Gọi a_j là số trận mà đội đã chơi từ ngày đầu tháng đến hết ngày j . Khi đó

$$1 \leq a_1 < a_2 < \dots < a_{30} < 45$$

$$15 \leq a_1 + 14 < a_2 + 14 < \dots < a_{30} + 14 < 59$$

Sáu mươi số nguyên $a_1, a_2, \dots, a_{30}, a_1 + 14, a_2 + 14, \dots, a_{30} + 14$ nằm giữa 1 và 59. Do đó theo nguyên lý Dirichlet có ít nhất 2 trong 60 số này bằng nhau. Vì vậy tồn tại i và j sao cho $a_i = a_j + 14 (j < i)$. Điều này có nghĩa là từ ngày $j + 1$ đến hết ngày i đội đã chơi đúng 14 trận. ■

Ví dụ 1.15. Chứng tỏ rằng trong $n + 1$ số nguyên dương không vượt quá $2n$, tồn tại ít nhất một số chia hết cho số khác.

Lời giải. Ta viết mỗi số nguyên a_1, a_2, \dots, a_{n+1} dưới dạng $a_j = 2^{k_j} q_j$ trong đó k_j là số nguyên không âm còn q_j là số dương lẻ nhỏ hơn $2n$. Vì chỉ có n số nguyên dương lẻ nhỏ hơn $2n$ nên theo nguyên lý Dirichlet tồn tại i và j sao cho $q_i = q_j = q$. Khi đó $a_i = 2^{k_i} q$ và $a_j = 2^{k_j} q$. Vì vậy, nếu $k_i \leq k_j$ thì a_j chia hết cho a_i còn trong trường hợp ngược lại ta có a_i chia hết cho a_j . ■

Thí dụ cuối cùng trình bày cách áp dụng nguyên lý Dirichlet vào lý thuyết tổ hợp mà vẫn quen gọi là lý thuyết Ramsey, tên của nhà toán học người Anh. Nói chung, lý thuyết Ramsey giải quyết những bài toán phân chia các tập con của một tập các phần tử.

Ví dụ 1.16. Giả sử trong một nhóm 6 người mỗi cặp hai hoặc là bạn hoặc là thù. Chứng tỏ rằng trong nhóm có ba người là bạn lẫn nhau hoặc có ba người là kẻ thù lẫn nhau.

Lời giải. Gọi A là một trong 6 người. Trong số 5 người của nhóm hoặc là có ít nhất ba người là bạn của A hoặc có ít nhất ba người là kẻ thù của A , điều này suy ra từ nguyên lý Dirichlet tổng quát, vì $\lceil \frac{5}{2} \rceil = 3$. Trong trường hợp đầu ta gọi B, C, D là bạn của A . Nếu trong ba người này có hai người là bạn thì họ cùng với A lập thành một bộ ba người bạn lẫn nhau, ngược lại, tức là nếu trong ba người B, C, D không có ai là bạn ai cả thì chứng tỏ họ là bộ ba người thù lẫn nhau. Tương tự có thể chứng minh trong trường hợp có ít nhất ba người là kẻ thù của A . ■

1.3

Chỉnh hợp và tổ hợp

1.3.1. Chính hợp có lặp

Một cách sắp xếp có thứ tự k phần tử có thể lặp lại của một tập n phần tử được gọi là một chỉnh hợp lặp chập k từ tập n phần tử. Nếu A là tập gồm n phần tử đó thì mỗi chỉnh hợp như

1. Giải tích tổ hợp

thế là một phần tử của tập A_n^k . Ngoài ra, mỗi chỉnh hợp lặp chap k từ tập n phần tử là một hàm từ tập k phần tử vào tập n phần tử. Vì vậy số chỉnh hợp lặp chap k từ tập n phần tử là n^k .

1.3.2. Tổ hợp có lặp.

Một tổ hợp lặp chap k của một tập hợp là một cách chọn không có thứ tự k phần tử có thể lặp lại của tập đã cho. Như vậy một tổ hợp lặp kiểu này là một dãy không kể thứ tự gồm k thành phần lấy từ tập n phần tử. Do đó có thể là $k > n$.

Định lý 1.3. Số tổ hợp lặp chap k từ tập n phần tử bằng $\binom{k}{n+k-1}$.

Chứng minh. Mỗi tổ hợp lặp chap k từ tập n phần tử có thể biểu diễn bằng một dãy n ô và k ngôi sao. Ô thứ i chứa thêm một ngôi sao mỗi lần khi phần tử thứ i của tập xuất hiện trong tổ hợp. Chẳng hạn, tổ hợp lặp chap 6 của 4 phần tử được biểu thị bằng một dãy gồm 4 ô.

**	*		***
----	---	--	-----

Hình trên mô tả tổ hợp chứa đúng 2 phần tử thứ nhất, 1 phần tử thứ hai, không có phần tử thứ 3 và 3 phần tử thứ tư của tập hợp. Mỗi dãy n ô và k ngôi sao ứng với một chuỗi nhị phân độ dài $n+k-1$ với k số 1. Do đó số các dãy n ô và k ngôi sao chính là số tổ hợp chap k từ tập $n+k-1$ phần tử. Đó là điều cần chứng minh. ■

Ví dụ 1.17. Có bao nhiêu cách chọn 5 tờ giấy bạc từ một két đựng tiền gồm những tờ 1000đ, 2000đ, 5000đ, 10.000đ, 20.000đ, 50.000đ, 100.000đ.

Lời giải. Giả sử thứ tự mà các tờ tiền được chọn là không quan trọng, các tờ tiền cùng loại là không phân biệt và mỗi loại có ít nhất 5 tờ. Vì ta không kể tới thứ tự chọn tờ tiền và vì ta chọn đúng 5 lần, mỗi lần lấy một tờ trong 7 loại tiền nên mỗi cách chọn 5 tờ giấy bạc này chính là một tổ hợp lặp chap 5 từ 7 phần tử. Do đó số cần tìm là $\binom{5}{5+7-1} = 462$. ■

Ví dụ 1.18. Phương trình $x_1 + x_2 + x_3 = 15$ có bao nhiêu nghiệm nguyên không âm?

Lời giải. Chúng ta nhận thấy mỗi nghiệm của phương trình ứng với một cách chọn 15 phần tử từ một tập có 3 loại, sao cho có x_1 phần tử loại 1, x_2 phần tử loại 2 và x_3 phần tử loại 3 được chọn. Vì vậy số nghiệm bằng số tổ hợp lặp chap 15 từ tập có 3 phần tử và bằng $\binom{15}{15+3-1} = 136$. ■

1.3.3. Hoán vị của tập hợp có các phần tử giống nhau

Trong bài toán đếm, một số phần tử có thể giống nhau. Khi đó cần phải cẩn thận, tránh đếm chúng hơn một lần. Ta xét thí dụ sau.

Ví dụ 1.19. Có thể nhận được bao nhiêu chuỗi khác nhau bằng cách sắp xếp lại các chữ cái của từ SUCCESS?

1. Giải tích tổ hợp

Lời giải. Vì một số chữ cái của từ SUCCESS là như nhau nên câu trả lời không phải là số hoán vị của 7 chữ cái được. Từ này chứa 3 chữ S, 2 chữ C, 1 chữ U và 1 chữ E. Để xác định số chuỗi khác nhau có thể tạo ra được ta nhận thấy có $\binom{3}{7}$ cách chọn 3 chỗ cho 3 chữ S, còn lại 4 chỗ trống. Có $\binom{2}{4}$ cách chọn 2 chỗ cho 2 chữ C, còn lại 2 chỗ trống. Có thể đặt chữ U bằng $\binom{1}{2}$ cách và $\binom{1}{1}$ cách đặt chữ E vào chuỗi. Theo nguyên lý nhân, số các chuỗi khác nhau có thể tạo được là:

$$\binom{3}{7} \binom{2}{4} \binom{1}{2} \binom{1}{1} = \frac{7!4!2!1!}{3!4!2!2!1!1!0!} = \frac{7!}{3!2!1!1!} = 420$$

■

Định lý 1.4. Số hoán vị của n phần tử trong đó có n_1 phần tử như nhau thuộc loại 1, n_2 phần tử như nhau thuộc loại 2,..., và n_k phần tử như nhau thuộc loại k , là

$$\frac{n!}{n_1!n_2!\dots n_k!}$$

Chứng minh. Để xác định số hoán vị trước tiên chúng ta nhận thấy có $\binom{n_1}{n}$ cách giữ n_1 chỗ cho n_1 phần tử loại 1, còn lại $n - n_1$ chỗ trống. Sau đó có $\binom{n_2}{n-n_1}$ cách đặt n_2 phần tử loại 2 vào hoán vị, còn lại $n - n_1 - n_2$ chỗ trống. Tiếp tục đặt các phần tử loại 3, loại 4,..., loại $k - 1$ vào chỗ trống trong hoán vị. Cuối cùng có $\binom{n_k}{n-n_1-n_2-\dots-n_{k-1}}$ cách đặt n_k phần tử loại k vào hoán vị. Theo quy tắc nhân tất cả các hoán vị là

$$\binom{n_1}{n} \binom{n_2}{n-n_1} \dots \binom{n_k}{n-n_1-\dots-n_{k-1}} = \frac{n!}{n_1!n_2!\dots n_k!}$$

■

1.3.4. Sự phân bố các đồ vật vào trong hộp

Ví dụ 1.20. Có bao nhiêu cách chia những xấp bài 5 quân cho mỗi một trong 4 người chơi từ một cỗ bài chuẩn 52 quân?

Lời giải. Người đầu tiên có thể nhận được 5 quân bài bằng $\binom{5}{52}$ cách. Người thứ hai có thể nhận được chia 5 quân bài bằng $\binom{5}{47}$ cách, vì chỉ còn 47 quân bài. Người thứ ba có thể nhận được 5 quân bài bằng $\binom{5}{42}$ cách. Cuối cùng, người thứ tư nhận được 5 quân bài bằng $\binom{5}{37}$ cách. Vì vậy, theo nguyên lý nhân tổng cộng có

$$\binom{5}{52} \binom{5}{47} \binom{5}{42} \binom{5}{37} = \frac{52!}{5!5!5!32!}$$

cách chia cho 4 người mỗi người một xấp 5 quân bài.

■

Thí dụ trên là một bài toán điển hình về việc phân bố các đồ vật khác nhau vào các hộp khác nhau. Các đồ vật là 52 quân bài, còn 4 hộp là 4 người chơi và số còn lại để trên bàn. Số cách sắp xếp các đồ vật vào trong hộp được cho bởi mệnh đề sau

1. Giải tích tổ hợp

Định lý 1.5. Số cách phân chia n đồ vật khác nhau vào trong k hộp khác nhau sao cho có n_i vật được đặt vào trong hộp thứ i , với $i = 1, 2, \dots, k$ là

$$\frac{n!}{n_1!n_2!\dots n_k!(n - n_1 - n_2 - \dots - n_k)!}$$

1.4

Sinh các hoán vị và tổ hợp

1.4.1. Sinh các hoán vị

Có nhiều thuật toán đã được phát triển để sinh ra $n!$ hoán vị của tập $\{1, 2, \dots, n\}$. Ta sẽ mô tả một trong các phương pháp đó, phương pháp liệt kê các hoán vị của tập $\{1, 2, \dots, n\}$ theo thứ tự từ điển.

Khi đó, hoán vị $a_1a_2\dots a_n$ được gọi là đi trước hoán vị $b_1b_2\dots b_n$ nếu tồn tại k ($1 \leq k \leq n$), $a_1 = b_1, a_2 = b_2, \dots, a_{k-1} = b_{k-1}$ và $a_k < b_k$.

Thuật toán sinh các hoán vị của tập $\{1, 2, \dots, n\}$ dựa trên thủ tục xây dựng hoán vị kế tiếp, theo thứ tự từ điển, từ hoán vị cho trước $a_1a_2\dots a_n$.

Đầu tiên nếu $a_{n-1} < a_n$ thì rõ ràng đổi chỗ a_{n-1} và a_n cho nhau thì sẽ nhận được hoán vị mới đi liền sau hoán vị đã cho. Nếu tồn tại các số nguyên a_j và a_{j+1} sao cho $a_j < a_{j+1}$ và $a_{j+1} > a_{j+2} > \dots > a_n$, tức là tìm cặp số nguyên liền kề đầu tiên tính từ bên phải sang bên trái của hoán vị mà số đầu nhỏ hơn số sau. Sau đó, để nhận được hoán vị liền sau ta đặt vào vị trí thứ j số nguyên nhỏ nhất trong các số lớn hơn a_j của tập $a_{j+1}, a_{j+2}, \dots, a_n$, rồi liệt kê theo thứ tự tăng dần của các số còn lại của $a_j, a_{j+1}, a_{j+2}, \dots, a_n$ vào các vị trí $j+1, \dots, n$. Để thấy không có hoán vị nào đi sau hoán vị xuất phát và đi trước hoán vị vừa tạo ra.

Ví dụ 1.21. Tìm hoán vị liền sau theo thứ tự từ điển của hoán vị 4736521.

Lời giải. Cặp số nguyên đầu tiên tính từ phải qua trái có số trước nhỏ hơn số sau là $a_3 = 3$ và $a_4 = 6$. Số nhỏ nhất trong các số bên phải của số 3 mà lại lớn hơn 3 là số 5. Đặt số 5 vào vị trí thứ 3. Sau đó đặt các số 3, 6, 1, 2 theo thứ tự tăng dần vào bốn vị trí còn lại. Hoán vị liền sau hoán vị đã cho là 4751236. ■

Mã python phát sinh hoán vị liền sau:

```
def DoiCho(a, i, j):
    a[i], a[j] = a[j], a[i]

def HoanViLienSau(a):
    n = len(a)
    j = n - 2
    while a[j] > a[j+1]:
        j -= 1
```

1. Giải tích tổ hợp

```
k = n - 1
while a[j] > a[k]:
    k -= 1
DoiCho(a, j, k)
r = n - 1
l = j + 1
while l < r:
    DoiCho(a, l, r)
    l += 1
    r -= 1

a = [4, 7, 3, 6, 5, 2, 1]
HoanViLienSau(a)
print(a)
```

kết quả chạy sẽ in ra màn hình

```
[4, 7, 5, 1, 2, 3, 6]
```

1.4.2. Sinh các tổ hợp

Làm thế nào để tạo ra tất cả các tổ hợp các phần tử của một tập hữu hạn? Vì tổ hợp chính là một tập con, nên ta có thể dùng phép tương ứng 1-1 giữa các tập con của $\{a_1, a_2, \dots, a_n\}$ và chuỗi nhị phân độ dài n . Ta thấy một chuỗi nhị phân độ dài n cũng là khai triển nhị phân của một số nguyên nằm giữa 0 và $2^n - 1$. Khi đó 2ⁿ chuỗi nhị phân có thể liệt kê theo thứ tự tăng dần của số nguyên trong biểu diễn nhị phân của chúng. Chúng ta sẽ bắt đầu từ chuỗi nhị phân nhỏ nhất 00...00 (n số 0). Mỗi bước để tìm chuỗi liền sau ta tìm vị trí đầu tiên tính từ phải qua trái mà ở đó là số 0, sau đó thay tất cả số 1 ở bên phải số này bằng 0 và đặt số 1 vào chính vị trí này.

Mã python để phát sinh chuỗi nhị phân liên sau:

```
def ChuoiNhiPhanLienSau(b):
    i = 0
    while b[i] == 1:
        b[i] = 0
        i += 1
    b[i] = 1

b = [0, 1, 1, 0, 0]
ChuoiNhiPhanLienSau(b)
print(b)
```

kết quả chạy sẽ in ra màn hình

```
[1, 1, 1, 0, 0]
```

1. Giải tích tổ hợp

Tiếp theo chúng ta sẽ trình bày thuật toán tạo các tổ hợp chập k từ n phần tử $\{1, 2, \dots, n\}$. Mỗi tổ hợp chập k có thể biểu diễn bằng một chuỗi tăng. Khi đó có thể liệt kê các tổ hợp theo thứ tự từ điển. Có thể xây dựng tổ hợp liền sau tổ hợp $a_1a_2\dots a_k$ bằng cách sau. Trước hết, tìm phần tử đầu tiên a_i trong dãy đã cho kể từ phải qua trái sao cho $a_i \neq n - k + i$. Sau đó thay a_i bằng a_{i+1} và a_j bằng $a_i + j - i + 1$ với $j = i + 1, i + 2, \dots, k$.

Ví dụ 1.22. Tìm tổ hợp chập 4 từ tập $\{1, 2, 3, 4, 5, 6\}$ đi liền sau tổ hợp $\{1, 2, 5, 6\}$.

Lời giải. Ta thấy từ phải qua trái $a_2 = 2$ là số hạng đầu tiên của tổ hợp đã cho thỏa mãn điều kiện $a_i \neq 6 - 4 + i$. Để nhận được tổ hợp tiếp sau ta tăng a_i lên một đơn vị, tức $a_2 = 3$, sau đó đặt $a_3 = 3 + 1 = 4$ và $a_4 = 3 + 2 = 5$. Vậy tổ hợp liền sau tổ hợp đã cho là $\{1, 3, 4, 5\}$. ■

Thủ tục này được cài đặt bằng mã python như sau:

```
def ToHopLienSau(a, n, k):
    i = k - 1
    while a[i] == n - k + i + 1:
        i -= 1
    a[i] = a[i] + 1
    for j in range(i + 1, k):
        a[j] = a[i] + j - i

a = [1, 2, 5, 6]
ToHopLienSau(a, 6, 4)
print(a)
```

kết quả chạy sẽ in ra màn hình

```
[1, 3, 4, 5]
```

Tóm tắt chương

Chương này đã trình bày những vấn đề cơ bản nhất của lý thuyết tổ hợp, bao gồm:

- Việc sử dụng hai nguyên lý đếm căn bản là nguyên lý cộng và nguyên lý nhân để giải quyết những bài toán cơ bản
- Vận dụng các công thức liên quan đến chỉnh hợp và tổ hợp để đếm trong các trường hợp phức tạp hơn
- Áp dụng nguyên lý Dirichlet để giải quyết những bài toán liên quan đến vấn đề tồn tại

1. Giải tích tổ hợp

Bài tập

Tập hợp

B 1.1. Cho dãy tập hợp $A_1, A_2, \dots, A_n, \dots$. Chứng minh rằng luôn luôn tồn tại dãy tập hợp $B_1, B_2, \dots, B_n, \dots$, sao cho:

1. Các B_i từng đôi một rời nhau;
2. $\cup_{i=1}^{\infty} A_i = \cup_{k=1}^{\infty} B_k$.

B 1.2. Chứng minh rằng các hệ thức sau đây tương đương nếu A và B là tập hợp con của Ω :
 $A \cup B = \Omega, A \subset B, B \subset A$.

B 1.3. Khẳng định cho rằng nếu A, B, C là tập hợp con của tập hợp Ω sao cho $A \subset B \cup C$ và $B \subset A \cup C$, thì $B = \emptyset$, có đúng không?

B 1.4. Chứng minh rằng nếu A, B, C là các tập hợp con của tập hợp Ω , sao cho $A \cap B \subset C$ và $A \cup C \subset B$, thì $A \cap C = \emptyset$

B 1.5. Tìm biểu thức đơn giản của các biểu thức sau:

1. $(A \cup B)(A \cup C)$
2. $(A \cup B)(A \cup \overline{B})$
3. $(A \cup B)(\overline{A} \cup B)(A \cup \overline{B})$
4. $(A \cup B)(\overline{A} \cup B)(\overline{A} \cup \overline{B})$
5. $(A \cup B)(B \cup C)$

B 1.6. Hệ thức nào trong các hệ thức sau đây đúng

1. $A \cup B \cup C = A \cup (B \setminus AB) \cup (C \setminus AC)$
2. $A \cup B = (A \setminus AB) \cup B$
3. $(A \cup B) \setminus A = B$
4. $(A \cup B) \setminus C = A \cup (B \setminus C)$
5. $ABC = AB(C \cup B)$
6. $AB \cup BC \cup CA \supset ABC$
7. $(AB \cup BC \cup CA) \subset (A \cup B \cup C)$
8. $ABC \subset A \cup B$

1. Giải tích tổ hợp

9. $A \cup BC = AC \cup BC$
10. $A \cup BC = C \setminus (C(A \cup B))$

B 1.7. Chứng minh rằng:

1. $\overline{A \cup \overline{B}} \cup \overline{\overline{A} \cup B} = A$
2. $(A \cup B)\overline{AB} = A\overline{B} \cup B\overline{A}$

B 1.8. Chứng minh

1. Nếu $A \cup B = AB$ thì $A = B$
2. $A \cup BC \supset (A \cup B)C$
3. Nếu $A_1 \subset A, B_1 \subset B$ và $A \cap B = \emptyset$ thì $A_1 \cap B_1 = \emptyset$

Giải tích tổ hợp

B 1.9. Một lô hàng có 50 sản phẩm.

1. Có bao nhiêu cách chọn ngẫu nhiên cùng lúc 5 sản phẩm để kiểm tra?
2. Có bao nhiêu cách chọn ngẫu nhiên lần lượt 5 sản phẩm?

B 1.10. Trong một hệ thống điện thoại nội bộ 3 số

1. có bao nhiêu máy có các chữ số khác nhau?
2. Có bao nhiêu máy có số 9 ở cuối còn các chữ số còn lại đều khác nhau?

B 1.11. Một lớp học có 40 học sinh gồm 20 nam và 20 nữ. Có bao nhiêu cách chia để trong mỗi nửa lớp có 10 nam sinh và 10 nữ sinh?

B 1.12. Nếu một người có 6 đôi vớ khác nhau và 4 đôi giày khác nhau. Có bao nhiêu cách kết hợp giữa vớ và giày?

B 1.13. Năm người A, B, C, D, E sẽ phát biểu trong một hội nghị. Có bao nhiêu cách sắp xếp để:

1. Người B phát biểu sau A.
2. Người A phát biểu xong thì đến lượt B.

B 1.14. Có 6 học sinh được sắp xếp ngồi vào 6 chỗ đã ghi số thứ tự trên một bàn dài. Tìm số cách xếp

1. 6 học sinh vào bàn.

1. Giải tích tổ hợp

2. 6 học sinh này vào bàn sao cho 2 học sinh A, B ngồi cạnh nhau.
3. 6 học sinh này ngồi vào bàn sao cho 2 học sinh A, B không ngồi cạnh nhau.

B 1.15. Một lớp có 40 học sinh. Giáo viên chủ nhiệm muốn chọn ra một ban cán sự lớp: 1 lớp trưởng, 1 lớp phó, 1 thủ quỹ. Hỏi giáo viên chủ nhiệm có bao nhiêu cách chọn ban cán sự lớp?

B 1.16. Một hộp có 8 bi đỏ, 6 bi trắng, 4 bi vàng. Người ta chọn ra 6 bi từ hộp đó. Hỏi có bao nhiêu cách chọn nêu:

1. Không yêu cầu gì thêm.
2. Phải có 2 bi đỏ, 2 bi trắng, 2 bi vàng.
3. Có đúng 2 bi vàng.

B 1.17. Một đồn cảnh sát khu vực có 9 người. Trong ngày cần cử 3 người làm nhiệm vụ ở địa điểm A, 2 người ở địa điểm B còn 4 người trực tại đồn. Hỏi có bao nhiêu cách phân công?

B 1.18. Một tổ sản xuất có 12 người, trong đó có 4 nữ, cần chia thành 4 nhóm đều nhau. Hãy tìm số cách phân chia sao cho mỗi nhóm có 1 nữ?

B 1.19. Xếp 12 hành khách lên 4 toa tàu. Tìm số cách sắp xếp:

1. Mỗi toa có 3 hành khách.
2. Một toa có 6 hành khách, một toa có 4 hành khách, 2 toa còn lại mỗi toa có 1 hành khách.

B 1.20. Có bao nhiêu chuỗi 8 bit đối xứng?

B 1.21. Có 6 sách tin học, 4 sách toán học, 3 sách vật lý

1. Có bao nhiêu cách sắp xếp chúng lên giá sách?
2. Có bao nhiêu cách sắp xếp chúng lên giá sách sao cho tất cả các sách cùng nhóm thì được xếp kề nhau?
3. Có bao nhiêu cách sắp xếp để 2 sách toán không kề nhau?

B 1.22. Một tổ chức gồm 7 nam và 8 nữ

1. Có bao nhiêu cách chọn một hội đồng gồm 5 người?
2. Có bao nhiêu cách chọn hội đồng gồm 2 nam và 3 nữ?
3. Có bao nhiêu cách chọn hội đồng gồm 4 người và phải có ít nhất 1 nữ?
4. Có bao nhiêu cách chọn hội đồng gồm 5 người và phải có cả nam và nữ?

1. Giải tích tổ hợp

B 1.23. Trong tổng số 2504 sinh viên của một khoa công nghệ thông tin, có 1876 theo học môn ngôn ngữ lập trình Pascal, 999 học môn ngôn ngữ Fortran và 345 học ngôn ngữ C. Ngoài ra còn biết 876 sinh viên học cả Pascal và Fortran, 232 học cả Fortran và C, 290 học cả Pascal và C. Nếu 189 sinh viên học cả 3 môn Pascal, Fortran và C thì trong trường hợp đó có bao nhiêu sinh viên không học môn nào trong 3 môn ngôn ngữ lập trình kể trên.

B 1.24. Một cuộc họp gồm 12 người tham dự để bàn về 3 vấn đề. Có 8 người phát biểu về vấn đề I, 5 người phát biểu về vấn đề II và 7 người phát biểu về vấn đề III. Ngoài ra, có đúng 1 người không phát biểu vấn đề nào. Hỏi nhiều lầm là có bao nhiêu người phát biểu cả 3 vấn đề.

B 1.25. Chỉ ra rằng có ít nhất 4 người trong số 25 triệu người có cùng tên họ viết tắt bằng 3 chữ cái sinh cùng ngày trong năm (không nhất thiết trong cùng một năm).

B 1.26. Một tay đô vật tham gia thi đấu giành chức vô địch trong 75 giờ. Mỗi giờ anh ta có ít nhất một trận đấu, nhưng toàn bộ anh ta có không quá 125 trận. Chứng tỏ rằng có những giờ liên tiếp anh ta đã đấu đúng 24 trận.

B 1.27. Cho n là số nguyên dương bất kỳ. Chứng minh rằng luôn lấy ra được từ n số đã cho một số số hạng thích sao cho tổng của chúng chia hết cho n .

B 1.28. Trong một cuộc lấy ý kiến về 7 vấn đề, người được hỏi ghi vào một phiếu trả lời sẵn bằng cách để nguyên hoặc phủ định các câu trả lời tương ứng với 7 vấn đề đã nêu. Chứng minh rằng với 1153 người được hỏi luôn tìm được 10 người trả lời giống hệt nhau.

B 1.29. Có 17 nhà bác học viết thư cho nhau trao đổi 3 vấn đề. Chứng minh rằng luôn tìm được 3 người cùng trao đổi một vấn đề.

B 1.30. Trong kỳ thi kết thúc học phần toán học rời rạc có 10 câu hỏi. Có bao nhiêu cách gán điểm cho các câu hỏi nếu tổng số điểm bằng 100 và mỗi câu ít nhất được 5 điểm.

B 1.31. Phương trình $x_1 + x_2 + x_3 + x_4 + x_5 = 21$ có bao nhiêu nghiệm nguyên không âm?

B 1.32. Có bao nhiêu số nguyên trong tập $\{1, 2, \dots, 100000\}$ có tổng chữ số là 16

B 1.33. Có bao nhiêu chuỗi khác nhau có thể lập được từ các chữ cái trong từ MISSISSIPI, yêu cầu phải dùng tất cả các chữ?

B 1.34. Một giáo sư cất bộ sưu tập gồm 40 số báo toán học vào 4 chiếc ngăn tủ, mỗi ngăn đựng 10 số. Có bao nhiêu cách có thể cất các tờ báo vào các ngăn nếu:

1. Mỗi ngăn được đánh số sao cho có thể phân biệt được;
2. Các ngăn là giống hệt nhau?

B 1.35. Có thể nối năm máy tính với nhau sao cho có chính xác hai máy tính được nối trực tiếp đến cùng số máy? Hãy giải thích lý do.

1. Giải tích tổ hợp

B 1.36. Một bàn cờ kích thước 3×7 gồm 21 ô hình vuông, mỗi hình vuông được tô màu đen hoặc màu trắng. Chứng minh rằng bàn cờ chứa một hình chữ nhật không *tầm thường* (có kích thước $1 \times k$ hoặc $k \times 1$) sao cho 4 hình vuông ở bốn góc hoặc tất cả tô màu đen hoặc tất cả tô màu trắng.

B 1.37. Tìm hệ thức truy hồi cho số mất thứ tự \bar{N} .

B 1.38. Tìm hệ thức truy hồi cho số các chuỗi nhị phân chứa chuỗi 01.

B 1.39. Tìm hệ thức truy hồi cho số cách đi lên n bậc thang nếu một người có thể bước một, hai hoặc ba bậc một lần.

B 1.40. Tìm hệ thức truy hồi mà R_n thoả mãn, trong đó R_n là số miền của mặt phẳng bị phân chia bởi n đường thẳng nếu không có hai đường nào song song và không có 3 đường nào cùng đi qua một điểm. Tính R_n bằng phương pháp lặp.

B 1.41. Tìm nghiệm của hệ thức truy hồi $a_n = 2a_{n-1} + 5a_{n-2} - 6a_{n-3}$ với $a_0 = 7, a_1 = -4, a_2 = 8$.

B 1.42. Giả sử m, n, r là các số nguyên dương. Chứng minh rằng

$$C_m^0 C_{n-m}^r + C_m^1 C_{n-m}^{r-1} + \dots + C_m^r C_{n-m}^0 = C_n^r$$

B 1.43. Chứng minh rằng

$$1. \quad C_n^1 + 2C_n^2 + \dots + nC_n^n = n2^{n-1}$$

$$2. \quad 2.1.C_n^2 + 3.2C_n^3 + \dots + n.(n-1).C_n^n = n(n-1)2^{n-2}$$

B 1.44. Cho m, n, r là các số nguyên dương. Chứng minh rằng

$$1. \quad \sum_{k=0}^m C_{n-k}^r = C_{n+1}^{r+1} - C_{n-m}^{r+1}$$

$$2. \quad \sum_{k=0}^m (-1)^k C_n^k = (-1)^m C_{n-1}^m$$

B 1.45. Chứng minh rằng

$$(C_n^0)^2 + (C_n^1)^2 + \dots + (C_n^n)^2 = C_{2n}^n$$

B 1.46. Chứng minh rằng

$$\sum_{k=0}^n \frac{2n!}{k!^2(n-k)!^2} = (C_{2n}^n)^2$$

□ Lập trình

B 1.47. Viết chương trình in ra tất cả các hoán vị của n số tự nhiên $\{1, 2, \dots, n\}$

B 1.48. Viết chương trình in ra tất cả các chỉnh hợp k của n số tự nhiên $\{1, 2, \dots, n\}$

B 1.49. Viết chương trình in ra tất cả các tổ hợp k của n số tự nhiên $\{1, 2, \dots, n\}$

Lý thuyết xác suất là công cụ cơ bản và là tiền đề cho các lĩnh vực khoa học khác.

2.1 Khái niệm xác suất

2.1.1. Phép thử và sự kiện

Khái niệm 2.1. *Phép thử là một thử nghiệm cho kết quả là một sự kiện hay biến cố (event).* Ví dụ, tung một con xúc xắc 6 mặt được coi là một phép thử, kết quả thu được là xuất hiện mặt 1 chấm, 2 chấm, ... 6 chấm, và các kết quả này được gọi là các sự kiện thu được từ phép thử tung con xúc xắc.

Như vậy ta có thể phân sự kiện thành 3 dạng chính sau:

- *Sự kiện chắc chắn:* là sự kiện luôn luôn xảy ra
- *Sự kiện bất khả:* là sự kiện không bao giờ xảy ra
- *Sự kiện ngẫu nhiên:* là sự kiện có thể xảy ra hoặc không

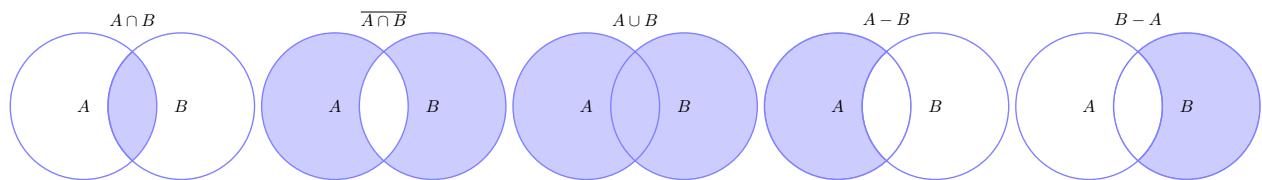
Các sự kiện trong cùng một phép thử có thể có những quan hệ chính sau:

- *Sự kiện đối:* là 2 sự kiện không xảy ra đồng thời. Sự kiện đối của A được kí hiệu là \bar{A} . Sự kiện này còn được gọi là sự kiện bù của A và được kí hiệu là A^C .
- *Sự kiện hợp:* là sự kiện xảy ra khi có ít nhất một trong những sự kiện thành phần xảy ra. Sự kiện hợp của A và B được kí hiệu là $\bigcup_{i=1}^n A_i$ hoặc $A + B$. Trường hợp tổng quát, hợp của các sự kiện $\{A_i\}, i \in \{1, 2, \dots, n\}$ là $\bigcup_{i=1}^n A_i$ hoặc $\sum_{i=1}^n A_i$.
- *Sự kiện giao:* Là sự kiện xảy ra khi tất cả các sự kiện thành phần cùng xảy ra. Giao của 2 sự kiện A và B được kí hiệu là $\bigcap_{i=1}^n A_i$ hoặc AB . Trường hợp tổng quát, giao của các sự kiện $\{A_i\}, i \in \{1, 2, \dots, n\}$ là $\bigcap_{i=1}^n A_i$ hoặc $\prod_{i=1}^n A_i$.
- *Sự kiện xung khắc:* Là các sự kiện không đồng thời xảy ra. Các sự kiện $\{A_i\}, i \in \{1, 2, \dots, n\}$ xung khắc đôi một khi và chỉ khi $\bigcap_{i=1}^n A_i = \emptyset$.

2. Xác suất

- **Sự kiện độc lập:** các sự kiện được gọi là độc lập khi và chỉ khi việc xảy ra sự kiện này không ảnh hưởng tới việc xảy ra tập sự kiện còn lại. Như vậy có thể thấy nếu 2 sự kiện A, B là độc lập thì $A, \bar{B}; \bar{A}, B; \bar{A}, \bar{B}$ cũng là độc lập.

Khái niệm 2.2. Không gian sự kiện là tập hợp của tất cả các sự kiện độc lập có thể xảy ra.
Không gian sự kiện được kí hiệu là Ω .



Tính chất. Các phép toán trên các sự kiện.

1. Giao hoán:

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

2. Kết hợp:

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$$A \cap (B \cap C) = (A \cap B) \cap C$$

3. Phân phối:

$$A \cap (B \cup C) = A \cap B \cup A \cap C$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

4. Phản bù:

$$\overline{\overline{A}} = A$$

$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$

2. Xác suất

2.1.2. Định nghĩa xác suất

Khái niệm 2.3. *Không gian xác suất* (probability space) là bộ ba (Ω, \mathcal{F}, P)

- Ω là tập không rỗng, không gian mẫu hay không gian sự kiện.
- \mathcal{F} là một tập hợp mà các thành viên của nó được gọi là các “biến cố” (tập con của Ω).
- P là một độ đo xác suất.

thỏa được các tính chất

1. Xác suất của sự kiện A bất kì luôn nằm trong khoảng 0, 1:

$$P(A) \in [0, 1] \quad \forall A \subseteq \Omega \text{ hay } A \in \mathcal{F}$$

2. Xác suất của sự kiện bất khả bằng 0:

$$P(\emptyset) = 0$$

3. Xác suất của sự kiện chắc chắn hay không gian sự kiện bằng 1:

$$P(\Omega) = 1$$

4. Xác suất hợp của 2 sự kiện độc lập A, B là tổng của chúng:

$$P(A \cup B) = P(A) + P(B)$$

5. Xác suất kéo theo $A \subseteq B$ thì:

$$P(A) \leq P(B)$$

Khái niệm 2.4. *Tỉ lệ cờ hối* (odds) của một sự kiện A là tỉ lệ giữa xác suất xảy ra sự kiện A (p) và xác suất không xảy ra sự kiện A ($1 - p$)

$$\text{odds}(A) = \frac{p}{1 - p} \tag{2.1}$$

Lưu ý: $\text{odds}(A) \in [0, \infty)$

Định nghĩa đồng khả năng

Khái niệm 2.5. Xét một không gian sự kiện Ω gồm n sự kiện sơ cấp độc lập đồng khả năng $\{A_1, \dots, A_n\}$, sự kiện A bao gồm m sự kiện sơ cấp khi đó xác suất của sự kiện A được tính bằng

$$P(A) = \frac{m}{n} \tag{2.2}$$

2. Xác suất

Ví dụ. Tìm xác suất khi đẻ khi gieo một con xúc xắc thì xuất hiện mặt chẵn.

Lời giải. Giả sử con xúc xắc là cân xứng thì không gian sự kiện $\Omega = \{A_1, A_2, A_3, A_4, A_5, A_6\}$ bao gồm $n = 6$ sự kiện sơ cấp độc lập và đồng khả năng; trong đó, A_i là sự kiện xuất hiện mặt i . Gọi A là sự kiện xuất hiện mặt chẵn vậy

$$A = A_2 + A_4 + A_6 \implies m = 3$$

Do đó, xác suất

$$P(A) = \frac{m}{n} = \frac{3}{6} = 0.5$$

■

Định nghĩa theo hình học

Trong thực tế, nhiều bài toán có không gian sự kiện Ω là một tập hợp vô hạn các sự kiện sơ cấp độc lập đồng khả năng. Nếu Ω có thể được biểu diễn bằng miền hình học thì

Khái niệm 2.6. Xác suất sự kiện $A \subseteq \Omega$ được tính bằng

$$P(A) = \frac{\text{Diện tích}(A)}{\text{Diện tích}(\Omega)} \quad (2.3)$$

Ví dụ. Hai người A và B hẹn gặp nhau ở một địa điểm từ lúc 12h đến 13h. Họ thống nhất rằng người đến trước sẽ đợi người đến sau 20', nếu không gặp thì sẽ bỏ đi. Giả sử rằng thời gian đến của mỗi người là ngẫu nhiên. Tìm xác suất để A gặp B.

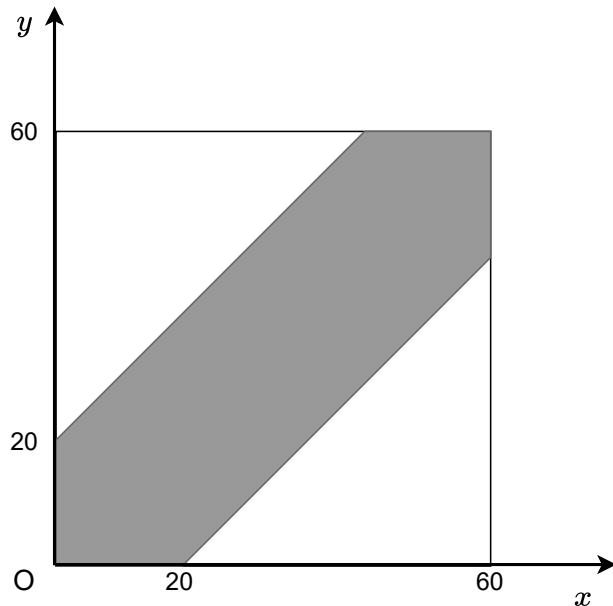
Lời giải. Giả sử x và y là thời điểm đến của A và B (tính từ 12h và đơn vị là phút), như vậy

$$0 \leq x \leq 60$$

$$0 \leq y \leq 60$$

Biểu diễn (x, y) trên mặt phẳng tọa độ xOy

2. Xác suất



- Tập hợp các điểm có thể xảy ra là hình vuông cạnh 60
- Tập hợp các điểm tương ứng với sự kiện A và B gắp nhau là

$$|x - y| \leq 20$$

phần hình tô màu xám

Vậy xác suất A và B gắp nhau là

$$p = \frac{60^2 - 40^2}{60^2} = \frac{5}{9}$$

■

Định nghĩa theo tần suất

Trong định nghĩa cổ điển, đòi hỏi tính đồng khả năng của các sự kiện độc lập sơ cấp. Đó là một yêu cầu khá ngặt nghèo, vì trong thực tế tính đồng khả năng thường bị vi phạm.

Khái niệm 2.7. *Tần suất* của một sự kiện A là tỉ lệ của số xuất hiện n_A của nó sau n lần thực hiện phép thử.

$$f_n(A) = \frac{n_A}{n} \quad (2.4)$$

Khái niệm 2.8. *Xác suất* (probability) của một sự kiện A là tần suất sự kiện khi số lần thử lên tới vô hạn.

$$P(A) = \lim_{n \rightarrow \infty} f_n(A) \quad (2.5)$$

2. Xác suất

Trên thực tế ta không đủ thời gian và điều kiện để thực hiện vô hạn số lần phép thử. Ta chấp nhận n đủ lớn thì tần số $f_n(A)$ sẽ tiến tới một giá trị gần như không biến thiên nhiều; nghĩa là $|P(A) - f_n(A)| < \epsilon$ với ϵ là một số dương rất bé.

Khái niệm 2.9. Xác suất của một sự kiện A là tần suất sự kiện khi số lần n là đủ lớn

$$P(A) \approx f_n(A) \quad (2.6)$$

Ví dụ. Giả sử thực nghiệm gieo một đồng xu 10 lần có kết quả là 4 lần mặt sấp và 6 lần mặt ngửa. Vậy xác suất để có mặt sấp khi gieo đồng xu sẽ là

$$P(\text{mặt sấp}) = \frac{4}{10}$$

2.2 Các phép toán xác suất

Tương tự như quan hệ của các sự kiện, ta cũng có quan hệ của các xác suất.

2.2.1. Tổng xác suất

Khái niệm 2.10. Cho tập sự kiện $\{A_i\}, i \in \{1, 2, \dots, n\}$, khi đó ta có **tổng xác suất** là xác suất của sự kiện hợp.

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n P(A_i) \\ &\quad - \sum_{1 \leq i_1 < i_2 \leq n} P(A_{i_1} A_{i_2}) \\ &\quad + \sum_{1 \leq i_1 < i_2 < i_3 \leq n} P(A_{i_1} A_{i_2} A_{i_3}) \\ &\quad - \dots + (-1)^{n+1} P(A_1 A_2 \dots A_n) \end{aligned} \quad (2.7)$$

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n (-1)^{i+1} \sum_{1 \leq k_1 < \dots < k_i \leq n} P\left(\bigcap_{j=1}^i A_{k_j}\right) \quad (2.8)$$

Trong đó, tổng $\sum_{1 \leq k_1 < \dots < k_i \leq n} P\left(\bigcap_{j=1}^i A_{k_j}\right)$ là tổng của tất cả các xác suất giao của tập con gồm i phần tử từ tập $\{1, 2, \dots, n\}$. Như vậy ta có thể thấy rằng mỗi tổng này sẽ gồm $\binom{n}{i} = \frac{n!}{i!(n-i)!}$ phần tử.

- Một số công thức cho trường hợp đơn giản

$$P(A + B) = P(A) + P(B) - P(AB)$$

$$P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

2. Xác suất

- Từ công thức tổng ta có thể thấy suy ra

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$$

- Dấu bằng đạt được khi tập sự kiện này xung khắc đôi một

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

- Nếu các sự kiện này tạo thành không gian sự kiện Ω thì

$$P(\Omega) = \sum_{i=1}^n P(A_i) = 1$$

- Hai sự kiện A và \bar{A} tạo thành không gian sự kiện Ω nên ta có

$$\begin{aligned} P(A) + P(\bar{A}) &= 1 \\ \iff P(A) &= 1 - P(\bar{A}) \\ \iff P(\bar{A}) &= 1 - P(A) \end{aligned}$$

2.2.2. Xác suất có điều kiện

Khái niệm 2.11. Là xác suất của một sự kiện xảy ra khi biết xác suất của sự kiện khác đã xảy ra. Xác suất của sự kiện A khi biết B đã xảy ra được kí hiệu là $P(A | B)$. Công thức tính xác suất có điều kiện

$$P(A | B) = \frac{P(AB)}{P(B)} \quad \forall P(B) > 0 \quad (2.9)$$

- Nếu A và B là độc lập, tức A không phụ thuộc vào B thì $P(A | B) = P(A)$ và $P(B | A) = P(B)$.
- Xác suất có điều kiện cũng có các tính chất hệt như xác suất thông thường

$$\begin{aligned} 1. \quad P\left(\bigcup_{i=1}^n A_i | B\right) &= \sum_{i=1}^n (-1)^{i-1} \sum_{k_1 \leq \dots \leq k_i} P\left(\bigcap_{j=1}^i A_{k_j} | B\right) \\ 2. \quad P(\bar{A} | B) &= 1 - P(A | B) \end{aligned}$$

2.2.3. Tích xác suất

Khái niệm 2.12. Tích xác suất là xác suất của sự kiện giao. Từ công thức xác suất có điều kiện ta có thể tính được xác suất giao như sau

$$P(AB) = P(B)P(A | B) = P(A)P(B | A) \quad (2.10)$$

2. Xác suất

Khái niệm 2.13. Trường hợp tổng quát, cho $\{A_i\}, i \in \{1, 2, \dots, n\}$ thì tích xác suất của chúng được tính như sau

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1)P(A_2 | A_1)P(A_3 | A_1A_2) \dots P(A_n | A_1A_2 \dots A_{n-1}) \quad (2.11)$$

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P\left(A_i | \bigcap_{j=1}^{i-1} A_j\right) \quad (2.12)$$

- Nếu $\{A_i\}$ là độc lập từng đôi một thì ta có:

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

- Do $0 \leq P(A_i) \leq 1$ nên xác suất của tích không thể nào lớn hơn xác suất thành phần được

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

2.2.4. Công thức Bayes

Khái niệm 2.14. Xác suất của A khi biết B được tính bằng công thức Bayes

$$P(A | B) = \frac{P(A)P(B | A)}{P(B)} \quad (2.13)$$

- Trường hợp mở rộng, cho **tập hợp sự kiện đầy đủ** $\{A_i\}, i \in \{1, 2, \dots, n\}$ ($P\left(\bigcup_{i=1}^n A_i\right) = 1$), với mỗi sự kiện B bất kì ta có:

$$\begin{aligned} P(B) &= P\left(B \bigcup_{i=1}^n A_i\right) \\ &\iff P(B) = P\left(\bigcup_{i=1}^n BA_i\right) \\ &\iff P(B) = \sum_{i=1}^n P(BA_i) \\ &\iff P(B) = \sum_{i=1}^n P(A_i)P(B | A_i) \end{aligned} \quad (2.14)$$

- Công thức trên được gọi là công thức xác suất đầy đủ. Nếu $P(B) > 0$ thì với bất kì $A \in \{A_i\}$, ta tính được xác suất của A sau khi quan sát B như sau

$$P(A | B) = \frac{P(A)P(B | A)}{\sum_{i=1}^n P(A_i)P(B | A_i)}$$

2. Xác suất

- Công thức Bayes odds cho sự kiện A và dữ liệu \mathcal{D}

$$odds(A | \mathcal{D}) = odds(A) \frac{p(\mathcal{D} | A)}{p(\mathcal{D} | \bar{A})} \quad (2.15)$$

Mô hình thống kê

Trong xây dựng **mô hình thống kê** cho dữ liệu, với \mathcal{D} là tập hợp dữ liệu và θ tham số **mô hình** (*model*), công thức Bayes trở thành

$$P(\theta | \mathcal{D}) = \frac{P(\theta)P(\mathcal{D} | \theta)}{P(\mathcal{D})} \quad (2.16)$$

- $P(\theta)$: xác suất tiên nghiệm (*prior probability*)
- $P(\theta | \mathcal{D})$: xác suất hậu nghiệm (*a posterior probability*)
- $P(\mathcal{D})$: hằng số chuẩn hóa (*normalizing constant*)
- $P(\mathcal{D} | \theta)$: khả năng (*likelihood*)

2.2.5. Công thức Bernoulli

Khái niệm 2.15. Một phép thử mà kết quả chỉ có 2 sự kiện là xảy ra A với xác suất $P(A) = p$ hoặc không xảy ra A với xác suất $P(\bar{A}) = 1 - p = q$ được gọi là phép thử Bernoulli.

Định lý 2.1. Xác suất để xảy ra sự kiện A đúng k lần được tính bằng công thức Bernoulli như sau

$$P(A^k) = \binom{n}{k} p^k q^{n-k} \quad (2.17)$$

Chứng minh. Đặt B là sự kiện mà A xảy ra đúng k lần. Vì ta không quan tâm tới thứ tự xảy ra của các sự kiện A nên ta có cả thảy là $\binom{n}{k}$ kết quả. Hay nói cách khác B là tổng của $\binom{n}{k}$ sự kiện. Vì các kết quả này là độc lập với nhau nên ta có thể biểu diễn $P(B)$ như sau:

$$P(B) = \binom{n}{k} P(B_0)$$

Trong đó, B_0 là sự kiện mà A xảy ra đúng k lần và \bar{A} xảy ra $n - k$ lần. Do 2 sự kiện này là độc lập nên ta có $P(B_0) = P(A_k)P(\bar{A}_{k'})$. Ngoài ra, do A xảy ra k lần nên: $P(A_k) = p^k$, còn \bar{A} xảy ra $n - k$ lần nên: $P(\bar{A}_{k'}) = q^{n-k}$. Như vậy $P(B_0) = p^k q^{n-k}$. Thế $P(B_0)$ vào công thức phía trên ta sẽ có:

$$P(B) = \binom{n}{k} p^k q^{n-k}$$

2. Xác suất

Dễ thấy rằng để A xảy ra trong khoảng $[k_1, k_2]$ lần thì xác suất sẽ là tổng của từng xác suất thành phần:

$$\begin{aligned} P(A; k_1, k_2) &= \sum_{k=k_1}^{k_2} P(A^k) \\ &= \sum_{k=k_1}^{k_2} \binom{n}{k} p^k q^{n-k} \end{aligned}$$

■

2.3

Nguyên lý xác suất nhỏ, nguyên lý xác suất lớn

2.3.1. Nguyên lý xác suất nhỏ

Sự kiện không thể có xác suất bằng 0, một sự kiện có xác suất gần bằng 0 vẫn có thể xảy ra khi thực hiện một số lớn các phép thử. Tuy nhiên qua thực nghiệm và quan sát thực tế, người ta thấy rằng các sự kiện có xác suất nhỏ sẽ không xảy ra khi ta chỉ thực hiện một phép thử hay một vài phép thử. Từ đó ta thừa nhận nguyên lý sau đây

Nguyên lý (nguyên lý xác suất nhỏ). *Nếu một sự kiện có xác suất rất nhỏ thì thực tế có thể cho rằng trong một phép thử sự kiện đó sẽ không xảy ra.*

Mức xác suất được coi là nhỏ tùy thuộc vào từng bài toán cụ thể và gọi là **mức ý nghĩa**, ký hiệu là α .

2.3.2. Nguyên lý xác suất lớn

Tương tự như trên, ta có thể đưa ra

Nguyên lý (nguyên lý xác suất lớn). *Nếu sự kiện A có xác suất gần bằng 1 thì trên thực tế có thể cho rằng trong một phép thử sự kiện đó sẽ xảy ra.*

Mức xác suất đủ lớn gọi là **độ tin cậy**, ký hiệu là $\gamma = 1 - \alpha$. Việc quy định một mức xác suất thế nào là lớn sẽ tùy thuộc vào từng bài toán cụ thể.

Bài tập

Biên cỗ

B 2.1. Khi nào thì có các đẳng thức sau:

1. $A + B = A$
2. $AB = A$
3. $A + B = AB$

2. Xác suất

Hai sự kiện A và $A + B$ có xung khắc không?

B 2.2. Một chiếc tàu thủy gồm một bánh lái, 4 nồi hơi, 2 tuốc bin. Gọi A, B_i ($i = 1, \dots, 4$), C_j ($j = 1, 2$) lần lượt là các sự kiện bánh lái hoạt động tốt, nồi hơi thứ i hoạt động tốt, tuốc bin thứ j hoạt động tốt. Biết rằng tàu hoạt động tốt khi và chỉ khi bánh lái, ít nhất 1 nồi hơi và ít nhất một tuốc bin đều hoạt động tốt. Gọi D là sự kiện tàu hoạt động tốt. Hãy biểu diễn D và \overline{D} qua A, B_i, C_j .

B 2.3. Có 4 sinh viên làm bài thi. Kí hiệu B_i ($i = 1, \dots, 4$) là biến cố sinh viên thứ i làm bài thi đạt yêu cầu. Hãy biểu diễn các biến cố sau đây:

1. Có đúng một sinh viên đạt yêu cầu.
2. Có đúng ba sinh viên đạt yêu cầu.
3. Có ít nhất một sinh viên đạt yêu cầu.
4. Không có sinh viên nào đạt yêu cầu.

B 2.4. Xét phép thử: Gieo một xúc xắc 2 lần. Mô tả không gian biến cố sơ cấp ứng với phép thử trên?

Gọi A : “Tổng số nốt chia hết cho 3”, B : “Trị tuyệt đối của hiệu số nốt là số chẵn”. Biểu diễn A, B ?

B 2.5. Cho A, B là hai biến cố ngẫu nhiên đã biết. Tìm biến cố X từ hệ thức:

$$\overline{X + A} + \overline{X + \overline{A}} = B$$

B 2.6. Xét phép thử: Bắn không hạn chế vào 1 bia cho đến khi trúng bia lần đầu tiên thì dừng. Biểu diễn không gian biến cố sơ cấp của biến cố trên. Chỉ ra một hệ đầy đủ các biến cố.

B 2.7. Gieo hai con xúc xắc cân đối và đồng chất. Gọi A_i là biến cố xảy ra khi số nốt ở mặt trên con xúc xắc thứ nhất là i ($i = 1, \dots, 6$), B_k biến cố xảy ra khi số nốt ở mặt trên con xúc xắc thứ hai là k ($k = 1, \dots, 6$).

1. Hãy mô tả các biến cố A_6B_6, A_3B_5
2. Viết bằng kí hiệu các biến cố:
 - A : “hiệu giữa số nốt ở mặt trên con xúc xắc thứ nhất và thứ hai có trị số tuyệt đối bằng ba”.
 - B : “số nốt ở mặt trên hai con xúc xắc bằng nhau”.
3. Hãy chỉ ra một nhóm đầy đủ các biến cố.

2. Xác suất

Xác suất cổ điển

B 2.8. Một nhóm n người xếp ngẫu nhiên thành một hàng dài.

1. Tìm xác suất để 2 người định trước đứng cạnh nhau.
2. Tìm xác suất để 2 người đó đứng cách nhau 2 người.
3. Tìm xác suất để 2 người đó đứng cách nhau r người ($0 < r < n - 2$).
4. Xét trường hợp khi họ xếp thành một vòng tròn.

B 2.9. Thang máy của một tòa nhà 7 tầng, xuất phát từ tầng một với 3 người khách. Tính xác suất để:

1. Tất cả cùng ra ở tầng bốn.
2. Tất cả cùng ra ở một tầng.
3. Mỗi người ra một tầng khác nhau.

B 2.10. Có n quả cầu được phân ngẫu nhiên lần lượt vào n hộp, mỗi hộp có thể chứa nhiều quả cầu. Khi phân biệt hộp và cầu, tìm xác suất để mỗi hộp chứa một quả cầu.

B 2.11. Cho một lô hàng gồm n sản phẩm trong đó có m sản phẩm xấu. Lấy ngẫu nhiên từ lô hàng đó k sản phẩm. Tìm xác suất sao cho trong số sản phẩm lấy ra có đúng s sản phẩm xấu ($s < k$).

B 2.12. Ta gieo liên tiếp 4 lần một đồng tiền cân đối đồng chất. Tìm xác suất của các biến cố:

1. A : “Có hai mặt sấp”.
2. B : “Có ba mặt ngửa”.
3. C : “Có ít nhất một mặt sấp”.

B 2.13. Mười hai sản phẩm được sắp ngẫu nhiên vào ba hộp. Tìm xác suất để hộp thứ nhất có chứa ba sản phẩm.

B 2.14. Gieo đồng thời hai con xúc xắc đồng chất cân đối n lần liên tiếp. Tìm xác suất để xuất hiện ít nhất một lần hai mặt trên cùng có 6 nốt.

2. Xác suất

Xác suất hình học

B 2.15. Một thanh sắt thẳng được bẻ thành ba khúc một cách ngẫu nhiên. Tìm xác suất để ba khúc đó tạo được thành một tam giác. Biết rằng thanh sắt dài l (đơn vị dài.)

B 2.16. (Bài toán Buffon) Trên mặt phẳng có các đường thẳng song song cách đều nhau $2a$, gieo ngẫu nhiên một cây kim có độ dài $2l$ ($l < a$). Tìm xác suất để cây kim cắt một đường thẳng nào đó.

B 2.17. Trên đường tròn bán kính R có một điểm A cố định, chọn ngẫu nhiên một điểm B . Tìm xác suất để cung AB không quá R .

B 2.18. Trên đoạn thẳng OA ta gieo một cách ngẫu nhiên hai điểm B, C có tọa độ tương ứng là $OB = x, OC = y$ ($y \geq x$). Tìm xác suất sao cho độ dài của đoạn BC bé hơn độ dài của đoạn OB .

Các công thức tính xác suất cơ bản

B 2.19. Một hệ thống được cấu tạo bởi 3 bộ phận độc lập nhau. Hệ thống sẽ hoạt động nếu ít nhất 2 trong 3 bộ phận còn hoạt động. Nếu độ tin cậy của mỗi bộ phận là 0.95 thì độ tin cậy của hệ thống là bao nhiêu?

B 2.20. Một hộp có 7 bi đỏ và 3 bi đen.

1. Lấy ngẫu nhiên 1 viên bi từ hộp ra để kiểm tra. Tính xác suất nhận được bi đen.
2. Lấy ngẫu nhiên lần lượt có hoàn lại 2 bi. Tính xác suất để lấy được 2 bi đen.
3. Lấy ngẫu nhiên ra 2 viên bi từ hộp. Tính xác suất để lấy được 2 bi đen.

B 2.21. Cho $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{2}$ và $P(A+B) = \frac{3}{4}$. Tính $P(AB), P(\bar{A}.\bar{B}), P(\bar{A}+\bar{B}), P(A\bar{B}), P(\bar{A}B)$.

B 2.22. Tỷ lệ người mắc bệnh tim trong một vùng dân cư là 9%, mắc bệnh huyết áp là 12%, mắc cả hai bệnh là 7%. Chọn ngẫu nhiên một người trong vùng. Tính xác suất để người đó

1. Bị bệnh tim hay bị bệnh huyết áp.
2. Không bị bệnh tim cũng không bị bệnh huyết áp.
3. Không bị bệnh tim hay không bị bệnh huyết áp.
4. Bị bệnh tim nhưng không bị bệnh huyết áp.
5. Không bị bệnh tim nhưng bị bệnh huyết áp.

B 2.23. Bạn quên mất số cuối cùng trong số điện thoại cần gọi (số điện thoại gồm 6 chữ số) và bạn chọn số cuối cùng này một cách ngẫu nhiên. Tính xác suất để bạn gọi đúng số điện thoại này mà không phải thử quá 3 lần. Nếu biết số cuối cùng là số lẻ thì xác suất này là bao nhiêu ?

2. Xác suất

B 2.24. Xét

1. Cho A, B là hai biến cố độc lập. Chứng minh rằng $\overline{A}, \overline{B}$; A, \overline{B} và $\overline{A}, \overline{B}$ đều là các cặp biến cố độc lập.
2. Cho A_1, A_2, \dots, A_n là n biến cố độc lập. Chứng minh rằng A_1, A_2, \dots, A_n cũng là n biến cố độc lập. Từ đó suy ra rằng nếu xét n biến cố B_1, B_2, \dots, B_n với $B_i = A_i$ hoặc $B_i = \overline{A}_i$ thì B_1, B_2, \dots, B_n cũng là n biến cố độc lập.

B 2.25. Một đợt xổ số phát hành N vé, trong đó có M vé có thưởng. Một người mua r vé ($r < N - M$). Tính xác suất để người đó có ít nhất một vé trúng thưởng.

B 2.26. Một người có 3 con gà mái, 2 con gà trống nhốt chung một lồng. Một người đến mua, người bán bắt ngẫu nhiên ra một con. Người mua chấp nhận mua con đó.

1. Tìm xác suất để người đó mua được con gà mái.

Người thứ hai đến mua, người bán lại bắt ngẫu nhiên ra một con.

2. Tìm xác suất người thứ hai mua được gà trống, biết rằng người thứ nhất mua được gà mái.
3. Xác suất trên bằng bao nhiêu nếu người bán gà quên mất rằng con gà bán cho người thứ nhất là gà trống hay gà mái?

B 2.27. Có một nhóm n sinh viên, mỗi người có một áo mưa giống hệt nhau. Một hôm trời mưa, cả nhóm cùng đến lớp và treo áo ở mắc áo. Lúc ra về vì vội vàng mỗi người lấy hú họa một cái áo. Tính xác suất có ít nhất một sinh viên chọn đúng áo của mình.

B 2.28. Một người viết n lá thư và bỏ n lá thư này vào trong n phong bì đã viết sẵn địa chỉ. Tìm xác suất sao cho có ít nhất một lá thư được bỏ đúng vào phong bì của nó.

B 2.29. Ba xạ thủ, mỗi người bắn một viên đạn vào mục tiêu với xác suất trúng đích của mỗi người là 0.6; 0.7; 0.8. Tìm xác suất

1. chỉ có người thứ hai bắn trúng.
2. có đúng một người bắn trúng.
3. có ít nhất một người bắn trúng.
4. cả ba người đều bắn trúng.
5. có đúng hai người bắn trúng.
6. có ít nhất hai người bắn trúng.
7. có không quá hai người bắn trúng.

2. Xác suất

B 2.30. Cho hai biến cố xung khắc A và B , sao cho $P(A) \neq 0$, $P(B) \neq 0$. Chứng minh rằng A và B phụ thuộc nhau.

B 2.31. Ba con ngựa a, b, c trong một cuộc đua ngựa. Nếu xuất hiện bac có nghĩa là b đến đích trước, sau đó là a và về cuối là c . Khi đó tập hợp tất cả các khả năng xuất hiện là

$$\Omega = \{abc, acb, bac, bca, cab, cba\}.$$

Giả sử rằng $P[\{abc\}] = P[\{acb\}] = 1/18$ và bốn khả năng còn lại đều có xác suất xảy ra là $2/9$.

Hơn nữa, ta định nghĩa các biến cố

$A = "a$ đến đích trước $b"$ và $B = "a$ đến đích trước $c"$

1. Hai biến cố A và B có tạo thành một hệ đầy đủ của Ω ?

2. Hai biến cố A và B có độc lập nhau?

B 2.32. Có tồn tại hai biến cố xung khắc và độc lập không?

B 2.33. Một máy tính điện tử gồm có n bộ phận. Xác suất hỏng trong khoảng thời gian T của bộ phận thứ k bằng p_k ($k = 1, 2, \dots, n$). Nếu dù chỉ một bộ phận bị hỏng thì máy tính ngừng làm việc. Tìm xác suất để máy tính ngừng làm việc trong khoảng thời gian T .

B 2.34. Chứng minh rằng nếu $P(A | B) > P(A)$, thì $P(B | A) > P(B)$

B 2.35. Giả sử $P(AB) = 1/4$, $P(A | B) = 1/8$ và $P(B) = 1/2$. Tính $P(A)$.

B 2.36. Biết rằng ta đã nhận được ít nhất một mặt ngửa trong 3 lần tung đồng xu độc lập. Hỏi xác suất đạt được cả 3 mặt ngửa là bao nhiêu?

B 2.37. Tung một con xúc sắc hai lần độc lập nhau. Biết rằng lần tung thứ nhất được số nốt chẵn. Tính xác suất tổng số nốt hai lần tung bằng 4.

B 2.38. Giả sử $P(A) = P(B) = 1/4$ và $P(A | B) = P(B)$. Tính $P(AB)$.

B 2.39. Bắn liên tiếp vào một mục tiêu cho đến khi có một viên đạn đầu tiên rơi vào mục tiêu thì ngừng bắn. Tìm xác suất sao cho phải bắn đến viên thứ 6, biết rằng xác suất trúng đích của mỗi viên đạn là 0.2 và các lần bắn là độc lập.

B 2.40. Giả sử các biến cố A_1, \dots, A_n độc lập có xác suất tương ứng $P(A_k) = p_k$ ($k = 1, \dots, n$). Tìm xác suất sao cho:

1. không một biến cố nào trong các biến cố đó xuất hiện.

2. có ít nhất một biến cố trong các biến cố đó xuất hiện. Từ đó suy ra công thức khai triển tích

$$\prod_{k=1}^n (1 - p_k)$$

2. Xác suất

B 2.41. Có ba tiêu chí phổ biến cho việc chọn mua một chiếc xe hơi mới nào đó là A : hộp số tự động, B : động cơ V6, và C : điều hòa nhiệt độ. Dựa trên dữ liệu bán hàng trước đây, ta có thể giả sử rằng $P(A) = 0.7$, $P(B) = 0.75$, $P(C) = 0.80$, $P(A + B) = 0.80$, $P(A + C) = 0.85$, $P(B + C) = 0.90$ và $P(A + B + C) = 0.95$, với $P(A)$ là xác suất người mua bất kì chọn tiêu chí A , ... Tính xác suất của các biến cỗ sau:

1. người mua chọn ít nhất một trong 3 tiêu chí.
2. người mua không chọn tiêu chí nào trong 3 tiêu chí trên.
3. người mua chỉ chọn tiêu chí điều hòa nhiệt độ.
4. người mua chọn chính xác một trong 3 tiêu chí.

B 2.42. Giả sử $P(B | A_1) = 1/2$, $P(B | A_2) = 1/4$ với A_1 và A_2 là hai biến cỗ đồng khả năng và tạo thành một hệ đầy đủ các biến cỗ. Tính $P(A_1 | B)$.

B 2.43. Một hộp đựng 10 phiếu trong đó có 2 phiếu trúng thưởng. Có 10 người lần lượt rút thăm. Tính xác suất nhận được phần thưởng của mỗi người.

Công thức xác suất đầy đủ, công thức Bayes

B 2.44. Có hai hộp đựng bi. Hộp 1 đựng 20 bi trong đó có 5 bi đỏ và 15 bi trắng. Hộp 2 đựng 15 bi trong đó có 6 bi đỏ và 9 bi trắng. Lấy một bi ở hộp 1 bỏ vào hộp 2, trộn đều rồi lấy ra một bi. Tính xác suất nhận được bi đỏ? bi trắng?

B 2.45. Trong một vùng dân cư, cứ 100 người thì có 30 người hút thuốc lá. Biết tỷ lệ người bị viêm họng trong số người hút thuốc lá là 60%, trong số người không hút thuốc lá là 30%. Khám ngẫu nhiên một người và thấy người đó bị viêm họng.

1. Tìm xác suất để người đó hút thuốc lá.
2. Nếu người đó không bị viêm họng thì xác suất để người đó hút thuốc lá là bao nhiêu.

B 2.46. Một trung tâm chẩn đoán bệnh dùng một phép kiểm định T . Xác suất để một người đến trung tâm mà có bệnh là 0.8. Xác suất để người khám có bệnh khi phép kiểm định dương tính là 0.9 và xác suất để người khám không có bệnh khi phép kiểm định âm tính là 0.5. Tính các xác suất

1. phép kiểm định là dương tính.
2. phép kiểm định cho kết quả đúng.

B 2.47. Một cặp trẻ sinh đôi có thể do cùng một trứng (sinh đôi thật) hay do hai trứng khác nhau sinh ra (sinh đôi giả). Các cặp sinh đôi thật luôn luôn có cùng giới tính. Các cặp sinh đôi giả thì giới tính của mỗi đứa độc lập với nhau và có xác suất là 0.5. Thống kê cho thấy 34% cặp sinh đôi là trai; 30% cặp sinh đôi là gái và 36% cặp sinh đôi có giới tính khác nhau.

2. Xác suất

1. Tính tỷ lệ cắp sinh đôi thật.
2. Tìm tỷ lệ cắp sinh đôi thật trong số các cắp sinh đôi có cùng giới tính.

B 2.48. Có 10 hộp bi, trong đó có 4 hộp loại I, 3 hộp loại II, còn lại là hộp loại III. Hộp loại I có 3 bi trắng và 5 đỏ, hộp loại II có 4 bi trắng và 6 bi đỏ, hộp loại III có 2 bi trắng và 2 bi đỏ.

1. Chọn ngẫu nhiên một hộp và từ đó lấy hú họa 1 bi. Tìm xác suất để được bi đỏ.
2. Chọn ngẫu nhiên một hộp và từ đó lấy 1 bi thì được bi trắng. Tìm xác suất để bi lấy ra thuộc loại II.

B 2.49. Có hai lô sản phẩm, lô thứ nhất có 10 sản phẩm loại I và 2 sản phẩm loại II. Lô thứ hai có 16 sản phẩm loại I và 4 sản phẩm loại II. Từ mỗi lô ta lấy ngẫu nhiên một sản phẩm. Sau đó, từ 2 sản phẩm thu được lấy hú họa ra một sản phẩm. Tìm xác suất để sản phẩm lấy ra sau cùng là sản phẩm loại I.

B 2.50. Có 2 lô gà. Lô thứ nhất gồm 15 con, trong đó có 3 con gà trống. Lô thứ hai gồm 20 con, trong đó có 4 gà trống. Một con từ lô thứ hai nhảy sang lô thứ nhất. Sau đó từ lô thứ nhất ta bắt ngẫu nhiên ra một con. Tìm xác suất để con gà bắt ra là gà trống.

B 2.51. Ba máy tự động sản xuất cùng một loại chi tiết, trong đó máy I sản xuất 25%, máy II sản xuất 30% và máy III sản xuất 45% tổng sản lượng. Tỷ lệ phế phẩm của các máy lần lượt là 0.1%;0.2%;0.4%. Tìm xác suất để khi chọn ngẫu nhiên ra 1 sản phẩm từ kho thì

1. được chi tiết phế phẩm.
2. chi tiết phế phẩm đó do máy II sản xuất.

B 2.52. Giả sử 3 máy M_1, M_2, M_3 sản xuất lần lượt 500, 1000, 1500 linh kiện mỗi ngày với tỉ lệ phế phẩm tương ứng là 5%, 6% và 7%. Vào cuối ngày làm việc nào đó, người ta lấy một linh kiện được sản xuất bởi một trong 3 máy trên một cách ngẫu nhiên, kết quả là được một phế phẩm. Tìm xác suất linh kiện này được sản xuất bởi máy M_3 .

B 2.53. Ba khẩu pháo cùng bắn vào một mục tiêu với xác suất trúng đích của mỗi khẩu là 0.4;0.7;0.8. Biết rằng xác suất để mục tiêu bị tiêu diệt khi trúng một phát đạn là 30%, khi trúng 2 phát đạn là 70%, còn trúng 3 phát đạn thì chắc chắn mục tiêu bị tiêu diệt. Giả sử mỗi khẩu pháo bắn 1 phát.

1. Tính xác suất để mục tiêu bị tiêu diệt.
2. Biết rằng mục tiêu đã bị tiêu diệt. Tính xác suất để khẩu thứ 3 có đóng góp vào thành công đó.

B 2.54. Hộp I có 10 linh kiện trong đó có 3 bị hỏng. Hộp II có 15 linh kiện trong đó có 4 bị hỏng. Lấy ngẫu nhiên từ mỗi hộp ra một linh kiện.

2. Xác suất

1. Tính xác suất để cả 2 linh kiện lấy ra đều hỏng.
2. Số linh kiện còn lại trong 2 hộp đem bỏ vào hộp III. Từ hộp III lấy ngẫu nhiên ra 1 linh kiện. Tính xác suất để linh kiện lấy ra từ hộp III bị hỏng.
3. Biết linh kiện lấy ra từ hộp III là hỏng. Tính xác suất để 2 linh kiện lấy ra từ hộp I và II lúc ban đầu là hỏng.

B 2.55. Có 3 cửa hàng I, II, III cùng kinh doanh sản phẩm Y , trong đó thị phần của cửa hàng I, III như nhau và gấp đôi thị phần của cửa hàng II. Tỉ lệ sản phẩm loại A trong 3 cửa hàng lần lượt là 70%, 75% và 50%. Một khách hàng chọn ngẫu nhiên 1 cửa hàng và từ đó mua một sản phẩm.

1. Tính xác suất để khách hàng mua được sản phẩm loại A.
2. Giả sử khách hàng đã mua được sản phẩm loại A, hỏi khả năng người ấy đã mua được ở cửa hàng nào là nhiều nhất.

B 2.56. Cho T là một phép thử ngẫu nhiên với 3 biến cố sơ cấp có thể xảy ra là A , B và C . Giả sử ta tiến hành T vô hạn lần và độc lập nhau. Tính theo $P(A)$, $P(B)$ xác suất biến cố A xuất hiện trước B .

Biến ngẫu nhiên và hàm phân phối

3.1

Biến ngẫu nhiên

Khái niệm 3.1. *Biến ngẫu nhiên* (random variable) là các biến nhận 1 giá trị ngẫu nhiên đại diện cho kết quả của phép thử. Mỗi giá trị nhận được $x \in \mathcal{D}$ của biến ngẫu nhiên X được gọi là một thể hiện của X , đây cũng là kết quả của phép thử hay còn được hiểu là một sự kiện. Trong xác suất, thường tập giá trị \mathcal{D} được giới hạn là tập các giá trị số, ví dụ \mathbb{R} .

Biến ngẫu nhiên có 2 dạng (dựa trên tập giá \mathcal{D})

- **Rời rạc** (*discrete*): tập giá trị nó là rời rạc; ví dụ, như mặt chấm của con xúc xắc.
- **Liên tục** (*continuous*): tập giá trị là liên tục; ví dụ, giá thuê nhà ở TPHCM.

Khái niệm 3.2. *Biến ngẫu nhiên (giá trị thực)* là một hàm từ không gian xác suất vào không gian số thực

$$X : \text{khoảng gian xác suất} \longrightarrow \mathbb{R} \quad (3.1)$$

Lưu ý:

- Nếu X là một biến ngẫu nhiên (giá trị thực) và g là một hàm từ \mathbb{R} vào \mathbb{R} thì $g(X)$ cũng là một biến ngẫu nhiên (giá trị thực).
- Nếu X, Y là những biến ngẫu nhiên (giá trị thực) $X + Y, X - Y, XY, X/Y$ cũng là những biến ngẫu nhiên (giá trị thực).

3.2

Phân phối xác suất

3.2.1. Hàm khối xác suất của biến rời rạc

Khái niệm 3.3. Hàm xác suất như vậy đối với biến ngẫu nhiên rời rạc được gọi là **hàm khối xác suất** (PMF - Probability Mass Function). Giả sử miền xác định của X là \mathcal{D} , thì hàm khối xác suất được xác định như sau

$$p(x) = p_X(x) = \begin{cases} P(X = x) & \text{if } x \in \mathcal{D} \\ 0 & \text{if } x \notin \mathcal{D} \end{cases} \quad (3.2)$$

- Như vậy ta có thể thấy rằng hàm khối xác suất thực chất cũng là một xác suất nên nó mang đầy đủ tất cả các tính chất của xác suất như

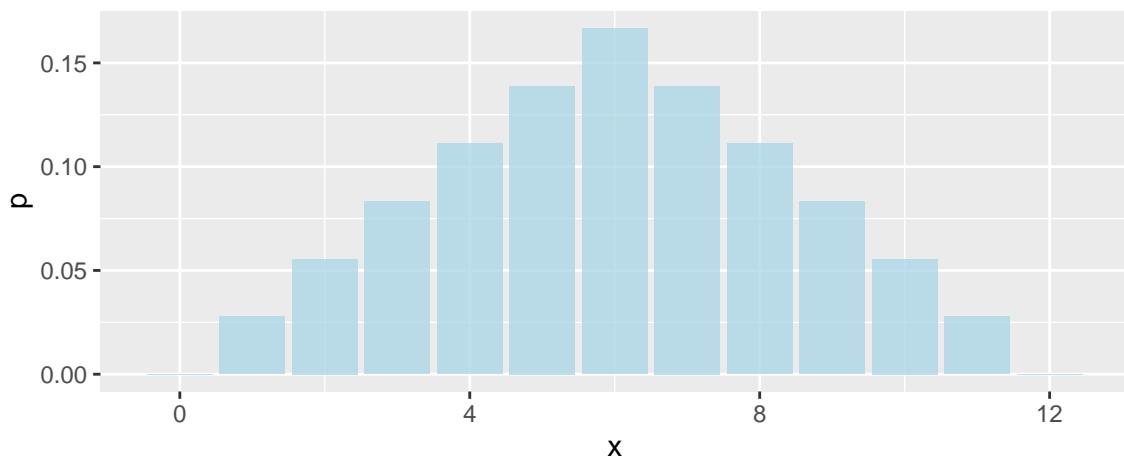
$$1. \ 0 \leq p(x) \leq 1$$

$$2. \ \sum_{x_i \in \mathcal{D}} p(x_i) = 1$$

Ví dụ. Xét hàm phân phối khối xác suất của biến rời rạc

$$p(x) = \begin{cases} \frac{x}{36} & \text{nếu } x \in \mathbb{N}, 0 \leq x \leq 6 \\ \frac{12-x}{36} & \text{nếu } x \in \mathbb{N}, 7 \leq x \leq 12 \\ 0 & \text{còn lại} \end{cases}$$

thì ta có thể biểu diễn bằng biểu đồ phân phối như sau



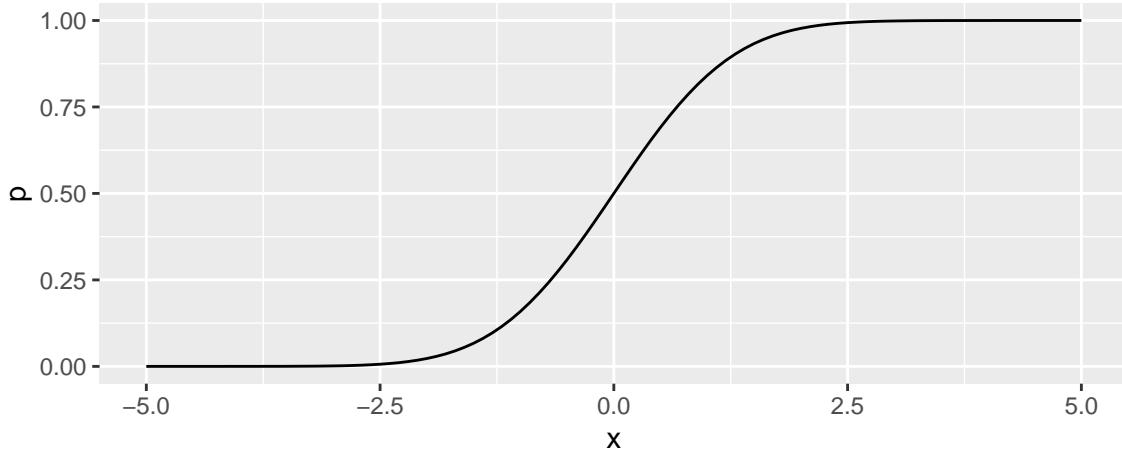
3.2.2. Hàm phân phối xác suất biến liên tục

Khái niệm 3.4. Hàm phân phối xác suất của biến ngẫu nhiên X được xác định như sau

$$F_X(x) = P(X \leq x) \quad , x \in \mathbb{R} \quad (3.3)$$

3. Biến ngẫu nhiên và hàm phân phối

- Hàm phân phối xác suất còn có tên là hàm phân phối tích lũy (CDF - Cumulative Distribution Function) do đặc trưng là lấy xác suất của biến ngẫu nhiên bên trái của một giá trị x bất kì nào đó.
- Hàm này có đặc điểm là một hàm không giảm, tức là nếu $a < b$ thì $F_X(a) \leq F_X(b)$ vì sự kiện $X \leq b$ đã bao gồm cả sự kiện $X \leq a$.



Hàm phân phối tích lũy F của biến ngẫu nhiên rời rạc có thể được biểu diễn qua hàm khối xác suất bằng cách lấy tổng

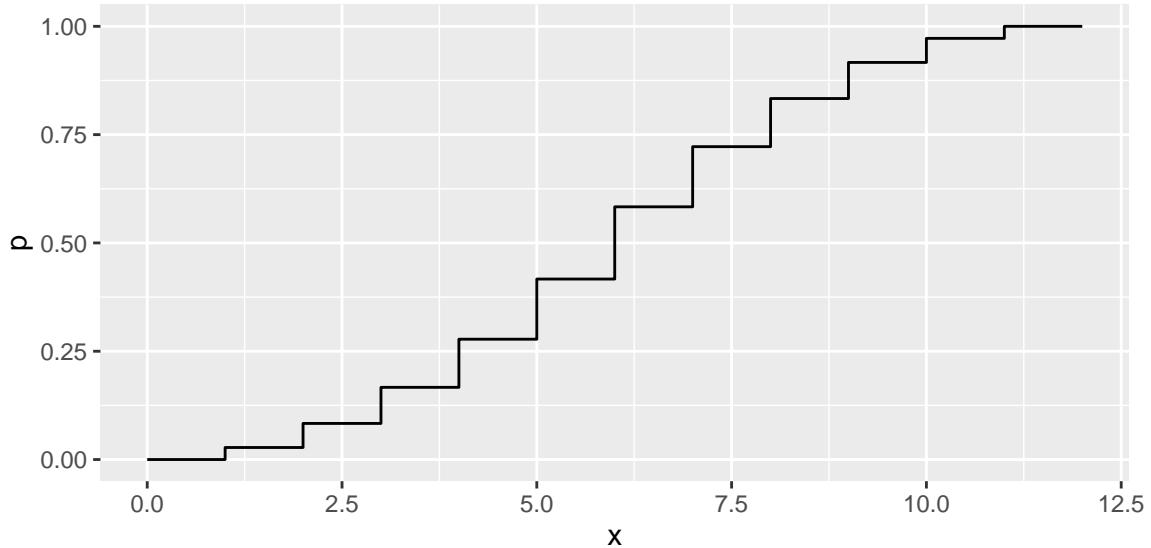
$$F_X(x) = \sum_{x_i \leq x} p(x_i), \quad x \in \mathbb{R}$$

Lúc này, hàm phân phối tích lũy sẽ có dạng bậc thang ứng với mỗi bậc là khoảng (x_i, x_{i+1}) .
Hàm phân phối tích lũy của ví dụ trước sẽ có dạng như sau

$$F(x) = \begin{cases} 0 & \text{nếu } x < 1 \\ 1/36 & \text{nếu } 1 \leq x < 2 \\ 3/36 & \text{nếu } 2 \leq x < 3 \\ 6/36 & \text{nếu } 3 \leq x < 4 \\ 10/36 & \text{nếu } 4 \leq x < 5 \\ 15/36 & \text{nếu } 5 \leq x < 6 \\ 21/36 & \text{nếu } 6 \leq x < 7 \\ \dots & \dots \end{cases}$$

và biểu đồ tương ứng là

3. Biến ngẫu nhiên và hàm phân phối



3.2.3. Hàm mật độ xác suất của biến liên tục

Với các biến ngẫu nhiên liên tục ta có khái niệm hàm mật độ xác suất (PDF - Probability Density Function) để ước lượng độ tập trung xác suất tại lân cận điểm nào đó.

Khái niệm 3.5. *Hàm mật độ xác suất $f(x)$ tại điểm x được xác định bằng cách lấy đạo hàm của hàm phân tích lũy $F(x)$ tại điểm đó*

$$f(x) = F'(x) \quad (3.4)$$

- Như vậy thì nơi nào $f(x)$ càng lớn thì ở đó mức độ tập xác suất càng cao. Từ đây ta cũng có thể biểu diễn hàm phân tích lũy như sau

$$F(x) = \int_{-\infty}^x f(t)dt \quad (3.5)$$

- Xác suất trong 1 khoảng (α, β) cũng có thể được tính bằng hàm mật độ xác suất

$$P(\alpha \leq X \leq \beta) = \int_{\alpha}^{\beta} f(x)dx \quad (3.6)$$

- Hàm mật độ xác suất cũng có 2 tính chất như xác suất như sau

- Không âm: $f(x) \geq 0, \quad \forall x \in \mathbb{R}$
- Tổng toàn miền bằng 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

Ví dụ. Thời gian tính bằng đơn vị giờ mà một máy tính hoạt động trước khi xảy ra lỗi được coi như một biến ngẫu nhiên liên tục và được xác định với hàm mật độ xác suất sau

$$f(x) = \begin{cases} \lambda e^{-x/100} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

3. Biến ngẫu nhiên và hàm phân phối

Hãy tính xác suất của

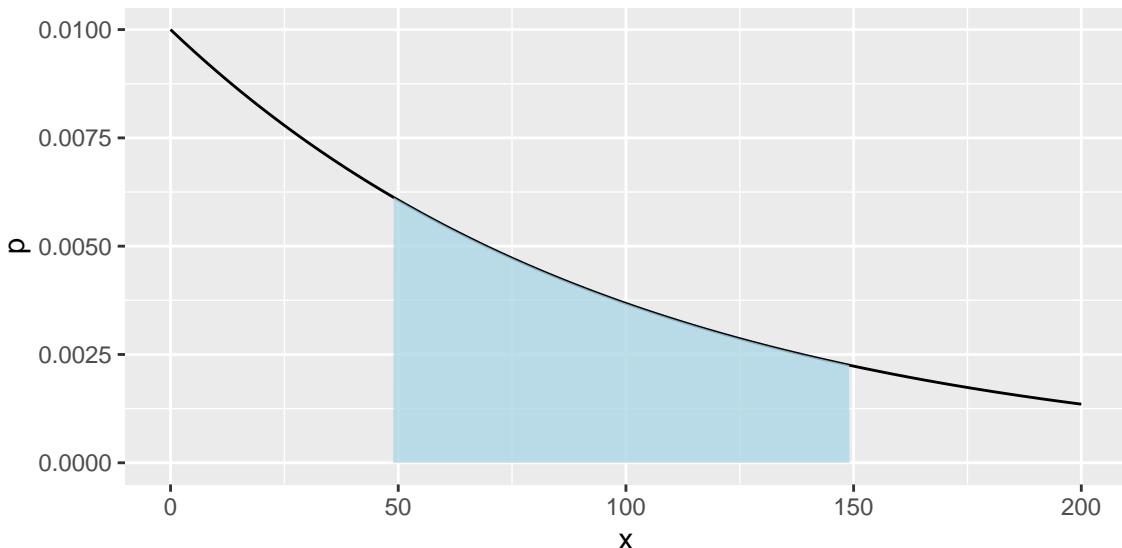
1. Một máy tính hoạt động từ 50 giờ tới 150 giờ trước khi xảy ra lỗi?
2. Một máy tính hoạt động dưới 100 giờ trước khi xảy ra lỗi?

Lời giải. Vì tổng xác suất toàn miền là 1 nên

$$\begin{aligned}
 & \int_{-\infty}^{\infty} f(x)dx = 1 \\
 \iff & \int_{-\infty}^{\infty} \lambda e^{-x/100} dx = 1 \\
 \iff & \lambda \int_{-\infty}^{\infty} e^{-x/100} dx = 1 \\
 \iff & \lambda \int_0^{\infty} e^{-x/100} dx = 1 \\
 \iff & -\lambda(100)e^{-x/100} \Big|_0^{\infty} = 1 \\
 \iff & 100\lambda = 1 \\
 \iff & \lambda = \frac{1}{100}
 \end{aligned}$$

1. Xác suất để 1 máy tính hoạt động được trong khoảng (50, 150) giờ là

$$\begin{aligned}
 P(50 < X < 150) &= \int_{50}^{150} \frac{1}{100} e^{-x/100} dx \\
 &= -e^{-x/100} \Big|_{50}^{150} \\
 &= e^{-1/2} - e^{-3/2} \\
 &\approx 0.384
 \end{aligned}$$

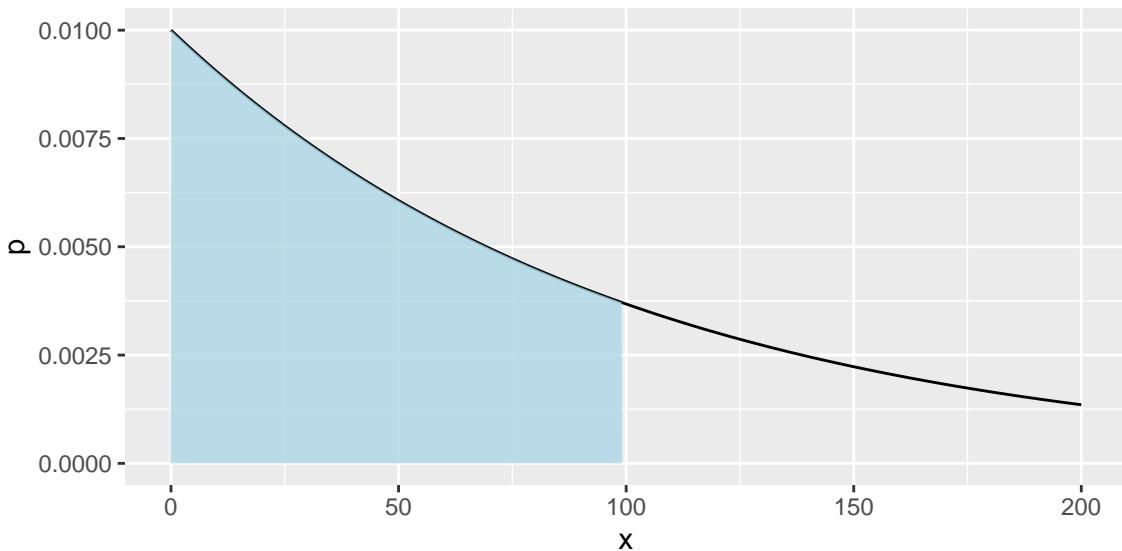


3. Biến ngẫu nhiên và hàm phân phối

Như vậy, xấp xỉ 38.4 phần trăm thời gian một máy tính sẽ hoạt động trước khi lỗi trong khoảng 50 tới 150 giờ.

- Xác suất để 1 máy tính hoạt động được trong vòng 100 trước khi lỗi là

$$\begin{aligned}
 P(X < 100) &= \int_0^{100} \frac{1}{100} e^{-x/100} dx \\
 &= -e^{-x/100} \Big|_0^{100} \\
 &= 1 - e^{-1} \\
 &\approx 0.633
 \end{aligned}$$



Nên xấp xỉ 63.3 phần trăm thời gian một máy tính sẽ lỗi sau 100 giờ sử dụng.

Nhìn vào các biểu đồ trên ta có thấy xác suất (1) là phần diện tích của hình thang cong phủ từ $50 < x < 150$, còn xác suất (2) là phần diện tích hình thang cong phủ tới $x < 100$. x càng lớn thì $f(x)$ cũng càng bé đi nên phần phần diện tích của nó càng hẹp dần đồng nghĩa với mật độ xác suất cũng giảm dần nên xác suất để máy tính hoạt động được ngày càng thấp đi. ■

- Lưu ý rằng khác với hàm xác suất, hàm mật độ xác suất tại 1 điểm bất kì luôn bằng 0

$$P(X = x) = \int_x^x f(t) dt = 0$$

- Ngoài ra, giá trị của hàm mật độ xác suất $f(x)$ có thể lớn hơn 1, miễn sao đảm bảo được rằng tổng xác suất toàn miền là 1

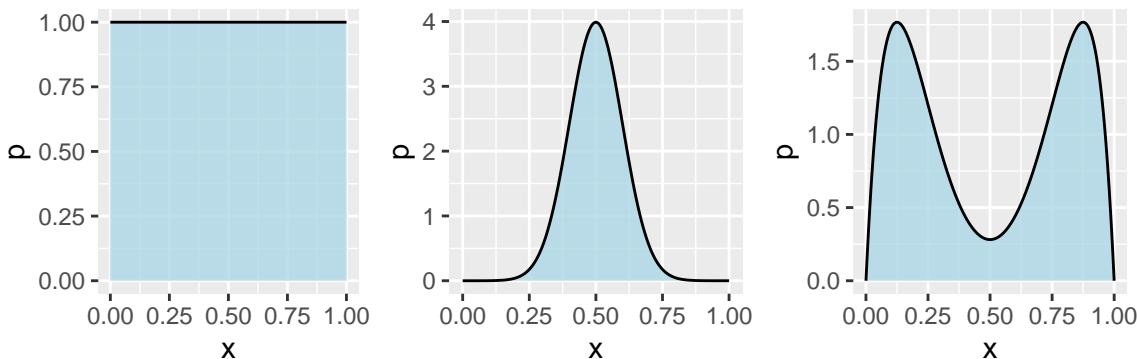
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

3.3**Các đặc trưng****3.3.1. Đồ thị hàm phân bố xác suất**

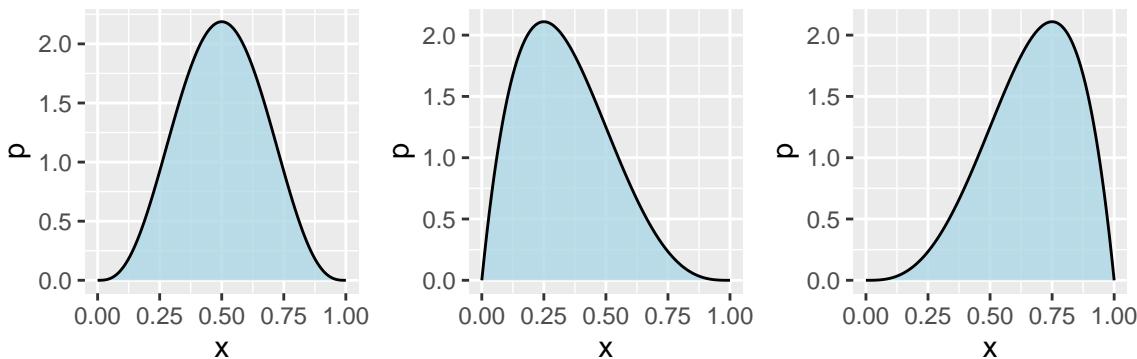
Đồ thị hàm phân bố xác suất là đặc trưng quan trọng nhất.

Các dạng đồ thị

- Đều hay có đỉnh (một đỉnh, hai đỉnh hay nhiều đỉnh)



- Đối xứng hay lệch (lệch trái hay lệch phải)



Trong thực tế ta quan tâm tới các đại lượng đặc trưng của phân bố xác suất như vị trí trung bình và độ phân tán ra sao.

3.3.2. Kỳ vọng

Định nghĩa. **Kỳ vọng** (*expectation*) của biến ngẫu nhiên là trung bình của biến ngẫu nhiên. Kỳ vọng của biến ngẫu nhiên X được kí hiệu là $\mathbb{E}[X]$

$$\mathbb{E}[X] = \begin{cases} \sum_{x_i} x_i p(x_i) & \text{nếu } X \text{ rời rạc} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{nếu } X \text{ liên tục} \end{cases} \quad (3.7)$$

3. Biến ngẫu nhiên và hàm phân phối

Kỳ vọng còn được biết tới với những tên gọi khác như giá trị trung bình (*mean*), giá trị trung bình có trọng lượng (*weighted average*), giá mong đợi (*expected value*) hay moment bậc một (*first moment*).

- Kỳ vọng có một số tính chất như sau

1. $\mathbb{E}(c) = c$ với c là hằng số
2. $\mathbb{E}(cX) = c\mathbb{E}(X)$ với c là hằng số
3. $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ với a, b là các hằng số
4. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
5. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ với X, Y là độc lập

Định nghĩa. Kỳ vọng của một hàm biến ngẫu nhiên $g(X)$ là

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x_i} g(x_i)p_X(x_i) & \text{nếu } x \text{ rời rạc} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{nếu } x \text{ liên tục} \end{cases} \quad (3.8)$$

3.3.3. Phương sai

Dựa vào kì vọng ta sẽ có được trung bình của biến ngẫu nhiên, tuy nhiên nó lại không cho ta thông tin về mức độ phân tán xác suất nên ta cần 1 phương pháp để đo được độ phân tán đó. Một trong những phương pháp đó là phương sai (*variance*).

Định nghĩa. Phương sai (*variance*) là trung bình của bình phương khoảng cách từ biến ngẫu nhiên tới giá trị trung bình. Phương sai của biến ngẫu nhiên X được kí hiệu là $\text{Var}(X)$

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad (3.9)$$

- Việc tính toán dựa vào công thức này khá phức tạp, nên trong thực tế người ta thường sử dụng công thức tương đương sau

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X] \quad (3.10)$$

Chứng minh. Ta có,

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}^2[X]] \\ &= \mathbb{E}[X^2] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}^2[X]], \quad \mathbb{E}[X] \text{ là hằng số} \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}^2[X] \\ &= \mathbb{E}[X^2] - \mathbb{E}^2[X] \end{aligned}$$

3. Biến ngẫu nhiên và hàm phân phối

■

Như vậy ta có thể thấy rằng phương sai luôn là một giá trị không âm và phương sai càng lớn thì nó thể hiện mức độ phân tán dữ liệu càng rộng hay nói cách khác mức độ ổn định càng nhỏ.

- Phương sai có một số tính chất sau
 1. $\text{Var}(c) = 0$ với c là hằng số
 2. $\text{Var}(cX) = c^2\text{Var}(X)$ với c là hằng số
 3. $\text{Var}(aX + b) = a^2\text{Var}(X)$ với a, b là các hằng số
 4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ với X, Y là độc lập

3.3.4. Độ lệch chuẩn

Định nghĩa. Vì đơn vị của phương sai là bình phương nên việc tính để khớp với đơn vị của biến ngẫu nhiên nên người ta đưa vào thêm khái niệm **độ lệch chuẩn** (*standard deviation*) bằng căn bậc 2 của phương sai. Độ lệch chuẩn ký hiệu là $\sigma(X)$.

$$\sigma(X) = \sqrt{\text{Var}(X)} \quad (3.11)$$

- Ta cũng có thể sử dụng $\sigma^2(X)$ để thể hiện phương sai của biến ngẫu nhiên X .
- Lưu ý với độ lệch chuẩn ta phải lấy trị tuyệt đối của hằng số khi nhân vì độ lệch chuẩn cũng là không âm

$$\sigma(cX) = |c|\sigma(X)$$

3.3.5. Trung vị

Định nghĩa. **Trung vị** (*median*) là điểm chia đều xác suất thành 2 phần giống nhau, ký hiệu là $\text{med}(X)$

$$P(X < \text{med}(X)) = P(X \geq \text{med}(X)) = 0.5 \quad (3.12)$$

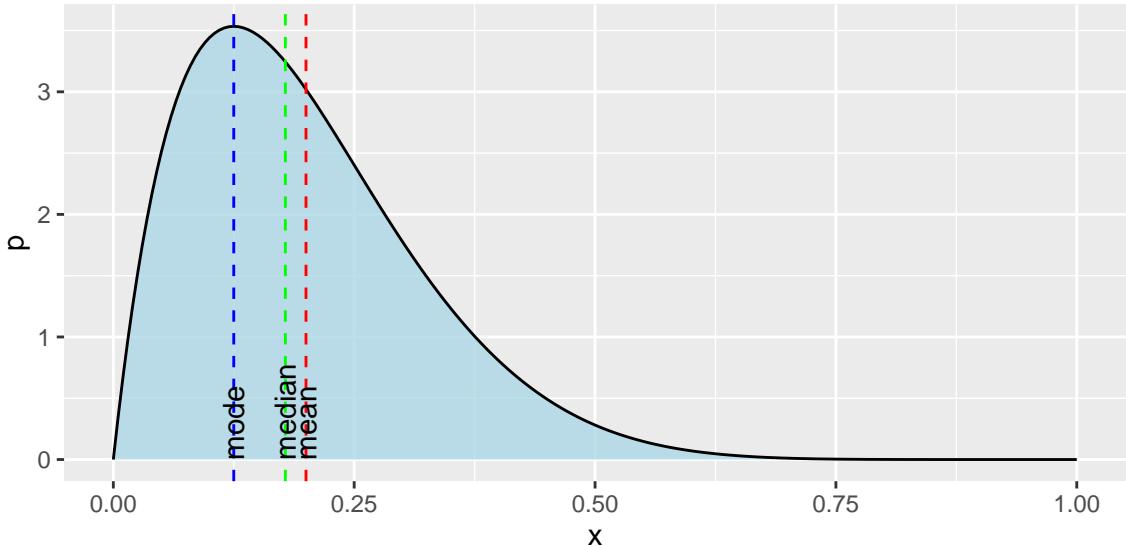
- Như vậy trung vị là nghiệm của phương trình hàm tích lũy xác suất

$$F_X(x) = 0.5$$

3.3.6. Yếu vị

Định nghĩa. **Yếu vị** (*mode*) là điểm có giá trị cực đại trong phân phối, ký hiệu là $\text{mod}(X)$. Lưu ý, Khác với *trung bình* và *trung vị*, phân bố xác suất có thể có rất nhiều điểm cực đại hoặc là không có điểm cực đại nào.

3. Biến ngẫu nhiên và hàm phân phối



3.3.7. Moment

Moment là khái niệm tổng quát của kỳ vọng và phương sai.

Định nghĩa. Một **moment** bậc k đối với a được định nghĩa như sau

$$m_k = \mathbb{E}[(X - a)^k] \quad (3.13)$$

- Kỳ vọng là moment bậc 1 với $a = 0$
- Phương sai là moment bậc 2 với $a = \mathbb{E}[X]$
- Khi $a = \mathbb{E}[X]$ người ta thường gọi là moment quy tâm, còn $a = 0$ gọi là moment gốc. Vậy nên ta có thể gọi kỳ vọng là moment gốc bậc 1 và phương sai là moment quy tâm bậc 2.

3.3.8. Entropy

Định nghĩa. Để đo sự hỗn loạn của hàm phân bố xác suất $p(x)$, người ta sử dụng độ đo entropy $H(X)$

$$H(X) = \begin{cases} -\sum_{x_i} p(x_i) \log_2 p(x_i) & \text{nếu } x \text{ rời rạc} \\ -\int_{-\infty}^{\infty} f(x) \log_2 f(x) dx & \text{nếu } x \text{ liên tục} \end{cases} \quad (3.14)$$

3.3.9. Cross entropy và KL divergence

Định nghĩa. Cross entropy và KL divergence dùng để đo sự khác biệt giữa hai phân bố xác suất $p(x)$ và $q(x)$

- Cross entropy

$$H(p, q) = -\sum_x p(x) \log_2 q(x) \quad (3.15)$$

3. Biến ngẫu nhiên và hàm phân phối

- KL divergence

$$KL(p, q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \quad (3.16)$$

3.4

Biến ngẫu nhiên nhiều chiều

Trong thực tế ta thường xuyên phải làm việc với nhiều biến ngẫu nhiên cùng lúc nên khảo sát việc kết hợp các biến như vậy là rất cần thiết. Lưu ý rằng ta cũng có thể coi các biến ngẫu nhiên này như 1 biến đa chiều hoặc 1 vector ngẫu nhiên có các phần tử là biến ngẫu nhiên. Sau này trong các bài toán chúng thường hay được biểu diễn dưới dạng vector nên ta cần nhớ tới điểm này.

3.4.1. Phân phối xác suất

Phân phối đồng thời

Định nghĩa. **Hàm phân phối xác suất đồng thời** (kết hợp) hay hàm phân phối tích lũy xác suất đồng thời (Joint CDF - Joint Cumulative Probability Distribution Function) của 2 biến ngẫu nhiên X, Y được định nghĩa như sau

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \quad x, y \in \mathbb{R} \quad (3.17)$$

Như vậy đây thực chất là hàm hợp xác suất của 2 biến ngẫu nhiên X, Y và tích lũy xác suất được lấy là phần giao tích lũy bên trái của X và bên trái của Y . Tương tự như với 1 biến ngẫu nhiên, hàm phân phối của 2 biến cũng là một hàm không giảm theo từng đối số 1 và ta còn có thể tính được tất cả các kiểu xác suất hợp của 2 biến X, Y thông qua hàm xác suất đồng thời.

- Ví dụ

$$\begin{aligned} P(X > x, Y > y) &= 1 - P(X \leq x, Y \leq y) \\ &= 1 - P(X \leq x \cup Y \leq y) \\ &= 1 - P(X \leq x \cup Y > y) \\ &= 1 - P(X \leq x) - P(Y \leq y) + P(X \leq x, Y \leq y) \\ &= 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y) \end{aligned}$$

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$

Định nghĩa. **Hàm khối xác suất đồng thời** (Joint PMF) của 2 biến ngẫu nhiên X, Y rời rạc sẽ có dạng

$$p_{X,Y}(x, y) = P(X = x, Y = y) \quad (3.18)$$

- Với mỗi $p(x_i, y_j)$ là hàm khối xác suất đồng thời, ta có:

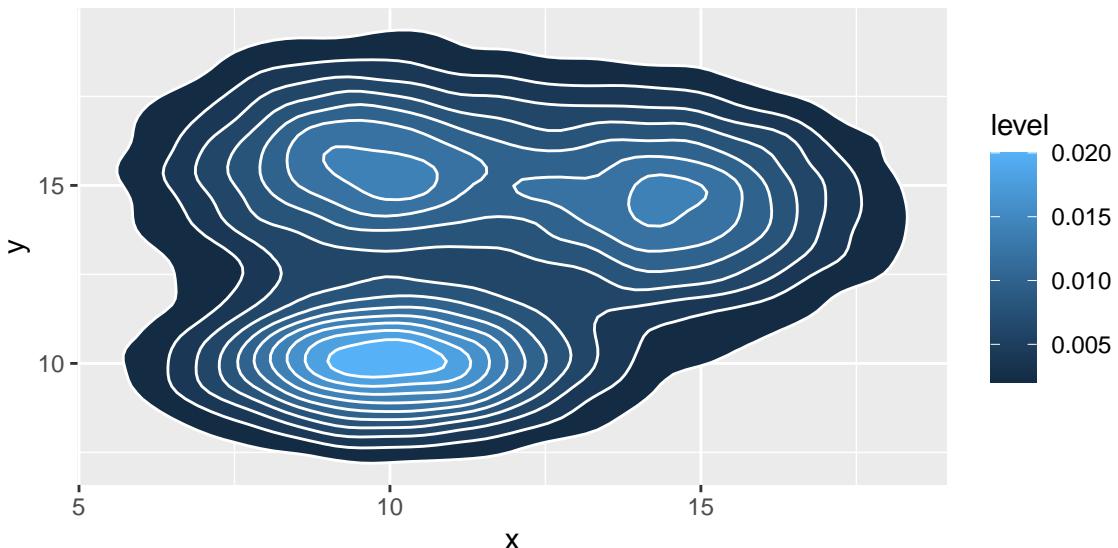
3. Biến ngẫu nhiên và hàm phân phối

1. $0 \leq p(x_i, y_j) \leq 1$
2. $\sum_{x_i} \sum_{y_j} p(x_i, y_j) = 1$
3. $F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j)$

Định nghĩa. **Hàm mật độ xác suất đồng thời** (Joint PDF) của 2 biến ngẫu nhiên X, Y cùng liên tục có dạng:

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x, y) \quad (3.19)$$

$$F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(u, v) du dv$$



- Tương tự như trường hợp 1 biến ta có:

1. $f(x, y) > 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$
3. $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y) dx dy$
4. $P(X = x, Y = y) = \int_y^y \int_x^x f(x, y) dx dy = 0$

Như vậy nếu để ý thì ta có thể nhớ 1 cách rằng trường hợp biến rời rạc ta lấy tổng còn biến là liên tục ta lấy tích phân. Natürlich là với biến rời rạc ta phải sử dụng hàm khối xác suất còn biến liên tục là hàm mật độ xác suất.

3. Biến ngẫu nhiên và hàm phân phối

Phân phối biên

Định nghĩa. Phân phối biên (Marginal Probability) là phân phối của riêng từng biến một.

$$\begin{aligned}
 F_X(x) &= P(X \leq x) \\
 &= P(X \leq x, Y < +\infty) \\
 &= F_{X,Y}(x, +\infty) \\
 F_Y(y) &= P(Y \leq y) \\
 &= P(+\infty, Y \leq y) \\
 &= F_{X,Y}(+\infty, y)
 \end{aligned} \tag{3.20}$$

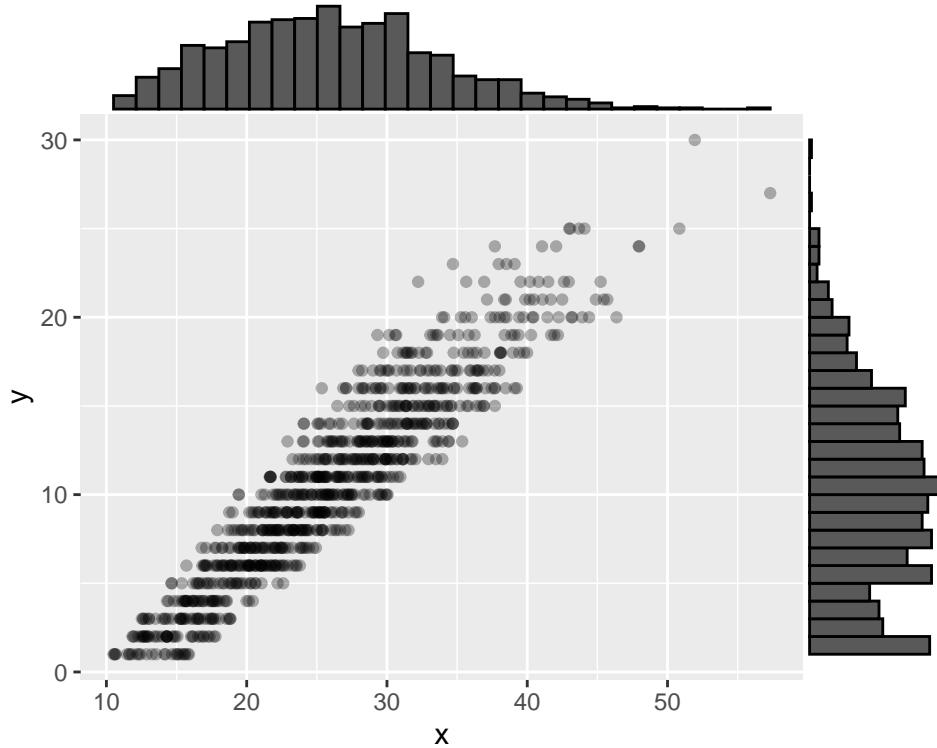
Định nghĩa. Đối với các biến rời rạc, ta có hàm khối xác suất biên (Marginal PMF)

$$\begin{aligned}
 p_X(x) &= P(X = x) \\
 &= \sum_{y_j} p(x, y_j) \\
 p_Y(y) &= P(Y = y) \\
 &= \sum_{x_i} p(x_i, y)
 \end{aligned} \tag{3.21}$$

Định nghĩa. Đối với các biến liên tục, ta có hàm mật độ xác suất biên (Marginal PDF):

$$\begin{aligned}
 f_X(x) &= P(X = x) \\
 &= \int_{-\infty}^{\infty} f(x, y) dy \\
 &= \frac{\partial}{\partial x} F_X(x) \\
 f_Y(y) &= P(Y = y) \\
 &= \int_{-\infty}^{\infty} f(x, y) dx \\
 &= \frac{\partial}{\partial y} F_Y(y)
 \end{aligned} \tag{3.22}$$

3. Biến ngẫu nhiên và hàm phân phối



Biến độc lập

Định nghĩa. 2 biến X, Y độc lập khi xác suất của chúng không phụ thuộc vào nhau. Như ta đã biết 2 sự kiện A, B độc lập khi và chỉ khi $P(AB) = P(A)P(B)$, tương tự với biến ngẫu nhiên chúng độc lập khi và chỉ khi

$$F_{X,Y}(x,y) = F_X(x)F_Y(y), \forall x, y \in \mathbb{R} \quad (3.23)$$

- Với trường hợp các biến ngẫu nhiên rời rạc

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \forall x, y \in D \quad (3.24)$$

- Với trường hợp các biến ngẫu nhiên liên tục

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R} \quad (3.25)$$

Như vậy từ đây ta có thể thấy rằng nếu các biến ngẫu nhiên là độc lập thì xác suất đồng thời của chúng có thể tính qua các xác suất riêng lẻ của chúng bằng cách lấy tích chúng lại với nhau.

Phân phối có điều kiện

Định nghĩa. Tương tự như xác suất có điều kiện của các sự kiện ta cũng có thể biểu diễn các phân phối có điều kiện của các biến ngẫu nhiên.

3. Biến ngẫu nhiên và hàm phân phối

- Với X, Y là rời rạc:

$$\begin{aligned}
 p_{X|Y}(x | y) &= P(X = x | Y = y) \\
 &= \frac{P(X = x, Y = y)}{P(Y = y)} \\
 &= \frac{p_{X,Y}(x, y)}{p_Y(y)} \\
 &= \frac{p_{X,Y}(x, y)}{\sum_{x_i} p(x_i, y)}
 \end{aligned} \tag{3.26}$$

- Tương tự với X, Y là liên tục, ta có:

$$\begin{aligned}
 f_{X|Y}(x | y) &= \frac{f_{X,Y}(x, y)}{f_Y(y)} \\
 &= \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f(x, y) dy}
 \end{aligned} \tag{3.27}$$

Qua các phép biến đổi trên ta thấy rằng hợp phân phối của các biến ngẫu nhiên có thể tính toán tương tự như các phép kết hợp của các sự kiện mà ta đã làm quen ở bài đầu tiên. Tôi sẽ không trình bày thêm các phép toán như nhân xác suất, công thức xác suất hậu nghiệm ở đây nữa vì chúng được áp dụng tương tự như các phép toán trên. Cụ thể hơn bạn xem lại bài đầu tiên nhé.

3.4.2. Các đặc trưng

Hiệp phương sai

Định nghĩa. **Hiệp phương sai** (*covariance*) của 2 biến ngẫu nhiên X, Y kí hiệu là $\text{Cov}(X, Y)$ được định nghĩa rằng:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{3.28}$$

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{3.29}$$

Nếu X, Y là độc lập thì $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ nên lúc này ta có $\text{Cov}(X, Y) = 0$, nhưng điều ngược lại chưa chắc đã đúng!

Tính chất.

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2. $\text{Cov}(X, X) = \text{Var}(X)$
3. $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$ với a, b là hằng số
4. $\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$

3. Biến ngẫu nhiên và hàm phân phối

$$5. \ Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) + 2 \sum_i \sum_{j>i} Cov(X_i, X_j)$$

Định nghĩa. Các giá trị hiệp phương sai thường được tập hợp lại thành **ma trận hiệp phương sai**

$$\begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{bmatrix} = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{bmatrix} \quad (3.30)$$

Tính chất.

1. Ma trận hiệp phương sai là ma trận đối xứng.
2. Ma trận hiệp phương sai là ma trận của dạng toàn phương không âm.

Hệ số tương quan

Định nghĩa. Hiệp phương sai có hạn chế cơ bản là khó xác định được miền biến thiên, nó thay đổi từ cặp biến thiên này sang cặp biến thiên khác. Chưa kể về mặt vật lý nó có đơn vị đo bằng bình phương đơn vị đo của biến ngẫu nhiên X, Y (nếu chúng cùng đơn vị đo). Vì thế cần đưa ra một số đặc trưng khác để khắc phục hạn chế này, đó là **hệ số tương quan** (Correlation) được kí hiệu là $\rho(X, Y)$ và định nghĩa như sau

$$\begin{aligned} \rho(X, Y) &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)} \end{aligned} \quad (3.31)$$

Tính chất.

1. $-1 \leq \rho(X, Y) \leq 1$.
2. Nếu $|\rho(X, Y)| = 1$ thì giữa X, Y có quan hệ tuyến tính tức là: $Y = a + bX$.
3. Nếu $\rho(X, Y) = 0$ thì chúng không tương quan với nhau.

Hệ số tương quan cho ta thấy được các biến ngẫu nhiên có quan hệ tuyến tính chặt tới đâu tức là khi 1 biến biến thiên thì biến còn lại cũng sẽ biến thiên tương ứng. $|\rho(X, Y)|$ càng lớn thì ta nói rằng 2 biến có quan hệ tuyến tính càng chặt chẽ. $\rho(X, Y) > 0$ ám chỉ rằng 2 biến là thuận biến với nhau, còn $\rho(X, Y) < 0$ ám chỉ rằng 2 biến là nghịch biến với nhau. Lưu ý rằng khi 2 biến độc lập thì chúng không tương quan nhưng điều ngược lại thì không đúng.

3.4.3. Đặc trưng có điều kiện

Các hàm khối xác suất của biến rời rạc và hàm mật độ xác suất của biến liên tục có điều kiện cũng có các đặc trưng như kỳ vọng và phương sai tương tự như các hàm khác chỉ khác 1 điều là thêm điều kiện tương ứng.

3. Biến ngẫu nhiên và hàm phân phối

Định nghĩa. Kỳ vọng có điều kiện (*conditional expectation*) được định nghĩa như sau

$$\mathbb{E}[X | Y = y] = \begin{cases} \sum_{x_i} x_i p_{X|Y}(x_i | y) & \text{cho } X, Y \text{ rời rạc} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx & \text{cho } X, Y \text{ liên tục} \end{cases} \quad (3.32)$$

Tính chất.

1. $\mathbb{E}[g(Y)X | Y] = g(Y)\mathbb{E}[X | Y]$ với $g(Y)$ là 1 hàm liên tục
2. $\mathbb{E}[X_1 + X_2 | Y] = \mathbb{E}[X_1 | Y] + \mathbb{E}[X_2 | Y]$
3. $\mathbb{E}[X | Y] = \mathbb{E}[X]$ nếu X, Y là độc lập

Một tính chất quan trọng nữa là

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X | Y]], \quad (3.33)$$

tức là kỳ vọng của 1 biến có thể lấy bằng cách tương tự như xác suất của nó. Ta có thể biểu diễn tương tự như vậy:

$$\mathbb{E}[X] = \begin{cases} \sum_j \mathbb{E}(X | Y = y_j) p_Y(y_j) & \text{cho } X, Y \text{ rời rạc} \\ \int_{-\infty}^{\infty} \mathbb{E}(X | Y = y) f_Y(y) dy & \text{cho } X, Y \text{ liên tục} \end{cases} \quad (3.34)$$

Định nghĩa. Phương sai có điều kiện (*conditional variance*) được định nghĩa như sau

$$\text{Var}[X | Y = y] = \begin{cases} \sum_{x_i} (x_i - \mathbb{E}[X | Y = y])^2 p_{X|Y}(x_i | y) & \text{cho } X, Y \text{ rời rạc} \\ \int_{-\infty}^{\infty} (x - \mathbb{E}[X | Y = y])^2 f_{X|Y}(x | y) dx & \text{cho } X, Y \text{ liên tục} \end{cases} \quad (3.35)$$

Tính chất.

1. $\text{Var}(X | Y) = \mathbb{E}[(X - \mathbb{E}[X])^2 | Y]$
2. $\text{Var}(X | Y) = \mathbb{E}[X^2 | Y] - \mathbb{E}^2[X | Y]$

Tính chất phân rã một phương sai

$$\text{Var}(X) = \mathbb{E}_Y[\text{Var}(X | Y)] + \text{Var}_Y(\mathbb{E}[X | Y]). \quad (3.36)$$

Chứng minh. Do $\text{Var}(X | Y) = \mathbb{E}[X^2 | Y] - \mathbb{E}^2[X | Y]$, nên

$$\begin{aligned} \mathbb{E}[\text{Var}(X | Y)] &= \mathbb{E}[\mathbb{E}[X^2 | Y]] - \mathbb{E}[\mathbb{E}^2[X | Y]] \\ &= \mathbb{E}[X^2] - \mathbb{E}[\mathbb{E}^2[X | Y]] \end{aligned}$$

3. Biến ngẫu nhiên và hàm phân phối

Ngoài ra,

$$\begin{aligned}\text{Var}(\mathbb{E}[X | Y]) &= \mathbb{E}[\mathbb{E}^2[X | Y]] - \mathbb{E}^2[\mathbb{E}[X | Y]] \\ &= \mathbb{E}[\mathbb{E}^2[X | Y]] - \mathbb{E}^2[X]\end{aligned}$$

Vậy nên,

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y])$$

■

Các đặc trưng có điều kiện là nền tảng để có tạo được mối quan hệ giữa các biến ngẫu nhiên tức là ta có thể vẽ một đường dự đoán giá trị của Y khi biết X bằng một hàm hồi quy của Y đối với X : $g(X) = \mathbb{E}[Y | X]$. Thực chất để ước lượng được Y ta cần tìm hàm hồi quy $g(X)$ sao cho kỳ vọng khoảng cách của chúng là nhỏ nhất có thể $\text{argmin} \mathbb{E}[(Y - g(X))^2]$. Và giá trị nhỏ nhất này chính là $g(X) = \mathbb{E}[Y | X]$.

3.5

Hàm và biến ngẫu nhiên

3.5.1. Hàm của biến ngẫu nhiên

Gọi X là một biến ngẫu nhiên, với hàm phân bố xác suất $F_X(x)$ và g là hàm đơn điệu từ \mathbb{R} đến \mathbb{R} . Xét biến ngẫu nhiên $Y = g(X)$ có phân phối $F_Y(y)$

Tính chất.

1. Nếu g là hàm đồng biến thì

$$F_Y(y) = F_X(g^{-1}(y)) \quad (3.37)$$

2. Nếu g là hàm nghịch biến thì

$$F_Y(y) = 1 - F_X(g^{-1}(y)) \quad (3.38)$$

3. Nếu X là biến ngẫu nhiên liên tục với hàm mật độ xác suất $f_X(x)$ thì

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \quad (3.39)$$

Trong trường hợp g không phải là hàm đơn điệu, ta có thể phân rã không gian mẫu ra thành từng phân hoạch sao cho g đơn điệu trên từng phân hoạch.

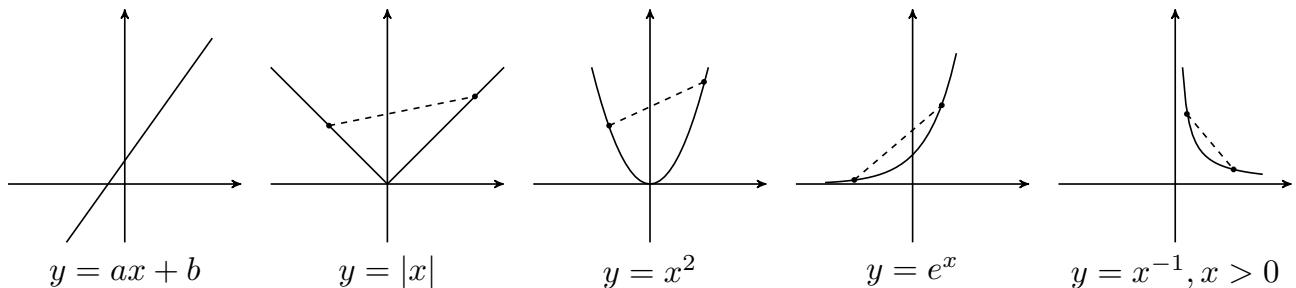
3. Biến ngẫu nhiên và hàm phân phối

3.5.2. Hàm lồi và biến ngẫu nhiên

Khái niệm 3.6. Một hàm số $f : \mathbb{R} \rightarrow \mathbb{R}$ được gọi là một hàm lồi (convex function) nếu với mọi x, y và $\lambda \in [0, 1]$ ta có

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad (3.40)$$

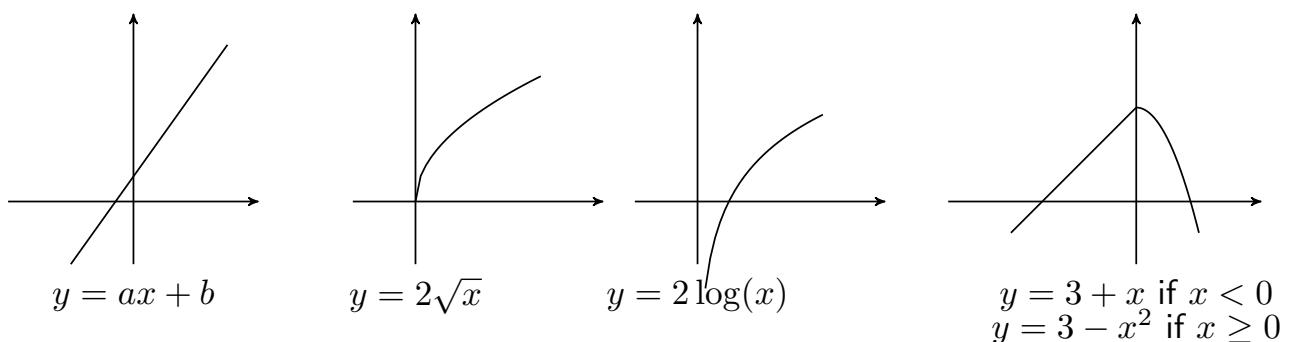
Một số hàm lồi đơn giản



Khái niệm 3.7. Một hàm số $f : \mathbb{R} \rightarrow \mathbb{R}$ được gọi là một hàm lõm (concave function) nếu với mọi x, y và $\lambda \in [0, 1]$ ta có

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) \quad (3.41)$$

Một số hàm lõm đơn giản



Định lý 3.1 (Bất đẳng thức Jensen). Cho X là biến ngẫu và f là một hàm lồi thì ta có

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)] \quad (3.42)$$

Chứng minh. Dành cho các bạn

3.6

Luật số lớn và định lý giới hạn trung tâm

3.6.1. Luật số lớn

Định lý 3.2 (Bất đẳng thức Markov). Cho X là biến ngẫu nhiên không âm. Khi đó với $\epsilon > 0$ tùy ý cho trước ta có:

$$P(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon} \quad (3.43)$$

3. Biến ngẫu nhiên và hàm phân phối

Chứng minh. Từ định nghĩa kỳ vọng của X ta có

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \int_0^{\infty} xf(x)dx \\
&\geq \int_{\epsilon}^{\infty} xf(x)dx \\
&\geq \int_{\epsilon}^{\infty} \epsilon f(x)dx \\
&= \epsilon \int_{\epsilon}^{\infty} f(x)dx \\
&= \epsilon P(X \geq \epsilon)
\end{aligned} \tag{3.44}$$

Suy ra điều phải chứng minh ■

Hệ quả 3.1 (Bất đẳng thức Chebyshev). *Cho X là biến ngẫu nhiên có $\mathbb{E}[X] = \mu$, $\text{Var}(X) = \sigma^2$ hữu hạn. Khi đó với $\epsilon > 0$ tùy ý cho trước ta có:*

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \tag{3.45}$$

hay tương đương

$$P(|X - \mu| < \epsilon) > 1 - \frac{\sigma^2}{\epsilon^2} \tag{3.46}$$

Chứng minh. Xét biến ngẫu nhiên $(X - \mathbb{E}[X])^2$ không âm, theo bất đẳng thức Markov ta có

$$P((X - \mathbb{E}[X])^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\epsilon^2}$$

Suy ra ta có

$$P(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2} \tag{3.47}$$

Vậy ta có điều phải chứng minh ■

Bổ đề 3.1. *X là biến ngẫu nhiên bị chặn trong đoạn $[a, b]$ thì*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda\mu + \frac{\lambda^2(b-a)^2}{8}\right) \tag{3.48}$$

Chứng minh. Dành cho các bạn ■

Định lý 3.3 (Bất đẳng thức Hoeffding). *Giả sử $\{X_i\}$ là dãy biến ngẫu nhiên độc lập bị chặn trong khoảng $[0, 1]$ với $i \in \{1, \dots, n\}$. Đặt $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Khi đó với n đủ lớn ta có*

$$P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \epsilon) \leq 2e^{-2\epsilon^2 n} \tag{3.49}$$

3. Biến ngẫu nhiên và hàm phân phối

Chứng minh. Dành cho các bạn ■

Định lý 3.4 (Luật số lớn yếu). *Giả sử $\{X_i\}$ là dãy biến ngẫu nhiên độc lập cùng phân phối với $\mathbb{E}[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$ với mọi $i \in \{1, \dots, n\}$. Đặt $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, khi đó ta có với $\epsilon > 0$*

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0 \quad (3.50)$$

Hay nói cách khác là biến ngẫu nhiên \bar{X} hội tụ theo xác suất về μ

Chứng minh. Dành cho các bạn ■

3.6.2. Định lý giới hạn trung tâm

Định lý 3.5. *Giả sử $\{X_i\}$ là dãy biến ngẫu nhiên độc lập cùng phân phối với $\mathbb{E}[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$ với mọi $i \in \{1, \dots, n\}$. Đặt $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, khi đó với n đủ lớn ta có*

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad (3.51)$$

Chứng minh. Ta có với mọi $i = 1, \dots, n$, $Y_i = \frac{X_i - \mu}{\sigma}$ có kỳ vọng $\mathbb{E}(Y_i) = 0$ và phương sai $\text{Var}(Y_i) = 1$, với *hàm đặc trưng* (*moment generating function*) số thực được khai triển dưới dạng

$$M_{Y_1}(t) = \mathbb{E}[e^{tY_1}] = 1 + \frac{t^2}{2} + o(t^2), \quad (t \rightarrow 0). \quad (3.52)$$

Đặt

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}}. \quad (3.53)$$

Từ các tính chất cơ bản của hàm đặc trưng, ta suy ra hàm đặc trưng của Z_n là

$$\begin{aligned} M_{Z_n}(t) &= \mathbb{E}[e^{tZ_n}] \\ &= \mathbb{E}\left[e^{t\frac{Y_1+Y_2+\dots+Y_n}{\sqrt{n}}}\right] \\ &= \mathbb{E}\left[e^{t\frac{Y_1}{\sqrt{n}}}\right] \mathbb{E}\left[e^{t\frac{Y_2}{\sqrt{n}}}\right] \dots \mathbb{E}\left[e^{t\frac{Y_n}{\sqrt{n}}}\right] \\ &= \mathbb{E}\left[e^{t\frac{Y_1}{\sqrt{n}}}\right]^n \\ &= \left[M_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \rightarrow e^{\frac{t^2}{2}} \quad \text{khi } n \rightarrow \infty. \end{aligned} \quad (3.54)$$

Giới hạn này chính là hàm đặc trưng của phân phối chuẩn $\mathcal{N}(0, 1)$. Vậy ta có điều phải chứng minh ■

3. Biến ngẫu nhiên và hàm phân phối

Bài tập

B 3.1. Cho biến ngẫu nhiên rời rạc X có bảng phân phối xác suất cho bởi bảng sau:

X	-2	-1	0	1	2
P	$1/8$	$2/8$	$2/8$	$2/8$	$1/8$

1. Tìm hàm phân phối xác suất $F(x)$.
2. Tính $P(-1 \leq X \leq 1)$ và $P(X \leq -1$ hoặc $X = 2)$.
3. Lập bảng phân phối xác suất của biến ngẫu nhiên $Y = X^2$.

B 3.2. Biến ngẫu nhiên rời rạc X có hàm xác suất cho bởi

$$f(x) = \frac{2x+1}{25}, x = 0, 1, 2, 3, 4$$

1. Lập bảng phân phối xác suất của X .
2. Tính $P(2 \leq X < 4)$ và $P(X > -10)$.

B 3.3. Gọi X là biến ngẫu nhiên rời rạc có bảng phân phối xác suất sau

X	-1	0	3
P	0.5	0.2	0.3

1. Tính độ lệch chuẩn của X .
2. Tính kỳ vọng của X^3 .
3. Tìm hàm phân phối của X .
4. Ta định nghĩa $Y = X^2 + X + 1$. Lập bảng phân phối xác suất của Y .

B 3.4. Biến ngẫu nhiên X có hàm mật độ $f(x)$ như sau

$$f(x) = \begin{cases} kx(2-x) & \text{khi } 1 < x < 2 \\ 0 & \text{nơi khác} \end{cases}$$

1. Xác định giá trị của k để $f(x)$ là hàm mật độ của biến ngẫu nhiên X . Với k vừa tìm được tính kỳ vọng và phương sai của biến ngẫu nhiên X .
2. Tìm hàm phân phối $F(x)$ của biến ngẫu nhiên X .
3. Tìm hàm phân phối $G(y)$ của biến ngẫu nhiên $Y = X^3$.

3. Biến ngẫu nhiên và hàm phân phối

B 3.5. Biến ngẫu nhiên liên tục X có hàm mật độ

$$f(x) = \begin{cases} e^{-x} & \text{khi } x > 0 \\ 0 & \text{khi } x \leq 0 \end{cases}$$

1. Tính $P(3 \leq X)$.
2. Tìm giá trị của a sao cho $P(X \leq a) = 0.1$.
3. Xác định hàm phân phối và mật độ xác suất của biến ngẫu nhiên $Y = \sqrt{X}$.

B 3.6. Tính $P(X \geq 8)$ nếu

$$f_X(x) = \begin{cases} \frac{1}{96}x^3e^{-x/2} & \text{nếu } x \geq 0 \\ 0 & \text{nếu khác} \end{cases}$$

B 3.7. Cho

$$f_X(x) = \sqrt{\frac{2}{\pi}} - x^2 \text{ với } -\sqrt{\frac{2}{\pi}} \leq x \leq \sqrt{\frac{2}{\pi}}$$

Tính $P(X < 0)$.

B 3.8. Biến ngẫu nhiên X có hàm mật độ

$$f(x) = \begin{cases} a \exp\left(-\frac{x}{2}\right) & \text{khi } x \geq 0 \\ 0 & \text{nơi khác} \end{cases}$$

Xác định:

1. Hằng số a .
2. Hàm phân phối xác suất $F(x)$
3. Kỳ vọng và phương sai của biến ngẫu nhiên X .
4. Kỳ vọng và phương sai của biến ngẫu nhiên $Y = (X/2) - 1$.

B 3.9. Cho X là biến ngẫu nhiên có hàm mật độ sau

$$f(x) = \begin{cases} c(1-x^2) & \text{nếu } -1 \leq x \leq 1 \\ 0 & \text{nếu } |x| \geq 1 \end{cases}$$

với c là một hằng số dương. Tìm

1. hằng số c

3. Biến ngẫu nhiên và hàm phân phối

2. trung bình của X
3. phương sai của X
4. hàm phân phối $F_X(x)$.

B 3.10. Biến ngẫu nhiên liên tục X có hàm mật độ

$$f(x) = \begin{cases} \frac{1}{2}x & \text{khi } 0 < x < 2 \\ 0 & \text{nơi khác} \end{cases}$$

Tìm hàm phân phối và hàm mật độ xác suất của các biến ngẫu nhiên sau:

1. $Y = X(2 - X)$.
2. $Z = 4 - X^3$.
3. $T = 3X + 2$.

B 3.11. Tính phương sai của \sqrt{X} nếu

$$p_X(x) = \begin{cases} 1/4 & \text{nếu } x = 0 \\ 1/2 & \text{nếu } x = 1 \\ 1/4 & \text{nếu } x = 4 \end{cases}$$

B 3.12. Tính phân vị mức 25% (tức là giá trị $x_{0.25}$ sao cho $P(X < x_{0.25}) = 0.25$) của biến ngẫu nhiên liên tục X có hàm mật độ sau:

$$f_X(x) = \begin{cases} xe^{-x^2/2} & \text{nếu } x \geq 0 \\ 0 & \text{nếu } x < 0 \end{cases}$$

B 3.13. Cho

$$p_X(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ x/2 & \text{nếu } 0 \leq x \leq 1 \\ x/6 + 1/3 & \text{nếu } 1 < x < 4 \\ 1 & \text{nếu } x \geq 4 \end{cases}$$

là hàm phân phối của biến ngẫu nhiên liên tục X .

1. Tính hàm mật độ của X .
2. Tìm phân vị mức 75% của X (tức là tìm $x_{0.75}$ sao cho $P(X < x_{0.75}) = 0.75$).
3. Tính kì vọng của X .

3. Biến ngẫu nhiên và hàm phân phối

4. Tính $\mathbb{E}(1/X)$.

5. Ta định nghĩa

$$Y = \begin{cases} -1 & \text{nếu } X \leq 1 \\ 1 & \text{nếu } X > 1 \end{cases}$$

a) Tìm $F_Y(0)$.

b) Tính phương sai của Y .

B 3.14. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất

$$f(x) = \begin{cases} \frac{3}{4}x(2-x) & \text{khi } 0 \leq x \leq 2 \\ 0 & \text{nơi khác} \end{cases}$$

1. Xác định hàm phân phối xác suất $F(x)$ của biến ngẫu nhiên X .

2. Tính $\mathbb{E}(X)$, $\text{Var}(X)$ và trung vị của biến ngẫu nhiên X .

3. Đặt $Y = \sqrt{X}$, xác định hàm phân phối và hàm mật độ xác suất của biến ngẫu nhiên Y .

B 3.15. Tuổi thọ của một loại côn trùng nào đó là một biến ngẫu nhiên liên tục X (đơn vị tháng) có hàm mật độ

$$f(x) = \begin{cases} kx^2(4-x) & \text{khi } 0 \leq x \leq 4 \\ 0 & \text{nơi khác} \end{cases}$$

1. Tìm hằng số k .

2. Tìm $F(x)$.

3. Tìm $\mathbb{E}(X)$, $\text{Var}(X)$ và $\text{Mod}(X)$.

4. Tính xác suất để côn trùng chết trước một tháng tuổi.

B 3.16. Biến ngẫu nhiên liên tục X có hàm mật độ

$$f(x) = \begin{cases} kx^2e^{-2x} & \text{khi } x \geq 0 \\ 0 & \text{nơi khác} \end{cases}$$

1. Tìm hằng số k .

2. Tìm hàm phân phối xác suất $F(x)$.

3. Tìm $\mathbb{E}(X)$, $\text{Var}(X)$ và $\text{Mod}(X)$.

B 3.17. Có hai thùng thuỷ tinh A và B , trong đó:

3. Biến ngẫu nhiên và hàm phân phối

- thùng A có 20 lọ gồm 2 lọ hỏng và 18 lọ tốt
 - thùng B có 20 lọ gồm 3 lọ hỏng và 17 lọ tốt.
1. Lấy ở mỗi thùng 1 lọ. Gọi X là số lọ hỏng trong hai lọ lấy ra. Tìm hàm mật độ của X .
 2. Lấy ở thùng B ra 3 lọ. Gọi Y là số lọ hỏng trong 3 lọ lấy ra. Tìm hàm mật độ của Y .

B 3.18. Một thùng đựng 10 lọ thuốc trong đó có 1 lọ hỏng. Ta kiểm tra từng lọ (không hoàn lại) cho tới khi phát hiện được lọ hỏng thì dừng. Gọi X là số lần kiểm tra. Tìm hàm mật độ của X . Tính kì vọng và phương sai.

B 3.19. Một biến ngẫu nhiên liên tục có hàm mật độ xác suất sau:

$$f_X(x) = \begin{cases} cxe^{-x/2} & \text{nếu } x \geq 0 \\ 1 & \text{nếu } x < 0 \end{cases}$$

1. Tìm hằng số c .
2. Tìm hàm phân phối xác suất $F_X(x)$.
3. Tìm trung bình của X
4. Tìm độ lệch chuẩn của X .
5. Tìm $\text{Med}(X)$.

B 3.20. Gọi X là tuổi thọ của con người. Một công trình nghiên cứu cho biết hàm mật độ của X là

$$f(x) = \begin{cases} cx^2(100-x)^2 & \text{khi } 0 \leq x \leq 100 \\ 0 & \text{khi } x < 0 \text{ hay } x > 100 \end{cases}$$

1. Xác định hằng số c .
2. Tính kì vọng và phương sai của X .
3. Tính xác suất của một người có tuổi thọ ≥ 60
4. Tính xác suất của một người có tuổi thọ ≥ 60 , biết rằng người đó hiện nay đã 50 tuổi.

B 3.21. Một thiết bị gồm 3 bộ phận hoạt động độc lập với nhau, xác suất trong khoảng thời gian t các bộ phận hỏng tương ứng bằng 0.2; 0.3; 0.25. Gọi X là số bộ phận bị hỏng trong khoảng thời gian t .

1. Lập bảng phân phối xác suất của X .
2. Viết biểu thức hàm phân phối của X .

3. Biến ngẫu nhiên và hàm phân phối

3. Tính $P(0 < X \leq 4)$ theo hai cách.

B 3.22. Một mẫu 4 sản phẩm được rút ra không hoàn lại từ 10 sản phẩm. Biết rằng trong 10 sản phẩm này có 1 thứ phẩm. Tính xác suất thứ phẩm có trong mẫu.

B 3.23. Một cái hộp chứa 100 transistor loại A và 50 transistor loại B.

1. Các transistor được rút ra lần lượt, ngẫu nhiên và được hoàn lại, cho đến khi lấy được transistor loại B đầu tiên. Tính xác suất 9 hoặc 10 transistor được rút ra.
2. Số lượng các transistor ít nhất phải rút ra, ngẫu nhiên và được hoàn lại, là bao nhiêu nếu ta muốn xác suất lấy được chỉ loại A nhỏ hơn $1/3$?

B 3.24. Gọi X là số lần măt nhất xuất hiện sau ba lần tung một con xúc xắc.

1. Lập bảng phân phối xác suất của X .
2. Tính xác suất có ít nhất một lần được măt nhất.
3. Tính xác suất có tối đa hai lần măt nhất.
4. Tính $\mathbb{E}(X)$, $\text{Var}(X)$

B 3.25. Xét trò chơi, tung một con xúc xắc ba lần: nếu cả ba lần được 6 nút thì lĩnh 6 ngàn đ, nếu hai lần 6 nút thì lĩnh 4 ngàn đ, một lần 6 nút thì lĩnh 2 ngàn đ, và nếu không có 6 nút thì không lĩnh gì hết. Mỗi lần chơi phải đóng A ngàn đ. Hỏi :

1. A là bao nhiêu thì người chơi về lâu về dài huê vốn (gọi là trò chơi công bằng).
2. A là bao nhiêu thì trung bình mỗi lần người chơi mất 1 ngàn đ.

B 3.26. Một hệ thống an ninh gồm có 10 thành phần hoạt động độc lập lẫn nhau. Hệ thống hoạt động nếu ít nhất 5 thành phần hoạt động. Để kiểm tra hệ thống có hoạt động hay không, người ta kiểm tra định kì 4 thành phần được chọn ngẫu nhiên (không hoàn lại). Hệ thống được báo cáo là hoạt động nếu ít nhất 3 trong 4 thành phần được kiểm tra hoạt động. Nếu thật sự chỉ có 4 trong 10 thành phần hoạt động, thì xác suất hệ thống được báo cáo là hoạt động là bao nhiêu?

B 3.27. Trong một trò chơi ném phi tiêu, người chơi hướng về một tâm bia lớn có vẽ một vòng tròn có bán kính 25 cm. Gọi X là khoảng cách (theo cm) giữa đầu phi tiêu cắm vào bia và tâm vòng tròn. Giả sử rằng

$$P(X \leq x) = \begin{cases} c\pi x^2 & \text{nếu } 0 \leq x < 25 \\ 1 & \text{nếu } x \geq 25 \end{cases}$$

với c là một hằng số nào đó.

1. Tính

3. Biến ngẫu nhiên và hàm phân phối

- a) hằng số c
 b) hàm mật độ, $f_X(x)$, của X
 c) trung bình của X
 d) xác suất $P(X \leq 10 | X \geq 5)$.
2. Người chơi sẽ mất 1 (đơn vị: ngàn đồng) cho mỗi lần phỏng và thắng

$$\begin{cases} 10 & \text{nếu } x \leq r \\ 1 & \text{nếu } r < x \leq 2r \\ 0 & \text{nếu } 2r < X < 25 \end{cases}$$

Với giá trị nào của r thì số tiền trung bình người chơi đạt được bằng 0.25?

B 3.28. Cho X là một đại lượng ngẫu nhiên có phân phối xác suất như sau

X	0	1	2	3	4	5	6	7
P	0	a	$2a$	$2a$	$3a$	a^2	$2a^2$	$7a^2 + a$

1. Xác định a
 2. Tính $P(X \geq 5)$, $P(X < 3)$.
 3. Tính k nhỏ nhất sao cho $P(X \leq k) \geq \frac{1}{2}$

B 3.29. Cho hàm mật độ của biến ngẫu nhiên X có dạng

$$f(x) = \begin{cases} Ax & \text{khi } x \in [0, 1] \\ 0 & \text{khi } x \notin [0, 1] \end{cases}$$

$$f(x) = \begin{cases} A \sin x & \text{khi } x \in [0, \pi] \\ 0 & \text{khi } x \notin [0, \pi] \end{cases}$$

$$f(x) = \begin{cases} A \cos \pi x & \text{khi } x \in [0, \frac{1}{2}] \\ 0 & \text{khi } x \notin [0, \frac{1}{2}] \end{cases}$$

$$f(x) = \begin{cases} \frac{A}{x^4} & \text{khi } x \geq 1 \\ 0 & \text{khi } x < 1 \end{cases}$$

Hãy xác định A . Tìm hàm phân phối xác suất của X . Tính $\mathbb{E}(X)$, $\text{Var}(X)$ nếu có.

3. Biến ngẫu nhiên và hàm phân phối

B 3.30. Cho biến ngẫu nhiên liên tục X có hàm phân phối

$$f(x) = \begin{cases} 0 & \text{khi } x < -\frac{\pi}{2} \\ a + b \sin x & \text{khi } -\frac{\pi}{2} \leq x \leq \frac{\pi}{2} \\ 1 & \text{khi } x > \frac{\pi}{2} \end{cases}$$

1. Tìm a và b .
2. Với a và b tìm được, tính hàm mật độ $f(x)$ của X và $\text{Mod}(X)$, $\text{Med}(X)$, $P(X > \frac{\pi}{4})$.

B 3.31. Cho X và Y là hai biến ngẫu nhiên độc lập và có phân phối xác suất tương ứng là

X	-1	0	1	2	và	Y	-1	0	1
P	0.2	0.3	0.3	0.2		P	0.3	0.4	0.3

Tìm phân phối xác suất của X^2 , $X + Y$. Tính kì vọng, phương sai của X , $X + Y$.

B 3.32. Một mẫu gồm 4 biến ngẫu nhiên X_1, X_2, X_3, X_4 độc lập với nhau từng đôi một. Mỗi biến ngẫu nhiên $X_i, i = 1, \dots, 4$ có hàm mật độ như sau:

$$f(x) = \begin{cases} 2x & \text{khi } 0 < x < 1 \\ 0 & \text{nơi khác} \end{cases}$$

Đặt $Y = \max\{X_1, X_2, X_3, X_4\}$ và $Z = \min\{X_1, X_2, X_3, X_4\}$. Tìm hàm mật độ của Y và Z .

B 3.33. Cho F_X là hàm phân phối xác suất của biến ngẫu nhiên X . Tìm hàm phân phối xác suất của biến ngẫu nhiên

$$Y = \begin{cases} \frac{X}{|X|} & \text{khi } X \neq 0 \\ 1 & \text{khi } X = 0 \end{cases}$$

B 3.34. Tìm hàm phân phối của $\frac{1}{2}(X + |X|)$ nếu hàm phân phối của X là F_X .

B 3.35. Giả sử X có hàm phân phối liên tục $F(x)$. Xác định hàm phân phối của $Y = F(X)$.

B 3.36. Giả sử $F(x)$ là hàm phân phối của biến ngẫu nhiên dương liên tục X , có tính chất $P(X < t + x \mid X > t) = P(X < x)$ với $x, t > 0$. Chứng minh rằng $F(x) = 1 - e^{-\lambda x}$ với $x > 0$.

Một số phân phối phổ biến

Cho tới thời điểm này ta đã có các khái niệm quan trọng trong xác suất như sự kiện, biến ngẫu nhiên, phân phối xác suất và các đặc trưng của phân phối. Giờ là lúc ta đề cập tới một số phân phối xác suất phổ biến để có thể áp dụng vào thực tế khi quan sát các mô hình xác suất.

4.1 Biến ngẫu nhiên rời rạc

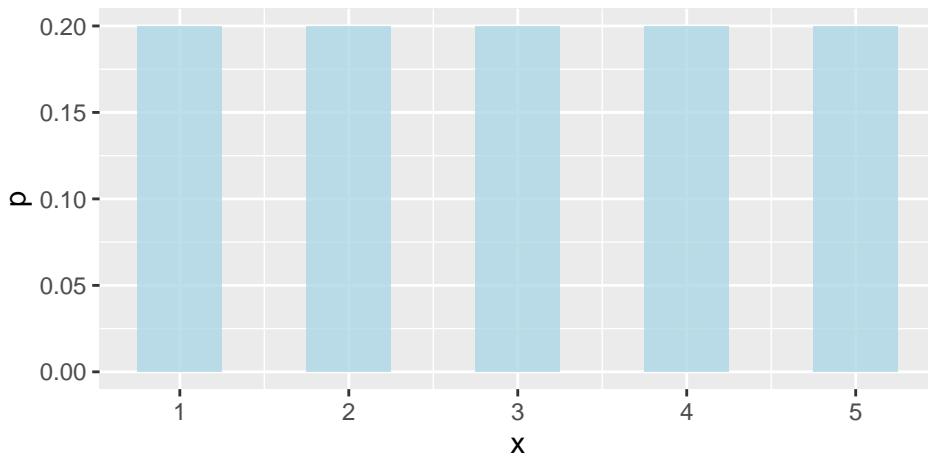
4.1.1. Phân phối đều (discrete uniform distribution)

Định nghĩa. Phân phối đều là phân phối mà xác suất xuất hiện của các sự kiện là như nhau. Biến ngẫu nhiên X tuân theo phân phối đều rời rạc $X \sim Uni(a, b)$ với tham số $a, b \in \mathbb{Z}; a < b$ là khoảng giá trị của X , đặt $n = b - a + 1$, ta có

Hàm/đại lượng	Công thức/giá trị
Tham số	$a, b \in \mathbb{N}$
PMF	$p(x) = \frac{1}{n}, \forall x \in [a, b]$
CDF	$F(x) = \frac{x - a + 1}{n}, \forall x \in [a, b]$
Kỳ vọng	$\mathbb{E}[X] = \frac{a + b}{2}$
Phương sai	$Var(X) = \frac{n^2 - 1}{12}$

- Thường người ta hay lấy $a = 1$ và khi đó phân phối đều của X sẽ được kí hiệu là $X \sim U(n)$. Lúc đó hàm phân phối xác suất CDF sẽ là: $F(x) = \frac{x}{n}$.

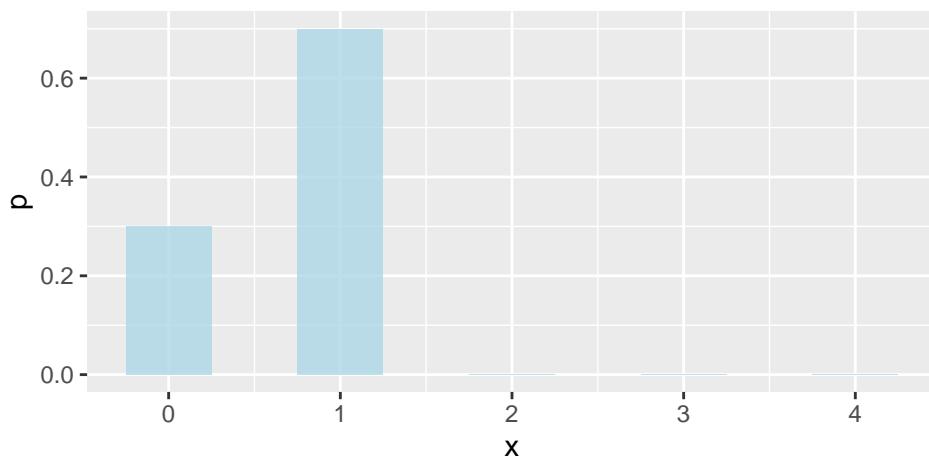
4. Một số phân phối phổ biến



4.1.2. Phân phối Bernoulli (Bernoulli distribution)

Định nghĩa. Biến ngẫu nhiên X tuân theo phân phối Bernoulli $X \sim \text{Bern}(p)$ với tham số $p \in \mathbb{R}, 0 \leq p \leq 1$ là xác suất xuất hiện của sự kiện A tại mỗi phép thử. Phân phối có những đặc điểm như sau:

Hàm/dại lượng	Công thức/giá trị
Tham số	p
PMF	$p(x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$
CDF	$F(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$
Kỳ vọng	$\mathbb{E}[X] = p$
Phương sai	$\text{Var}(X) = p(1-p)$



4. Một số phân phối phổ biến

4.1.3. Phân phối loại (categorical distribution)

Định nghĩa. Phân phối loại là sự tổng quát của phân phối Bernoulli trong đó biến ngẫu nhiên X sẽ có giá trị trong số k giá trị phân biệt $\{v_1, v_2, \dots, v_k\}$. Trong hầu hết các bài toán, tập giá trị $\{v_1, v_2, \dots, v_k\}$ được dùng là $\{1, 2, \dots, k\}$

Hàm/dại lượng	Công thức/giá trị
Tham số	$p_1, \dots, p_k \in \{0, 1\}, \sum p_i = 1$
PMF	$p(x) = \prod_{i=1}^k p_i^{[x=i]}, \quad x \in \{1, 2, \dots, k\}$
CDF	$F(x)$
Kỳ vọng	$\mathbb{E}[X]$
Phương sai	$\text{Var}(X)$

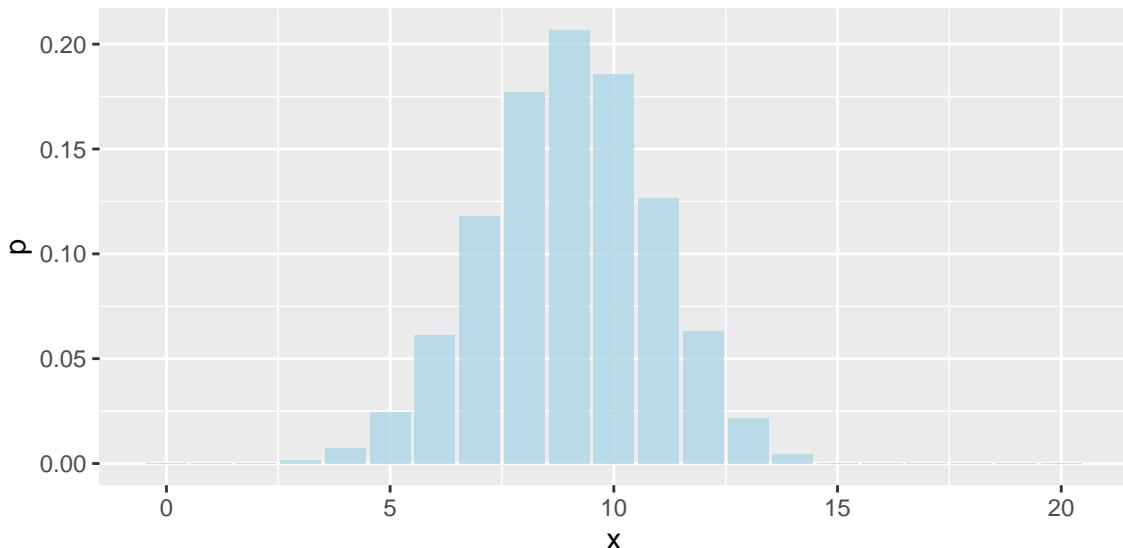
4.1.4. Phân phối nhị thức (Binomial distribution)

Định nghĩa. Biến ngẫu nhiên X tuân theo phân phối nhị thức $X \sim \mathcal{B}(n, p)$ với tham số $n \in \mathbb{N}$ là số lần xuất hiện của sự kiện A với $p \in \mathbb{R}, 0 \leq p \leq 1$ là xác suất xuất hiện của A tại mỗi phép thử. Phân phối có các đặc điểm sau:

Hàm/dại lượng	Công thức/giá trị
Tham số	n, p
PMF	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \dots, n\}$
CDF	$F(x) = \sum_{i=0}^x \binom{n}{i} p^i (1-p)^{n-i}$
Kỳ vọng	$\mathbb{E}[X] = np$
Phương sai	$\text{Var}(X) = np(1-p)$

- Với $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ được gọi là hệ số nhị thức và tên của phân phối này cũng xuất phát từ điểm này.
- Như vậy ta có thể thấy phép thử Bernoulli có thể coi là 1 trường hợp đặc biệt của phân phối nhị thức với $n = 1$, nên phân phối Bernoulli còn có thể kí hiệu là: $X \sim \mathcal{B}(1, p)$.

4. Một số phân phối phổ biến



4.1.5. Phân phối đa thức (multinomial distribution)

Định nghĩa. Phân phối đa thức là phân phối tổng quát hoá của phân phối nhị thức. Giả sử ta có n phép thử độc lập và mỗi phép thử sẽ cho kết quả là một trong số k nhóm với mỗi nhóm có xác suất tương ứng xác định. Khi đó, phân phối đa thức sẽ mô hình hoá phân phối xác suất của số lần thành công của sự kiện. Như vậy, khi $(n = 1, k = 2)$ ta sẽ có phân phối Bernoulli, còn khi $(n > 1, k = 2)$ ta có phân phối nhị thức.

- Giả sử $p_i, i = \{1, \dots, k\}$ là xác suất rơi vào nhóm i tương ứng trong k nhóm, ta có:

$$\sum_{i=1}^k p_i = 1$$

- Nếu biến ngẫu nhiên X_i có giá trị $x_i \in \{0, 1, \dots, n\}, i \in \{1, \dots, k\}$ thể hiện số lần xuất hiện của sự kiện nhóm i , ta có:

$$\sum_{i=1}^k x_i = n$$

- Đặt $\mathbf{X} = [X_1, X_2, \dots, X_k]^\top$ là vector ngẫu nhiên với giá trị $\mathbf{x} = \{x_1, x_2, \dots, x_k\}$ và xác suất tương ứng $\mathbf{p} = [p_1, p_2, \dots, p_k]^\top$. Khi đó, \mathbf{X} tuân theo phân phối $X \sim \mathcal{M}(n, \mathbf{p})$ với tham số $n \in \mathbb{N}$ có các đặc điểm sau:

4. Một số phân phối phô biến

Hàm/dại lượng	Công thức/giá trị
Tham số	p
PMF	$p(\mathbf{x}) = \binom{n}{x_1, x_2, \dots, x_k} \prod_{i=1}^k p_i^{x_i}$
CDF	$F(\mathbf{x})$
Kỳ vọng	$\mathbb{E}[\mathbf{X}] = np$
Phương sai	$\text{Var}(\mathbf{X}) = np \odot (1 - p)$

- Trong đó, $\binom{n}{x_1, x_2, \dots, x_k} = \frac{n!}{x_1!x_2!\dots x_k!}$ là hệ số đa thức và \odot phép nhân từng phần tử.

4.1.6. Phân phối Poisson (Poisson distribution)

Định nghĩa. Phân phối Poisson cho biết trung bình số lần xảy ra thành công của một sự kiện trong một khoảng thời gian nhất định. Giá trị trung bình này được gọi là λ . Phân phối Poisson $X \sim Poi(\lambda)$ sẽ có đặc tính

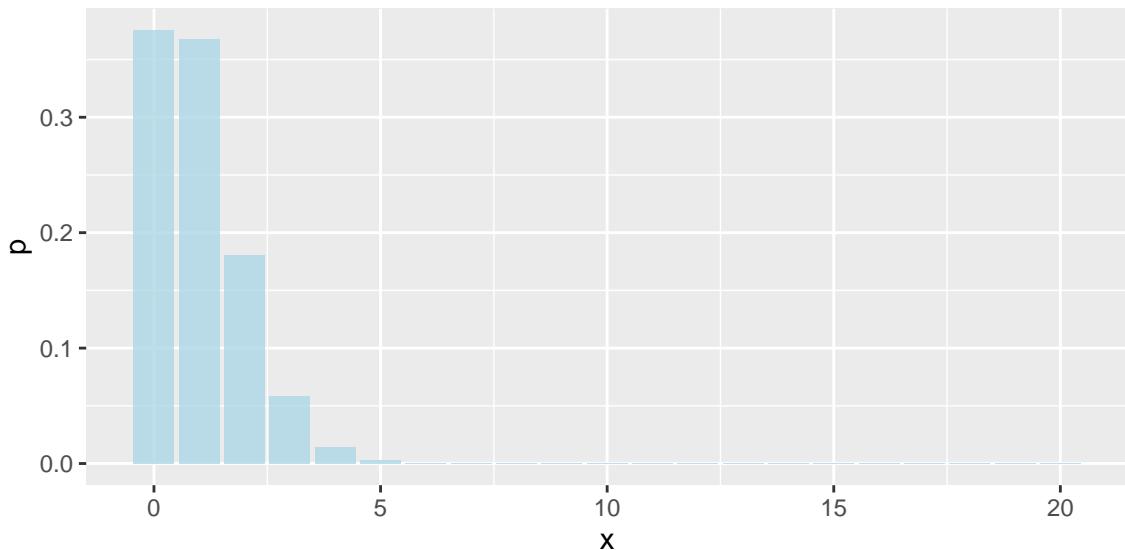
Hàm/dại lượng	Công thức/giá trị
Tham số	$\lambda \geq 0$
PMF	$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{N}$
CDF	$F(x) = e^{-\lambda} \sum_{i=0}^x \frac{\lambda^i}{i!}$
Kỳ vọng	$\mathbb{E}[X] = \lambda$
Phương sai	$\text{Var}(X) = \lambda$

- Phân phối Poisson là phân phối nhị thức đạt được khi n rất lớn và p rất nhỏ. Đặt $\lambda = np$, ta có

$$\begin{aligned}
 p(x) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
 &= \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
 &= \frac{n!}{n^x (n-x)!} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^{n-x}
 \end{aligned}$$

- Khi n rất lớn thì $\left(1 - \frac{\lambda}{n}\right)^x \approx 1$, $\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}$ và $\frac{n!}{n^x (n-x)!} \approx 1$ nên $p(x) \approx \frac{\lambda^x}{x!} e^{-\lambda}$

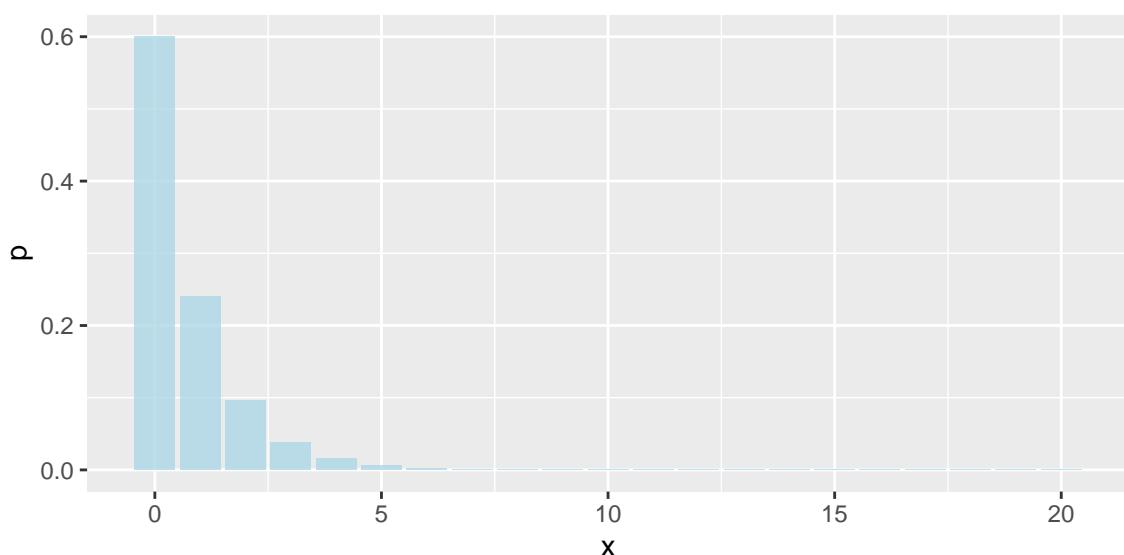
4. Một số phân phối phổ biến



4.1.7. Phân phối hình học (geometric distribution)

Định nghĩa. Phân phối hình học là phân phối của xác suất xuất hiện lần đầu tiên của sự kiện A trong phép thử Bernoulli. Phân phối hình học được kí hiệu là $X \sim Geo(p)$, trong đó tham số p là xác suất xuất hiện của sự kiện A trong mỗi phép thử.

Hàm/dại lượng	Công thức/giá trị
Tham số	p
PMF	$p(x) = p(1-p)^x$
CDF	$F(x) = 1 - (1-p)^{x+1}$
Kỳ vọng	$\mathbb{E}[X] = \frac{1-p}{p}$
Phương sai	$Var(X) = \frac{1-p}{p^2}$

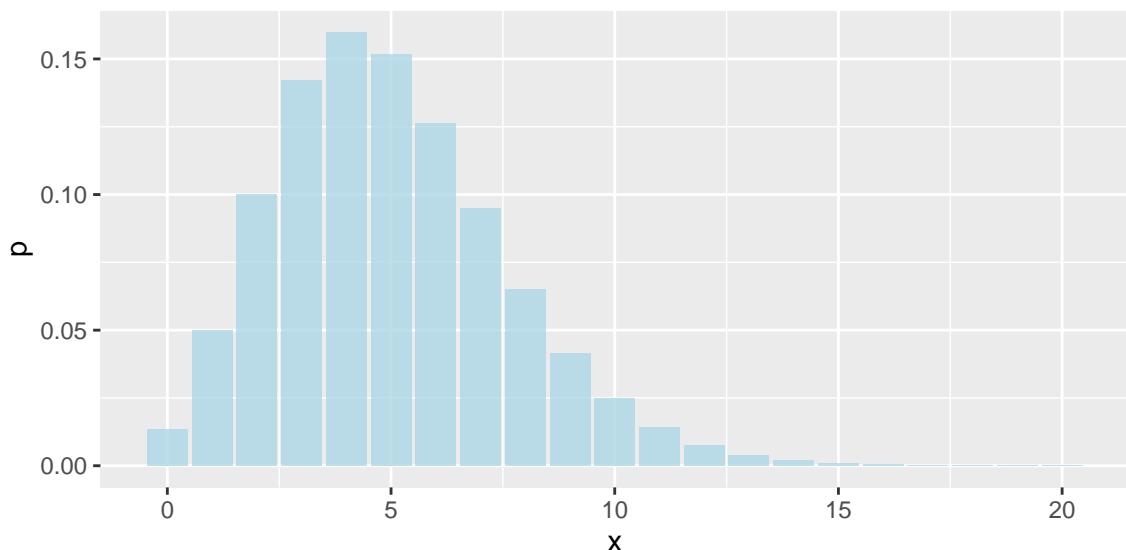


4. Một số phân phối phổ biến

4.1.8. Phân phối nhị thức âm (negative binomial distribution)

Định nghĩa. Là phân phối xác suất xuất hiện lần thứ r của sự kiện A trong phép thử Bernoulli. Như vậy đây là phân phối tổng quát của phân phối hình học và phân phối hình học là phân phối nhị thức âm với $r = 1$. Ta ký hiệu phân phối này là $X \sim \text{NegBin}(r, p)$ với tham số r là số lần xuất hiện của A cùng với p là xác suất xuất hiện của A trong mỗi phép thử.

Hàm/dại lượng	Công thức/giá trị
Tham số	r, p
PMF	$p(x) = \binom{x+r+1}{x} p^r (1-p)^x$
CDF	$F(x) = p^r \sum_{i=0}^x \binom{x+r+1}{x} (1-p)^x$
Kỳ vọng	$\mathbb{E}[X] = \frac{r(1-p)}{p}$
Phương sai	$\text{Var}(X) = \frac{r(1-p)}{p^2}$



4.2

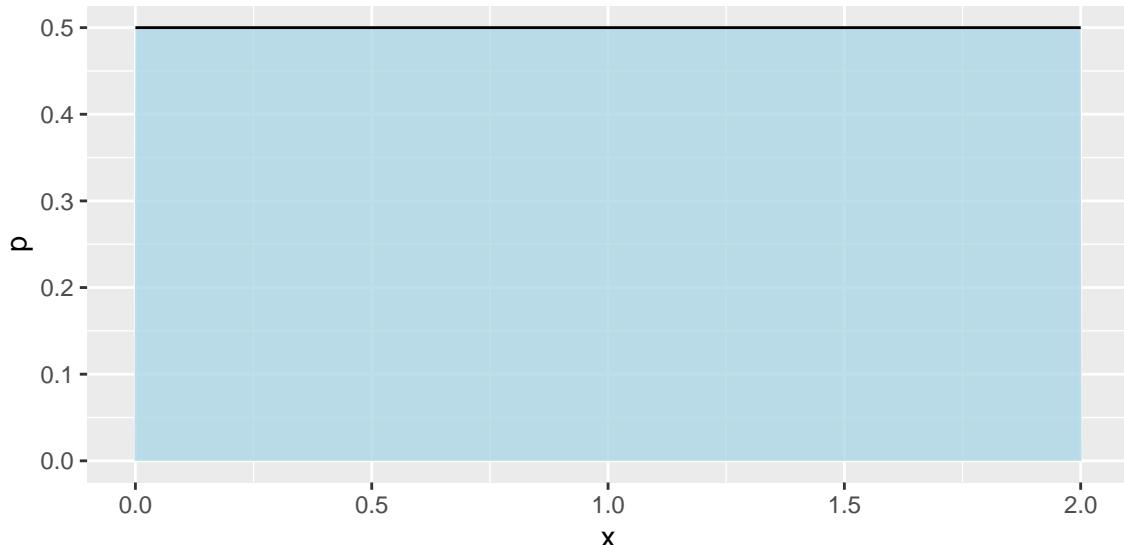
Biến ngẫu nhiên liên tục

4.2.1. Phân phối đều (continuous uniform distribution)

Định nghĩa. Tương tự như đối với trường hợp là biến rời rạc thì với phân phối đều liên tục, bất kì giá trị nào của biến ngẫu nhiên trong miền xác định cũng cho xác suất là như nhau. Biến ngẫu nhiên X tuân theo phân phối đều liên tục $X \sim \text{Uni}(a, b)$ với tham số $a, b \in \mathbb{R}; a < b$, ta sẽ có:

4. Một số phân phối phổ biến

Hàm/dại lượng	Công thức/giá trị
Tham số	$\begin{cases} \frac{1}{b-a}, & \text{nếu } x \in [a, b] \\ 0, & \text{còn lại} \end{cases}$
PDF	$f(x) = \begin{cases} \frac{1}{b-a}, & \text{nếu } x \in [a, b] \\ 0, & \text{còn lại} \end{cases}$
CDF	$F(x) = \begin{cases} 0, & \text{nếu } k < a \\ \frac{k-a}{b-a}, & \text{nếu } k \in [a, b) \\ 1, & \text{nếu } k \geq b \end{cases}$
Kỳ vọng	$\mathbb{E}[X] = \frac{a+b}{2}$
Phương sai	$\text{Var}(X) = \frac{(b-a)^2}{12}$



4.2.2. Phân phối chuẩn (normal distribution)

Phân phối chuẩn hay còn được gọi là phân phối Gauss là một trong những phân phối quan trọng nhất và được ứng dụng rất rộng rãi trong thực tế. Ở đây ta sẽ khảo sát phân phối chuẩn cho 1 biến ngẫu nhiên hay nói cách khác là biến ngẫu nhiên một chiều và cho cả nhiều biến ngẫu nhiên hay vector ngẫu nhiên - biến ngẫu nhiên nhiều chiều.

Đối với biến một chiều (univariate)

Định nghĩa. Biến ngẫu nhiên X tuân theo phân phối chuẩn $X \sim \mathcal{N}(\mu, \sigma^2)$ với tham số kỳ vọng μ và phương sai σ^2 , ta sẽ có:

4. Một số phân phối phổ biến

Hàm/dại lượng	Công thức/giá trị
Tham số	μ, σ^2
PDF	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
CDF	$F(x) = \frac{1}{2} + \Phi\left(\frac{x-\mu}{\sigma}\right)$
Kỳ vọng	$\mathbb{E}[X] = \mu$
Phương sai	$\text{Var}(X) = \sigma^2$

- Trong đó phân phối tích lũy chuẩn tắc $\Phi\left(\frac{x-\mu}{\sigma}\right) = F_X(x)$
- Biểu đồ của hàm mật độ xác suất tuân theo phân phối chuẩn có dạng như sau:
- Để ý rằng phương sai σ^2 càng lớn thì mức độ phân tán xác suất cũng càng rộng, đỉnh thấp hơn và trải rộng hơn. Đường màu đỏ với $\mu = 0$ và $\sigma^2 = 1$ thể hiện phân phối chuẩn tắc $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ (hàm Gauss). Phân phối này thường được dùng để tính các phân phối chuẩn khác qua các phép biến đổi tuyến tính.
- Thông thường các phân phối chuẩn được tính toán theo các phép biến đổi tuyến tính tức là dựa vào các phân phối chuẩn dễ tính và tính được từ trước (như phân phối chuẩn tắc) để ước lượng cho phân phối cần tính. Giờ ta sẽ tìm cách biểu diễn 1 phân phối chuẩn bất kì qua phân phối chuẩn tắc.
- Giả sử $Y = aX + b$ thì Y cũng sẽ là phân phối chuẩn có luật phân phối là: $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

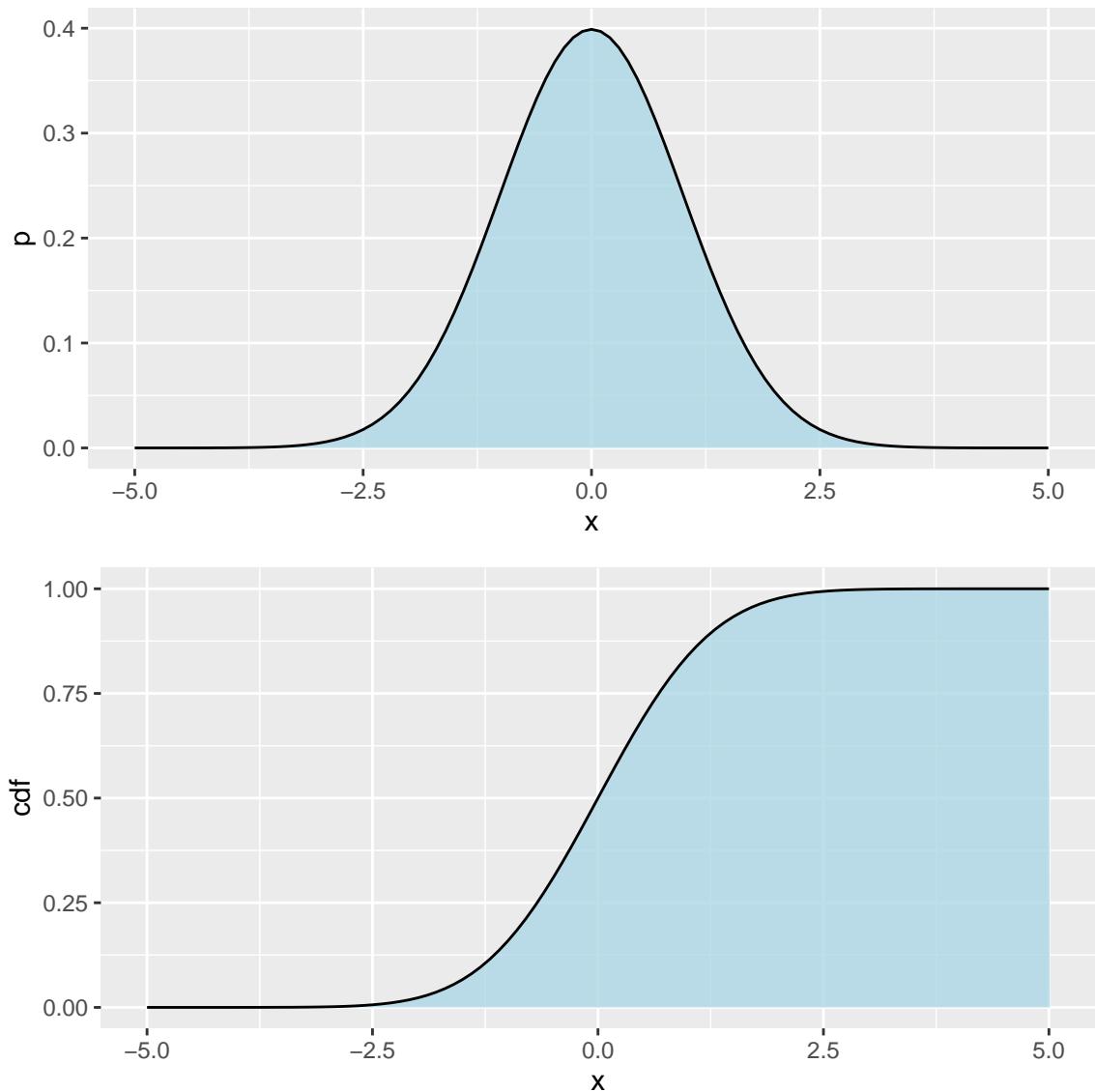
Định nghĩa. Phân phối Z là phân phối chuẩn đặc biệt được định nghĩa như sau

$$Z \sim \mathcal{N}(0, 1) \quad (4.1)$$

- Như vậy Z tuân theo phân phối chuẩn tắc nên ta có thể biến đổi ngược lại để thu được phép biểu diễn phân phối chuẩn qua phân phối của Z .

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

4. Một số phân phối phô biến

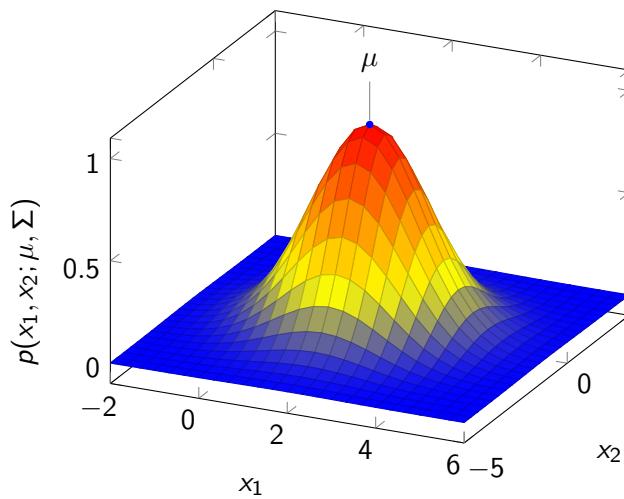


Đối với biến nhiều chiều (multivariate)

Khái niệm 4.1. Đây là tổng quát hóa của phân phối chuẩn đối với biến ngẫu nhiên một chiều và sử dụng cho hợp của nhiều biến ngẫu nhiên - vector ngẫu nhiên. Giả sử vector ngẫu nhiên có số chiều là k : $\mathbf{X} = [X_1, X_2, \dots, X_k]^\top$. Lúc đó phân phối chuẩn của nó sẽ được tham số hóa bởi $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Phân phối này sẽ được kí hiệu là: $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Hàm/dại lượng	Công thức/giá trị
Tham số	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$
PDF	$f(\mathbf{x}) = \frac{1}{\sqrt{ 2\pi\Sigma }} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$
CDF	$F(\mathbf{x})$
vector kỳ vọng	$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = [\mathbb{E}[X_1], \mathbb{E}[X_2], \dots, \mathbb{E}[X_k]]^\top$
Ma trận hiệp phương sai	$\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] = [\text{Cov}(X_i, X_j), 1 \leq i, j \leq k]$

4. Một số phân phối phô biến

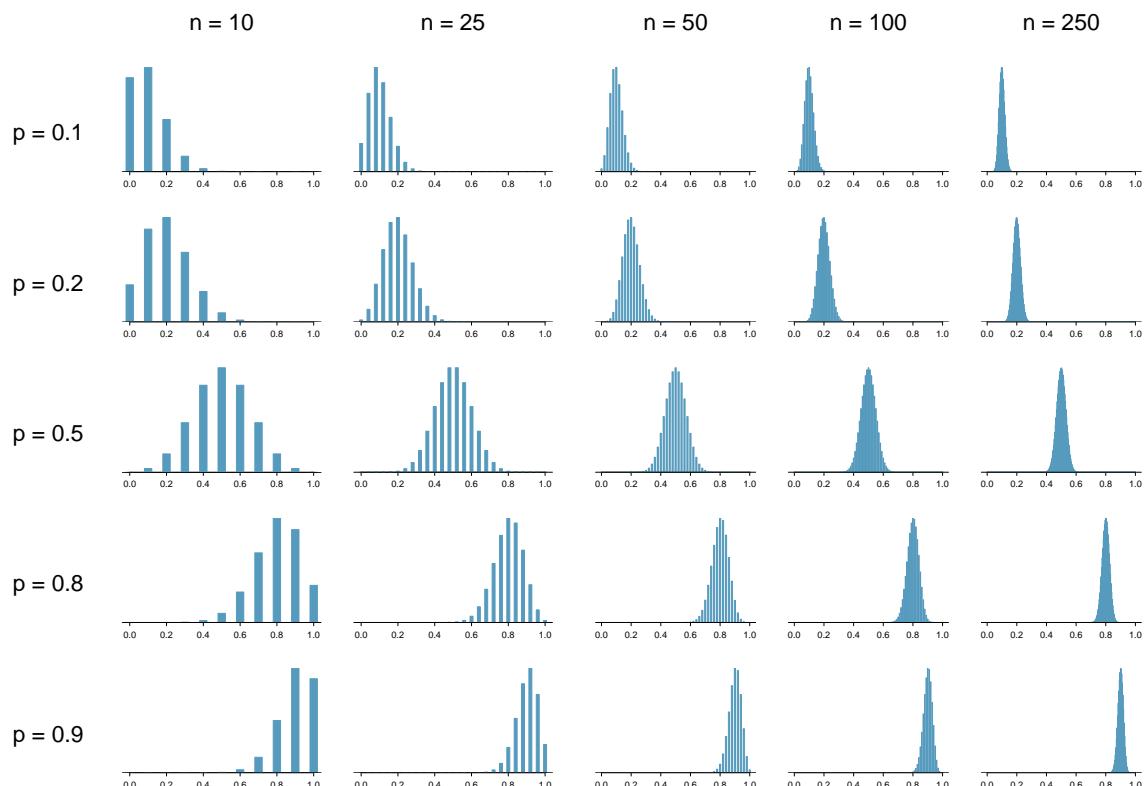


Xấp xỉ bằng phân phối chuẩn

Trong thực tế, đôi khi ta xấp xỉ phân phối nhị thức bằng phân phối chuẩn

- Khi np hoặc $n(1 - p)$ nhỏ, phân bố trông rác.
- Khi np hoặc $n(1 - p) < 10$, phân bố bị lệch
- Khi cả hai np và $n(1 - p)$ lớn, phân bố xấp xỉ phân bố chuẩn
- Khi cả hai np và $n(1 - p)$ rất lớn, phân bố trông trơn và chuẩn hơn

$$\text{Bin}(n, p) \approx \mathcal{N}(\mu = np, \sigma = np(1 - p)) \quad (4.2)$$



4. Một số phân phối phổ biến

Phân phối Poisson xấp xỉ bằng phân phối chuẩn khi λ rất lớn ($\lambda > 10$)

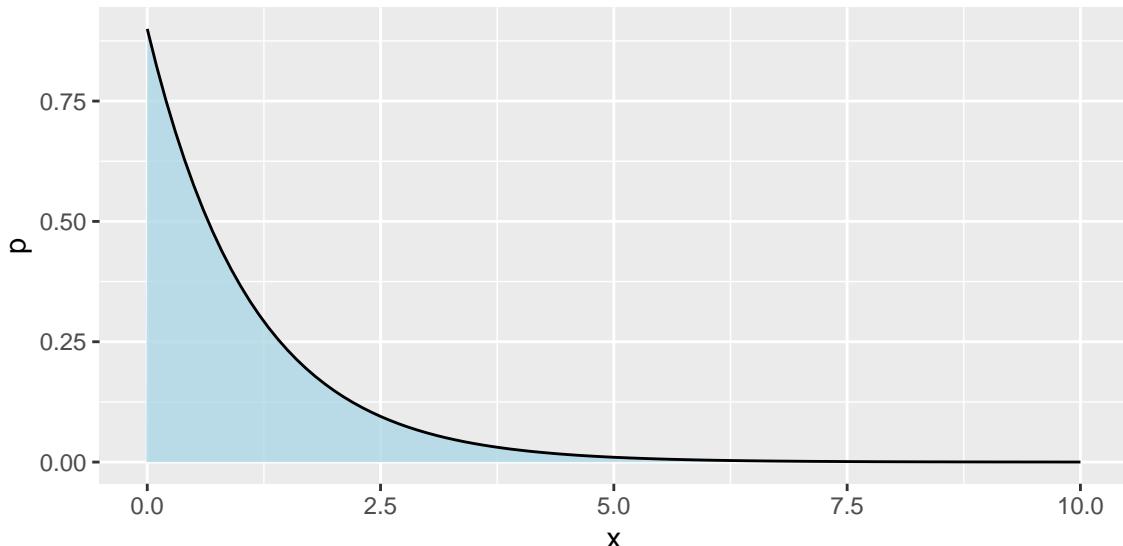
$$\mathcal{P}oi(\lambda) \approx \mathcal{N}(\mu = \lambda, \sigma = \lambda) \quad (4.3)$$

4.2.3. Phân phối mũ (exponential distribution)

Định nghĩa. Là phân phối biểu diễn xác suất thời gian giữa các lần một sự kiện xảy ra. Biến ngẫu nhiên X tuân theo phân phối mũ $X \sim \mathcal{E}xp(\lambda)$ với tham số λ là tần số xảy ra của sự kiện A .

Hàm/dại lượng	Công thức/giá trị
Tham số	λ
PDF	$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$
CDF	$F(x) = 1 - e^{-\lambda x}$
Kỳ vọng	$\mathbb{E}[X] = \frac{1}{\lambda}$
Phương sai	$\text{Var}(X) = \frac{1}{\lambda^2}$

- Nếu đặt $\beta = \frac{1}{\lambda}$ là kỳ vọng ta có thể sử dụng β là tham số của phân phối mũ. Khi đó phân phối này có thể ký hiệu là: $X \sim \mathcal{E}xp(\beta)$ và có $f(x) = \frac{1}{\beta} \exp(-\frac{x}{\beta})$.



4.2.4. Phân phối Student

Định nghĩa. Phân phối Student được ký hiệu là t -student. Được định nghĩa với tham số ν được gọi là bậc tự do (degree of freedom)

4. Một số phân phối phô biến

Hàm/dại lượng	Công thức/giá trị
Tham số	$\nu > 0$
PDF	$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$F(x)$
Kỳ vọng	$\mathbb{E}[X] = 0$
Phương sai	$\text{Var}(X) = \begin{cases} \frac{\nu}{\nu-2} & \nu > 2 \\ \infty & 1 < \nu \leq 2 \end{cases}$

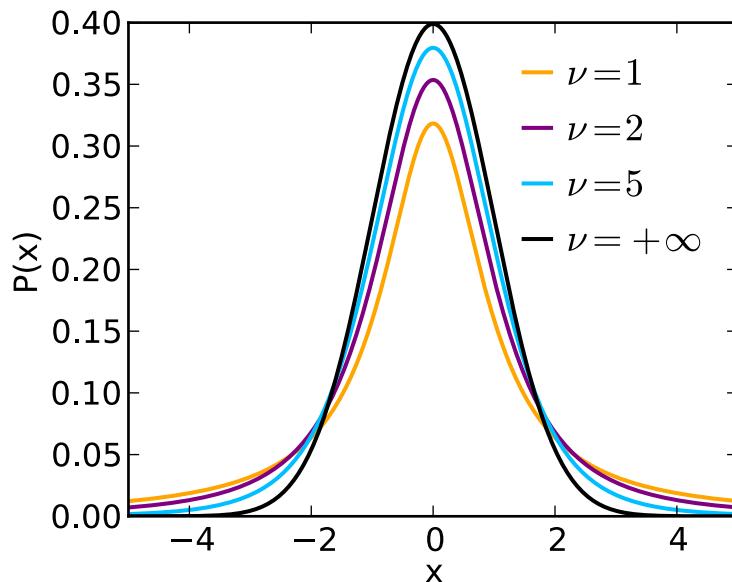
- Trong đó hàm Gamma

$$\Gamma(n) = (n-1)!, \text{ cho } n \in \mathbb{N}^+$$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx, \text{ cho } z > 0$$

- Khi bậc tự do $\nu \geq 30$ thì phân phối Student có thể xấp xỉ bằng phân phối chuẩn

$$t(\nu) \approx \mathcal{N}(0, 1), \text{ với } \nu \geq 30 \quad (4.4)$$

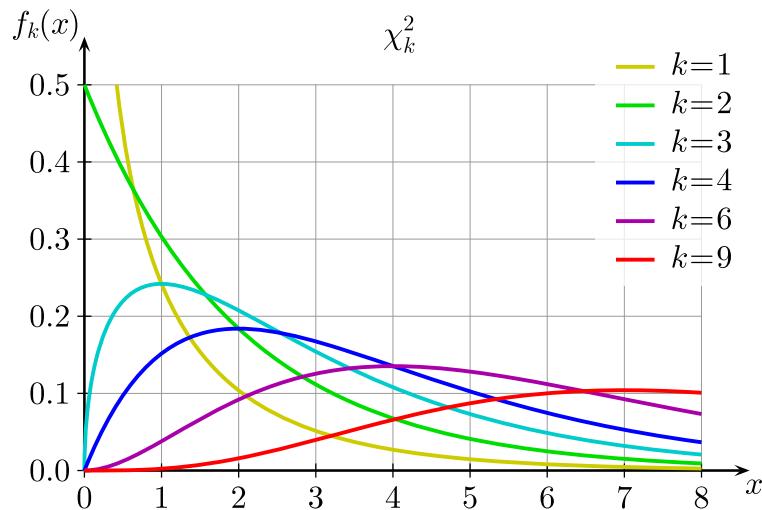


4.2.5. Phân phối Chi squared

Định nghĩa. Phân phối Chi squared được ký hiệu là χ_k^2 . Được định nghĩa với tham số k được gọi là **bậc tự do** (*degree of freedom*)

4. Một số phân phối phô biến

Hàm/dại lượng	Công thức/giá trị
Tham số	$k \in \mathbb{N}^+$
PDF	$f(x) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
CDF	$F(x)$
Kỳ vọng	$\mathbb{E}[X] = k$
Phương sai	$\text{Var}(X) = 2k$



Tính chất.

1. Nếu $X \sim \mathcal{N}(0, 1)$ thì $Y = X^2 \sim \chi_1^2$
2. Nếu $X \sim \chi_r^2$, $Y \sim \chi_s^2$, X và Y độc lập thì $Z = X + Y \sim \chi_{r+s}^2$
3. Nếu X_1, X_2, \dots, X_r độc lập và có cùng phân phối chuẩn $\mathcal{N}(0, 1)$ thì

$$Z = X_1^2 + X_2^2 + \dots + X_r^2 \sim \chi_r^2$$

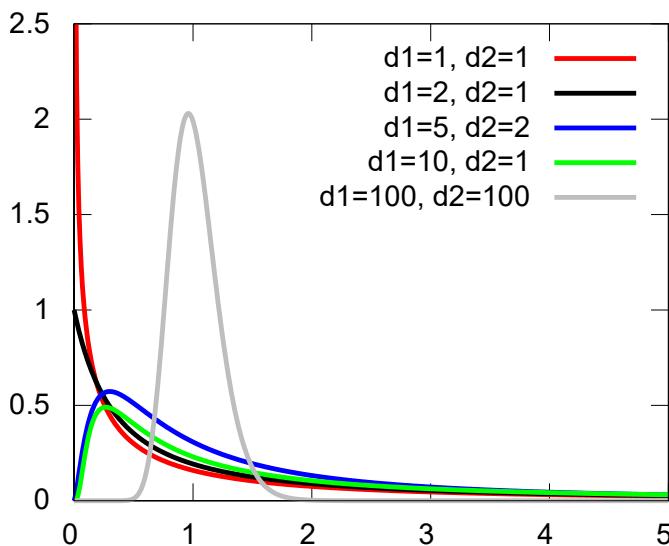
4.2.6. Phân phối Fisher

Định nghĩa. Phân phối Fisher được ký hiệu là F . Được định nghĩa với tham số d_1, d_2 được gọi là bậc tự do (degrees of freedom)

4. Một số phân phối phổ biến

Hàm/dại lượng	Công thức/giá trị
Tham số	$d_1, d_2 > 0$
PDF	$f(x) = \frac{\left(\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}\right)^{0.5}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$
CDF	$F(x)$
Kỳ vọng	$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}$, $d_1 > 2$
Phương sai	$\text{Var}(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)}$, $d_2 > 4$

- Trong đó hàm Beta $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$



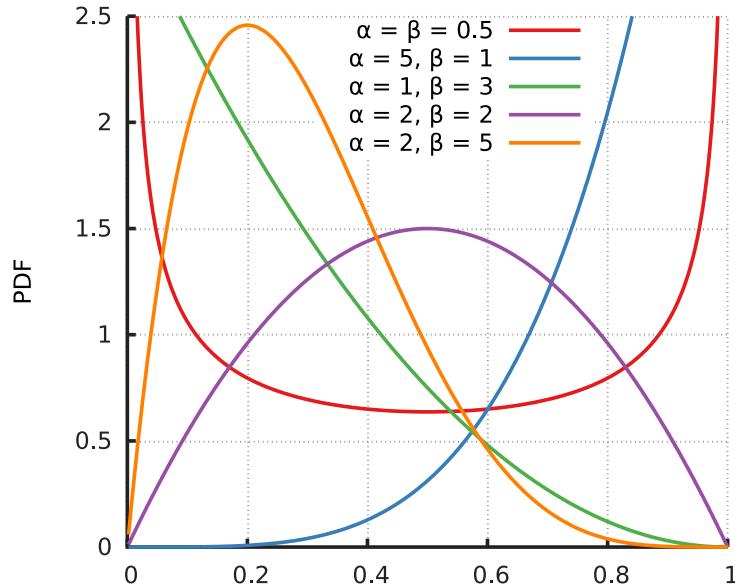
4.2.7. Phân phối Beta

Định nghĩa. Phân phối Beta được ký hiệu là $\mathcal{B}eta$. Được định nghĩa với tham số α, β được gọi là tham số hình dạng (shape parameters)

Hàm/dại lượng	Công thức/giá trị
Tham số	$\alpha, \beta > 0$
PDF	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$
CDF	$F(x)$
Kỳ vọng	$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}$
Phương sai	$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

- Trong đó $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

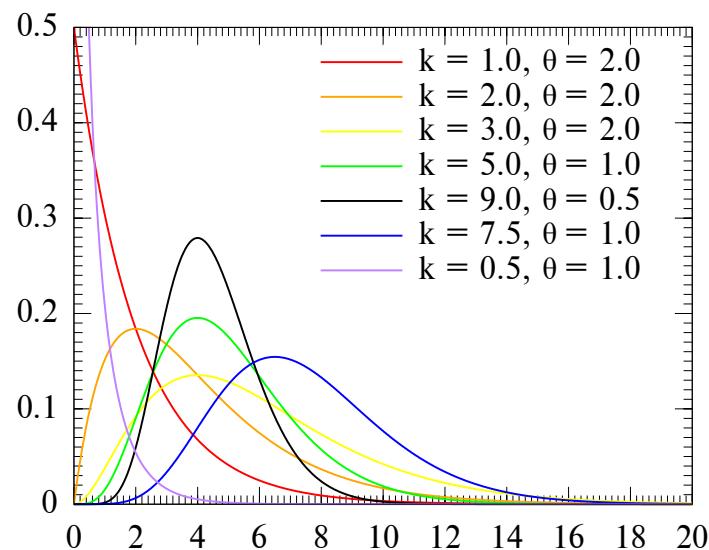
4. Một số phân phối phổ biến



4.2.8. Phân phối Gamma

Định nghĩa. Phân phối Gamma được ký hiệu là $\mathcal{G}amma$. Được định nghĩa với tham số k hình dạng và θ tỉ lệ

Hàm/dại lượng	Công thức/giá trị
Tham số	$k, \theta > 0$
PDF	$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$
CDF	$F(x)$
Kỳ vọng	$\mathbb{E}[X] = k\theta$
Phương sai	$\text{Var}(X) = k\theta^2$

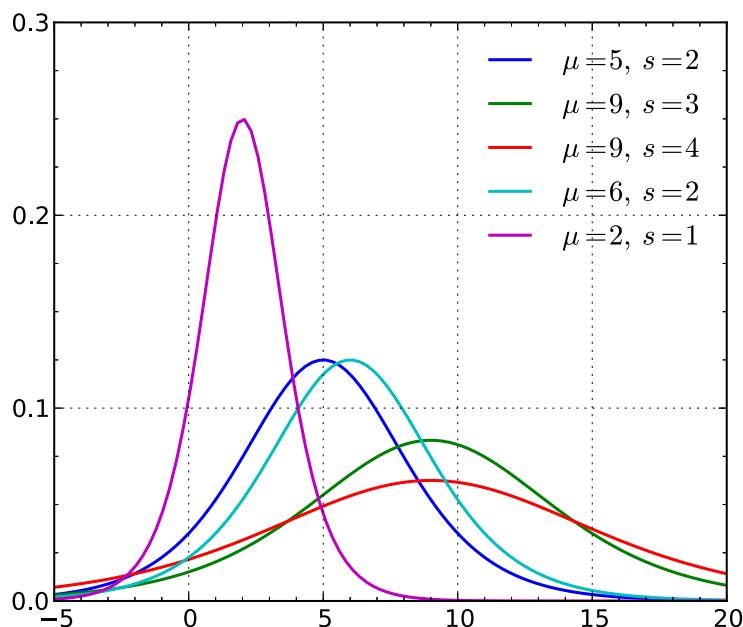


4. Một số phân phối phổ biến

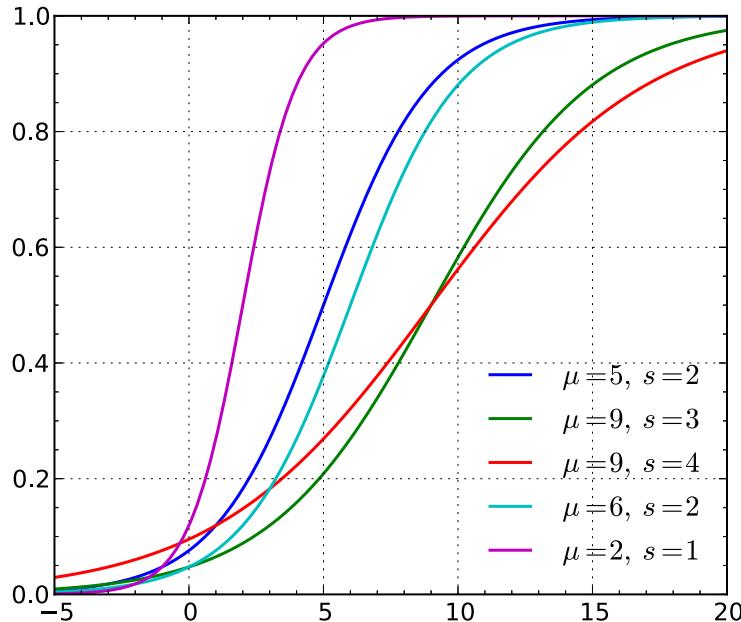
4.2.9. Phân phối Logistic

Định nghĩa. Là phân phối được sử dụng trong hồi quy và mạng neural. Phân phối có hình dạng giống như phân phối chuẩn nhưng mà có phần đuôi đậm hơn

Hàm/dại lượng	Công thức/giá trị
Tham số	$\mu, s > 0$
PDF	$f(x) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}$
CDF	$F(x) = \frac{1}{1+e^{-(x-\mu)/s}}$
Kỳ vọng	$\mathbb{E}[X] = \mu$
Phương sai	$\text{Var}(X) = \frac{\pi^2}{3}s^2$

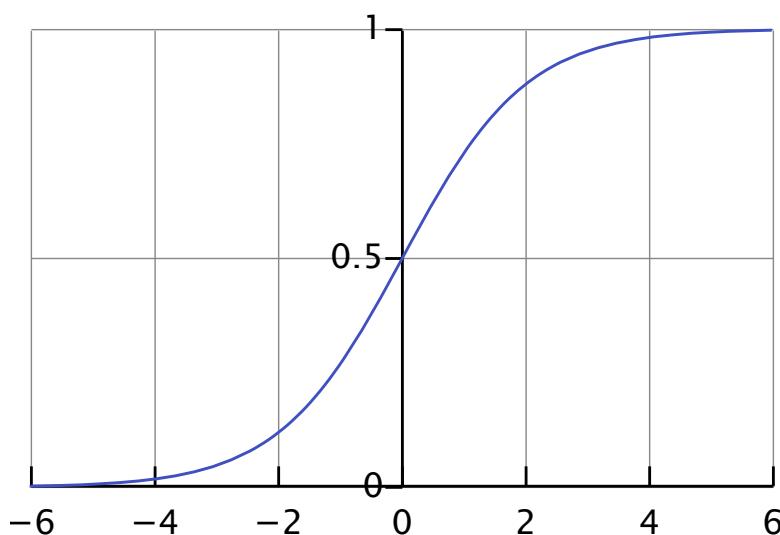


4. Một số phân phối phổ biến



Định nghĩa. Phân phối sigmoid là trường hợp đặc biệt của logistic khi $\mu = 0, s = 1$

Hàm/đại lượng	Công thức/giá trị
Tham số	$\mu, s > 0$
PDF	$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$
CDF	$F(x) = \frac{1}{1 + e^{-x}}$
Kỳ vọng	$\mathbb{E}[X] = 0$
Phương sai	$\text{Var}(X) = \frac{\pi^2}{3}$

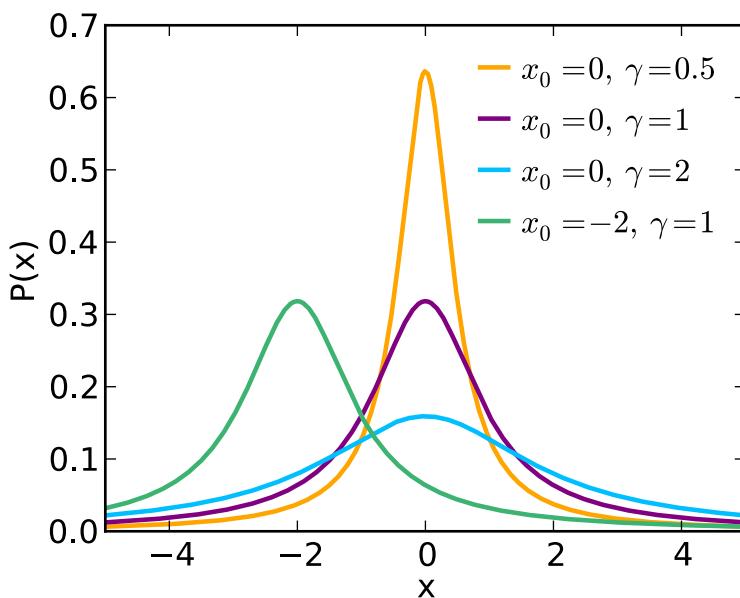


4.2.10. Phân phối Cauchy

Định nghĩa. Phân phối Cauchy là phân phối kỳ vọng và phương sai không xác định

4. Một số phân phối phổ biến

Hàm/đại lượng	Công thức/giá trị
Tham số	$x_0, \gamma > 0$
PDF	$f(x) = \frac{1}{\pi\gamma} \left[1 + \left(\frac{x-x_0}{\gamma} \right)^2 \right]$
CDF	$F(x) = \frac{1}{\pi} \arctan \left(\frac{x-x_0}{\gamma} \right) + \frac{1}{2}$
Kỳ vọng	$\mathbb{E}[X] = x_0$
Phương sai	$\text{Var}(X) = \gamma^2$



Bài tập

Phân phối Bernoulli, nhị thức

B 4.1. Có 8000 sản phẩm trong đó có 2000 sản phẩm không đạt tiêu chuẩn kỹ thuật. Lấy ngẫu nhiên (không hoàn lại) 10 sản phẩm. Tính xác suất để trong 10 sản phẩm lấy ra có 2 sản phẩm không đạt tiêu chuẩn.

B 4.2. Khi tiêm truyền một loại huyết thanh, trung bình có một trường hợp phản ứng trên 1000 trường hợp. Dùng loại huyết thanh này tiêm cho 2000 người. Tính xác suất để

1. có 3 trường hợp phản ứng,
2. có nhiều nhất 3 trường hợp phản ứng,
3. có nhiều hơn 3 trường hợp phản ứng.

B 4.3. Giả sử tỷ lệ sinh con trai và con gái là bằng nhau và bằng $\frac{1}{2}$. Một gia đình có 4 người con. Tính xác suất để 4 đứa con đó gồm

4. Một số phân phối phô biến

- 2 trai và 2 gái.
- 1 trai và 3 gái.
- 4 trai.

B 4.4. Một nhà máy sản xuất với tỷ lệ phế phẩm là 7%.

1. Quan sát ngẫu nhiên 10 sản phẩm. Tính xác suất để
 - a) có đúng một phế phẩm.
 - b) có ít nhất một phế phẩm.
 - c) có nhiều nhất một phế phẩm.
2. Hỏi phải quan sát ít nhất bao nhiêu sản phẩm để xác suất nhận được ít nhất một phế phẩm ≥ 0.9

B 4.5. Tỷ lệ một loại bệnh bẩm sinh trong dân số là $p = 0.01$. Bệnh này cần sự chăm sóc đặc biệt lúc mới sinh. Một nhà bảo sinh thường có 20 ca sinh trong một tuần. Tính xác suất để

1. không có trường hợp nào cần chăm sóc đặc biệt,
2. có đúng một trường hợp cần chăm sóc đặc biệt,
3. có nhiều hơn một trường hợp cần chăm sóc đặc biệt.

Tính bằng quy luật nhị thức rồi dùng quy luật Poisson để so sánh kết quả khi ta xấp xỉ phân phối nhị thức $B(n;p)$ bằng phân phối Poisson $P(np)$.

B 4.6. Tỷ lệ cử tri ủng hộ ứng cử viên A trong một cuộc bầu cử là 60%. Người ta hỏi ý kiến 20 cử tri được chọn một cách ngẫu nhiên. Gọi X là số người bỏ phiếu cho A trong 20 người đó.

1. Tìm giá trị trung bình, độ lệch chuẩn và Mod của X .
2. Tìm $P(X \leq 10)$
3. Tìm $P(X > 12)$
4. Tìm $P(X = 11)$

B 4.7. Giả sử tỷ lệ dân cư mắc bệnh A trong vùng là 10%. Chọn ngẫu nhiên 1 nhóm 400 người.

1. Viết công thức tính xác suất để trong nhóm có nhiều nhất 50 người mắc bệnh A .
2. Tính xấp xỉ xác suất đó bằng phân phối chuẩn.

B 4.8. Một máy sản xuất ra sản phẩm loại A với xác suất 0.485. Tính xác suất sao có trong 200 sản phẩm do máy sản xuất ra có ít nhất 95 sản phẩm loại A .

4. Một số phân phối phô biến

B 4.9. Dựa vào số liệu trong quá khứ, ta ước lượng rằng 85% các sản phẩm của một máy sản xuất nào đó là thứ phẩm. Nếu máy này sản xuất 20 sản phẩm mỗi giờ, thì xác suất 8 hoặc 9 thứ phẩm được sản xuất trong mỗi khoảng thời gian 30 phút là bao nhiêu?

B 4.10. Xác suất trúng số là 1%. Mỗi tuần mua một vé số. Hỏi phải mua vé số liên tiếp trong tối thiểu bao nhiêu tuần để có không ít hơn 95% hy vọng trúng số ít nhất 1 lần.

B 4.11. Trong trò chơi "bầu cua" có ba con xúc sắc, mỗi con có sáu mặt hình là: bầu, cua, huu, nai, tôm và gà. Giả sử có hai người, một người chơi và một người làm cái. Nếu mỗi ván người chơi chỉ đặt ở một ô (một trong các hình: bầu, cua, huu, nai, tôm và gà) sau khi chơi nhiều ván thì người nào sẽ thắng trong trò chơi này. Giả sử thêm mỗi ván người chơi đặt 1000 đ nếu thắng sẽ được 5000 đ, nếu thua sẽ mất 1000 đ. Hỏi trung bình mỗi ván người thắng sẽ thắng bao nhiêu?

B 4.12. Có ba lọ giống nhau: hai lọ loại I, mỗi lọ có 3 bi trắng và 7 bi đen; một lọ loại II có 4 bi trắng và 6 bi đen. Một trò chơi được đặt ra như sau: Mỗi ván, người chơi chọn ngẫu nhiên một lọ và lấy ra hai bi từ lọ đó. Nếu lấy được đúng hai bi trắng thì người chơi thắng, ngược lại người chơi thua.

1. Người A chơi trò chơi này, tính xác suất người A thắng ở mỗi ván.
2. Giả sử người A chơi 10 ván, tính số ván trung bình người chơi thắng được và số ván người A thắng tin chắc nhất.
3. Người A phải chơi ít nhất bao nhiêu ván để xác suất thắng ít nhất một ván không dưới 0,99.

B 4.13. Cho X và Y là hai đại lượng ngẫu nhiên độc lập.

1. Giả sử $X \sim B(1, \frac{1}{5})$, $Y \sim B(2, \frac{1}{5})$. Lập bảng phân phối xác suất của $X + Y$ và kiểm tra rằng $X + Y \sim B(3, \frac{1}{5})$
2. Giả sử $X \sim B(1, \frac{1}{2})$, $Y \sim B(2, \frac{1}{5})$. Tìm phân bố xác suất của $X + Y$. Chứng minh rằng $X + Y$ không có phân bố nhị thức.

B 4.14. Hai cầu thủ ném bóng vào rổ. Cầu thủ thứ nhất ném hai lần với xác suất trúng rổ của mỗi lần là 0.6. Cầu thủ thứ hai ném một lần với xác suất trúng rổ là 0.7. Gọi X là số lần trúng rổ của cả hai cầu thủ. Lập bảng phân phối xác suất của X , biết rằng kết quả của các lần ném rổ là độc lập với nhau.

B 4.15. Bưu điện dùng một máy tự động đọc địa chỉ trên bì thư để phân loại từng khu vực gởi đi, máy có khả năng đọc được 5000 bì thư trong 1 phút. Khả năng đọc sai 1 địa chỉ trên bì thư là 0,04% (xem như việc đọc 5000 bì thư này là 5000 phép thử độc lập).

1. Tính số bì thư trung bình mỗi phút máy đọc sai.

4. Một số phân phối phổ biến

2. Tính số bì thư tin chắc nhất trong mỗi phút máy đọc sai.
3. Tính xác suất để trong một phút máy đọc sai ít nhất 3 bì thư.

B 4.16. Một bài thi trắc nghiệm gồm có 10 câu hỏi, mỗi câu có 4 phương án trả lời, trong đó chỉ có một phương án đúng. Giả sử mỗi câu trả lời đúng được 4 điểm và câu trả lời sai bị trừ 2 điểm. Một sinh viên kém làm bài bằng cách chọn ngẫu nhiên một phương án cho mỗi câu hỏi.

1. Tính xác suất để học sinh này được 4 điểm.
2. Tính xác suất để học sinh này bị điểm âm.
3. Gọi X là số câu trả lời đúng, tính $\mathbb{E}(X)$ và $\text{Var}(X)$.
4. Tính số câu sinh viên này có khả năng trả lời đúng lớn nhất.

B 4.17. Các sản phẩm được sản xuất trong một dây chuyền. Để thực hiện kiểm tra chất lượng, mỗi giờ người ta rút ngẫu nhiên không hoàn lại 10 sản phẩm từ một hộp có 25 sản phẩm. Quá trình sản xuất được báo cáo là đạt yêu cầu nếu có không quá một sản phẩm là thứ phẩm.

1. Nếu tất cả các hộp được kiểm tra đều chứa chính xác hai thứ phẩm, thì xác suất quá trình sản xuất được báo cáo đạt yêu cầu ít nhất 7 lần trong một ngày làm việc 8 giờ là bao nhiêu?
2. Sử dụng phân phối Poisson để xấp xỉ xác suất được tính trong câu (1).
3. Biết rằng lần kiểm tra chất lượng cuối cùng trong câu (1), quá trình sản xuất được báo cáo đạt yêu cầu. Hỏi xác suất mẫu 10 sản phẩm tương ứng không chứa thứ phẩm là bao nhiêu?

Phân phối Poisson

B 4.18. Một trung tâm bưu điện nhận được trung bình 3 cuộc điện thoại trong mỗi phút. Tính xác suất để trung tâm này nhận được 1 cuộc, 2 cuộc, 3 cuộc gọi trong 1 phút, biết rằng số cuộc gọi trong một phút có phân phối Poisson.

B 4.19. Tính $P(X \geq 1 | X \leq 1)$ nếu $X \sim P(5)$

B 4.20. Cho X, Y là các biến ngẫu nhiên độc lập, $X \sim P(\lambda_1)$, $Y \sim P(\lambda_2)$

1. Tính xác suất $P(X + Y = n)$
2. Tính xác suất $P(X = k | X + Y = n)$

B 4.21. Một cửa hàng cho thuê xe ôtô nhận thấy rằng số người đến thuê xe ôtô vào ngày thứ bảy cuối tuần là một đại lượng ngẫu nhiên X có phân phối Poisson với tham số $\lambda = 2$. Giả sử cửa hàng có 4 chiếc ôtô.

4. Một số phân phối phổ biến

1. Tìm xác suất không phải tất cả 4 chiếc ôtô đều được thuê.
2. Tìm xác suất tất cả 4 chiếc ôtô đều được thuê.
3. Tìm xác suất cửa hàng không đáp ứng được yêu cầu.
4. Trung bình có bao nhiêu ôtô được thuê.
5. Cửa hàng cần có ít nhất bao nhiêu ôtô để xác suất không đáp ứng được nhu cầu thuê bé hơn 2%

B 4.22. Một tổng đài bưu điện có các cuộc điện thoại gọi đến xuất hiện ngẫu nhiên, độc lập với nhau và có tốc độ trung bình 2 cuộc gọi trong 1 phút. Tìm xác suất để

1. có đúng 5 cuộc điện thoại trong 2 phút,
2. không có cuộc điện thoại nào trong khoảng thời gian 30 giây,
3. có ít nhất 1 cuộc điện thoại trong khoảng thời gian 10 giây.

B 4.23. Các cuộc gọi điện đến tổng đài tuân theo phân phối Poisson với mức λ trên mỗi phút. Từ kinh nghiệm có được trong quá khứ, ta biết rằng xác suất nhận được chính xác một cuộc gọi trong một phút bằng ba lần xác suất không nhận được cuộc gọi nào trong cùng thời gian.

1. Gọi X là số cuộc gọi nhận được trong mỗi phút. Tính xác suất $P(2 \leq X \leq 4)$.
2. Ta xét 100 khoảng thời gian một phút liên tiếp và gọi U là số khoảng thời gian một phút không nhận được cuộc gọi điện nào. Tính $P(U \leq 1)$.

B 4.24. Tại một điểm bán vé máy bay, trung bình trong 10 phút có 4 người đến mua vé. Tính xác suất để:

1. Trong 10 phút có 7 người đến mua vé.
2. Trong 10 phút có không quá 3 người đến mua vé.

B 4.25. Các khách hàng đến quầy thu ngân, theo phân phối Poisson, với số lượng trung bình 5 người mỗi phút. Tính xác suất xuất hiện ít nhất 10 khách hàng trong khoảng thời gian 3 phút.

B 4.26. Số khách hàng đến quầy thu ngân tuân theo phân phối Poisson với tham số $\lambda = 1$ trong mỗi khoảng 2 phút. Tính xác suất thời gian đợi đến khi khách hàng tiếp theo xuất hiện (từ khách hàng trước đó) nhỏ hơn 10 phút.

B 4.27. Số lượng nho khô trong một cái bánh quy bất kì có phân phối Poisson với tham số λ . Hỏi giá trị λ là bao nhiêu nếu ta muốn xác suất có nhiều nhất hai bánh quy, trong một hộp có 20 bánh, không chứa nho khô là 0.925 ?

4. Một số phân phối phô biến

B 4.28. Một trạm cho thuê xe Taxi có 3 chiếc xe. Hàng ngày trạm phải nộp thuế 8 USD cho 1 chiếc xe (bất kể xe đó có được thuê hay không). Mỗi chiếc được cho thuê với giá 20 USD. Giả sử số xe được yêu cầu cho thuê của trạm trong 1 ngày là đại lượng ngẫu nhiên có phân phối Poisson với $\mu = 2.8$.

1. Tính số tiền trung bình trạm thu được trong một ngày.
2. Giải bài toán trên trong trường hợp trạm có 4 chiếc xe.
3. Theo bạn, trạm nên có 3 hay 4 chiếc xe?

B 4.29. Ta có 10 máy sản xuất (độc lập nhau), mỗi máy sản xuất ra 2% thứ phẩm (không đạt chuẩn).

1. Trung bình có bao nhiêu sản phẩm được sản xuất bởi máy đầu tiên trước khi nó tạo ra thứ phẩm đầu tiên?
2. Ta lấy ngẫu nhiên một sản phẩm từ mỗi máy sản xuất. Hỏi xác suất nhiều nhất hai thứ phẩm trong 10 sản phẩm này là bao nhiêu?
3. Làm lại câu (2) bằng cách sử dụng xấp xỉ Poisson.
4. Phải lấy ra ít nhất bao nhiêu sản phẩm được sản xuất bởi máy đầu tiên để xác suất đạt được ít nhất một thứ phẩm không nhỏ hơn $1/2$ (giả sử rằng các sản phẩm là độc lập với nhau)?

Phân phối chuẩn

B 4.30. Các kết quả của bài kiểm tra chỉ số thông minh (IQ) cho các học sinh của một trường tiểu học cho thấy điểm IQ của các học sinh này tuân theo phân phối chuẩn với các tham số là $\mu = 100$ và $\sigma^2 = 225$. Tỉ lệ học sinh có điểm IQ nhỏ hơn 91 hoặc lớn hơn 130 là bao nhiêu?

B 4.31. Giả sử chiều dài X (đơn vị tính m) của một nơi đỗ xe bất kì tuân theo phân phối chuẩn $N(\mu, 0.01\mu^2)$.

1. Một người đàn ông sở hữu một chiếc xe hơi cao cấp có chiều dài lớn hơn 15% chiều dài trung bình của một chỗ đậu xe. Hỏi tỉ lệ chỗ đậu xe có thể sử dụng là bao nhiêu?
2. Giả sử rằng $\mu = 4$. Hỏi chiều dài của xe là bao nhiêu nếu ta muốn chủ của nó có thể sử dụng 90% chỗ đậu xe?

B 4.32. Đường kính của một chi tiết máy do một máy tiện tự động sản xuất có phân phối chuẩn với trung bình $\mu = 50mm$ và độ lệch chuẩn $\sigma = 0.05mm$. Chi tiết máy được xem là đạt yêu cầu nếu đường kính không sai quá $0.1mm$.

1. Tính tỷ lệ sản phẩm đạt yêu cầu.

4. Một số phân phối phổ biến

2. Lấy ngẫu nhiên 3 sản phẩm. Tính xác suất có ít nhất một sản phẩm đạt yêu cầu.

B 4.33. Trọng lượng X (tính bằng gam) một loại trái cây có phân phối chuẩn $N(\mu, \sigma^2)$, với $\mu = 500(\text{gam})$ và $\sigma^2 = 16(\text{gam}^2)$. Trái cây thu hoạch được phân loại theo trọng lượng như sau:

1. loại 1 : trên 505 gam,
2. loại 2 : từ 495 đến 505 gam,
3. loại 3 : dưới 495 gam.

Tính tỷ lệ mỗi loại.

B 4.34. Một công ty kinh doanh mặt hàng A dự định sẽ áp dụng một trong 2 phương án kinh doanh. Ký hiệu X_1 là lợi nhuận thu được khi áp dụng phương án thứ 1, X_2 là lợi nhuận thu được khi áp dụng phương án thứ 2. X_1, X_2 đều được tính theo đơn vị triệu đồng/ tháng) và $X_1 \sim N(140, 2500)$, $X_2 \sim N(200, 3600)$. Nếu biết rằng, để công ty tồn tại và phát triển thì lợi nhuận thu được từ mặt hàng kinh doanh A phải đạt ít nhất 80 triệu đồng/tháng. Hãy cho biết công ty nên áp dụng phương án nào để kinh doanh mặt hàng A ? Vì sao?

B 4.35. Nghiên cứu chiều cao của những người trưởng thành, người ta nhận thấy rằng chiều cao đó tuân theo quy luật phân bố chuẩn với trung bình là 175cm và độ lệch tiêu chuẩn 4cm. Hãy xác định:

1. tỷ lệ người trưởng thành có tầm vóc trên 180cm.
2. tỷ lệ người trưởng thành có chiều cao từ 166cm đến 177cm.
3. tìm h_0 , nếu biết rằng 33% người trưởng thành có tầm vóc dưới mức h_0 .
4. giới hạn biến động chiều cao của 90% người trưởng thành xung quanh giá trị trung bình của nó.

B 4.36. Ta quan tâm đến tuổi thọ X (theo năm) của một thiết bị. Từ kinh nghiệm trong quá khứ, ta ước lượng xác suất thiết bị loại này còn hoạt động tốt sau 9 năm là 0.1.

1. Ta đưa ra mô hình sau cho hàm mật độ của X

$$f_X(x) = \frac{a}{(x+1)^b} \text{ với } x \geq 0$$

trong đó $a > 0$ và $b > 1$. Tìm hai hằng số a, b .

2. Nếu ta đưa ra một phân phối chuẩn với trung bình $\mu = 7$ cho X , thì giá trị tham số σ là bao nhiêu?
3. Ta xét 10 thiết bị loại này một cách độc lập. Tính xác suất 8 hoặc 9 thiết bị loại này có tuổi đời hoạt động ít hơn 9 năm.

4. Một số phân phối phổ biến

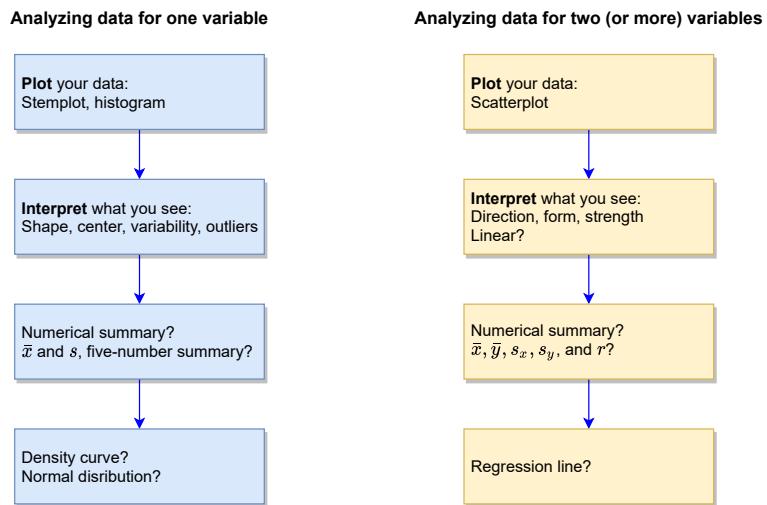
B 4.37. Entropy H của một biến ngẫu nhiên liên tục X được định nghĩa là $H = \mathbb{E}[-\ln f_X(X)]$ với f_X là hàm mật độ xác suất của biến ngẫu nhiên X và \ln là logarit tự nhiên. Tính entropy của biến ngẫu nhiên Gauss với trung bình 0 và phương sai $\sigma^2 = 2$.

Phân II.

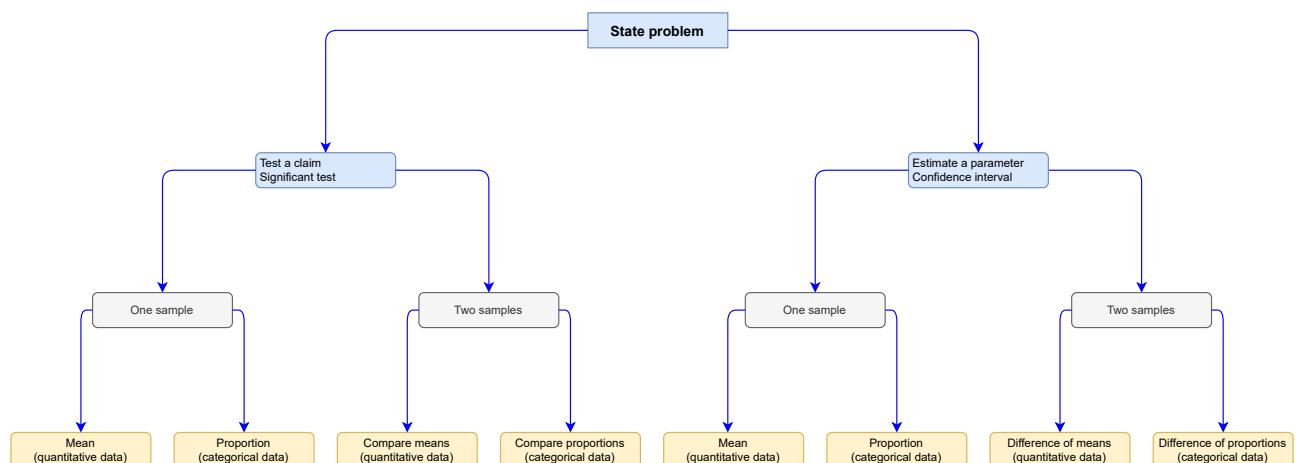
PHÂN TÍCH DỮ LIỆU

Phân tích dữ liệu thống kê là quá trình mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, thông báo kết luận và hỗ trợ ra quyết định.

Tóm tắt phân tích dữ liệu thống kê



Tóm tắt bài toán phân tích dữ liệu thống kê đơn giản



Nghịch lý Simpson

Nghịch lý Simpson là một nghịch lý trong xác suất và thống kê, trong đó một xu hướng xuất hiện trong nhiều nhóm của dữ liệu nhưng biến mất hoặc đảo ngược khi các nhóm đó được gộp lại.

Mẫu thống kê và ước lượng tham số

Thống kê toán là bộ môn toán học nghiên cứu quy luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý số liệu thống kê các kết quả quan sát về những hiện tượng ngẫu nhiên này. Nếu ta thu thập được các số liệu liên quan đến tất cả đối tượng cần nghiên cứu thì ta có thể biết được đối tượng này (phương pháp toàn bộ). Tuy nhiên, trong thực tế điều đó không thể thực hiện được vì quy mô của các đối tượng cần nghiên cứu quá lớn hoặc trong quá trình nghiên cứu đối tượng nghiên cứu bị phá hủy. Vì vậy cần lấy mẫu để nghiên cứu.

5.1

Tổng thể

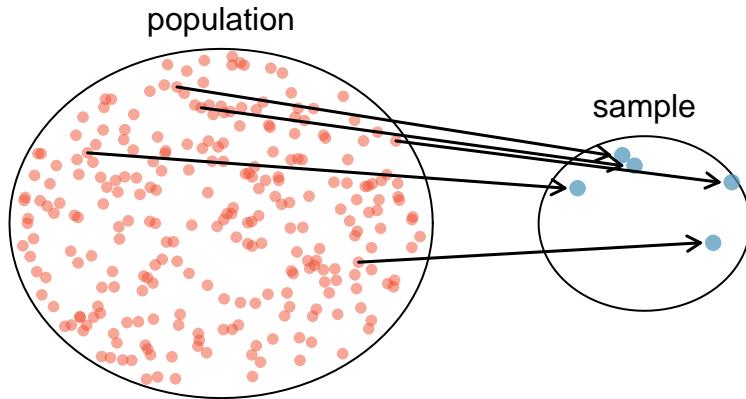
Khi nghiên cứu các vấn đề về kinh tế - xã hội, cũng như nhiều vấn đề thuộc các lĩnh vực vật lý, sinh vật, quân sự ...thường dẫn đến khảo sát một hay nhiều dấu hiệu (định tính hoặc định lượng) thể hiện bằng số lượng trên nhiều phần tử. Tập hợp tất cả các phần tử này gọi là **tổng thể** hay **quần thể** (*population*). Số phần tử trong tổng thể có thể là hữu hạn hoặc vô hạn. Cần nhấn mạnh rằng ta không nghiên cứu trực tiếp bản thân tổng thể mà chỉ nghiên cứu dấu hiệu nào đó của nó.

Ký hiệu N (có thể là ∞) là số phần tử của tổng thể; X là dấu hiệu cần khảo sát.

5.2

Mẫu thống kê

Tập mẫu (*sample*) là tập con của tổng thể và có tính chất tương tự như tổng thể. Số phần tử của tập mẫu được gọi là kích thước mẫu (cỡ mẫu), ký hiệu là n .



5.2.1. Một số cách chọn mẫu cơ bản

Một câu hỏi đặt ra là làm sao chọn được tập mẫu có tính chất tương tự như tổng thể để các kết luận của tập mẫu có thể dùng cho tổng thể? Ta sử dụng một trong những cách chọn mẫu sau:

1. Chọn mẫu ngẫu nhiên có hoàn lại: Lấy ngẫu nhiên một phần tử từ tổng thể và khảo sát nó. Sau đó trả phần tử đó lại tổng thể trước khi lấy một phần tử khác. Tiếp tục như thế n lần ta thu được một mẫu có hoàn lại gồm n phần tử.
2. Chọn mẫu ngẫu nhiên không hoàn lại: Lấy ngẫu nhiên một phần tử từ tổng thể và khảo sát nó rồi để qua một bên, không trả lại tổng thể. Sau đó lấy ngẫu nhiên một phần tử khác, tiếp tục như thế n lần ta thu được một mẫu không hoàn lại gồm n phần tử.
3. Chọn mẫu phân nhóm: Đầu tiên ta chia tập nền thành các nhóm tương đối thuần nhất, từ mỗi nhóm đó chọn ra một mẫu ngẫu nhiên. Tập hợp tất cả mẫu đó cho ta một mẫu phân nhóm. Phương pháp này dùng khi trong tập nền có những sai khác lớn. Hạn chế là phụ thuộc vào việc chia nhóm.
4. Chọn mẫu có suy luận: Dựa trên ý kiến của chuyên gia về đối tượng nghiên cứu để chọn mẫu.

5.2.2. Mẫu ngẫu nhiên

Mẫu ngẫu nhiên là tập các mẫu độc lập và có cùng một quy luật phân phối xác suất (i.i.d - *independently distributed*). Ví dụ ta cần thống kê mức độ xinh gái ảnh hưởng thế nào

5. Mẫu thống kê và ước lượng tham số

tới trí thông minh của chị em. Thì ta có thể coi độ xinh gái là một biến ngẫu nhiên. Lúc này ta lấy mẫu n người và mỗi người sẽ có độ xinh gái là X_i tương ứng. Khi đó ta có thể coi rằng X_i là độc lập đôi một với nhau và chúng có cùng một phân phối xác suất. Tập các mẫu này là mẫu ngẫu nhiên $X = [X_1, X_2, \dots, X_n]$ kích thước n .

Như vậy nếu gọi $p_X(x)$ là hàm trọng lượng xác suất đồng thời nếu X_i là rời rạc và $f_X(x)$ là hàm mật độ xác suất đồng thời nếu X_i là liên tục thì ta sẽ có:

$$p_X(x) = \prod_{i=1}^n p_{X_i}(x_i)$$

và

$$f_X(x) = \prod_{i=1}^n f_{X_i}(x_i)$$

5.2.3. Đại lượng thống kê

Ta đã chọn ra được mẫu ngẫu nhiên rồi và giờ là lúc ta cần xem các đặc trưng của nó hay là đại lượng thống kê. Về mặt định nghĩa hình thức, ta có thể định nghĩa một hàm $\theta = g(X)$ bất kì là một thống kê cho một mẫu ngẫu nhiên X . Ví dụ, $\bar{X} = g(X) = \frac{1}{n} \sum_{i=1}^n X_i$ có thể coi là một đại lượng thống kê.

Ở đây ta sẽ xét một số thống kê cơ bản cho mẫu ngẫu nhiên và gọi chúng là các đặc trưng mẫu. Giả sử ta có mẫu ngẫu nhiên $X = [X_1, X_2, \dots, X_n]$ kích thước n tuân theo một phân phối có kỳ vọng là μ và phương sai là σ^2 .

Trung bình mẫu

Định nghĩa. **Trung bình mẫu** (*mean*) hay còn gọi là **kỳ vọng mẫu** (*expectation*) của một mẫu ngẫu nhiên là giá trị trung bình của mẫu đó:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \tag{5.1}$$

- Rõ ràng, \bar{X} cũng sẽ là một biến ngẫu nhiên và ta có thể tính được các đặc trưng của biến ngẫu nhiên này như

đặc trưng	giá trị
Kỳ vọng	$E[\bar{X}]$
Phương sai	$Var(\bar{X})$

5. Mẫu thống kê và ước lượng tham số

Chứng minh. Ta có,

$$\begin{aligned}\mathbb{E}[\bar{X}] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \\ &= \frac{1}{n} \sum_{i=1}^n \mu \\ &= \mu\end{aligned}$$

$$\begin{aligned}Var[\bar{X}] &= Var\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2} Var\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n Var[X_i] \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned}$$

■

Như vậy là ta có thể thấy rằng giá trị kỳ vọng của biến ngẫu nhiên trung bình luôn là hằng số và bằng kỳ vọng của mẫu ngẫu nhiên. Tức là nếu ta lấy mẫu ngẫu nhiên từ 1 tập mẫu ra thì các tập mẫu ngẫu nhiên này luôn có cùng giá trị trung bình. Nói cách khác trung bình mẫu là không lệch (unbiased).

Phương sai mẫu

Định nghĩa. **Phương sai mẫu** S^2 (hoặc SE^2) là giá trị trung bình của phương sai của mẫu ngẫu nhiên:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (5.2)$$

- S^2 cũng sẽ là một biến ngẫu nhiên

đặc trưng	giá trị
Kỳ vọng	$\mathbb{E}[S^2]$

$\frac{n-1}{n} \sigma^2$

Như vậy là nó không còn bằng với phương sai của X nữa, nên người ta thường lấy một dạng phương sai khác sao cho kỳ vọng của nó là bằng σ^2 .

5. Mẫu thống kê và ước lượng tham số

Định nghĩa. Phương sai mẫu hiệu chỉnh, kí hiệu là s^2 hoặc se^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (5.3)$$

- s^2 cũng sẽ là một biến ngẫu nhiên

đặc trưng	giá trị
Kỳ vọng	$\mathbb{E}[s^2]$
Phương sai	$\text{Var}(s^2) = \frac{2\sigma^4}{n-1}$

Khái niệm 5.1. *Tỉ lệ mẫu*, là tỉ lệ số quan sát có tính chất A (hay dấu hiệu) trong mẫu ngẫu nhiên. Đặt

$$Y_i = \begin{cases} 1 & \text{nếu } X_i \text{ có tính chất } A \\ 0 & \text{nếu } X_i \text{ không có tính chất } A \end{cases} \quad (5.4)$$

Tỉ lệ mẫu được tính như sau

$$F_n = \frac{1}{n} \sum_{i=1}^n Y_i \quad (5.5)$$

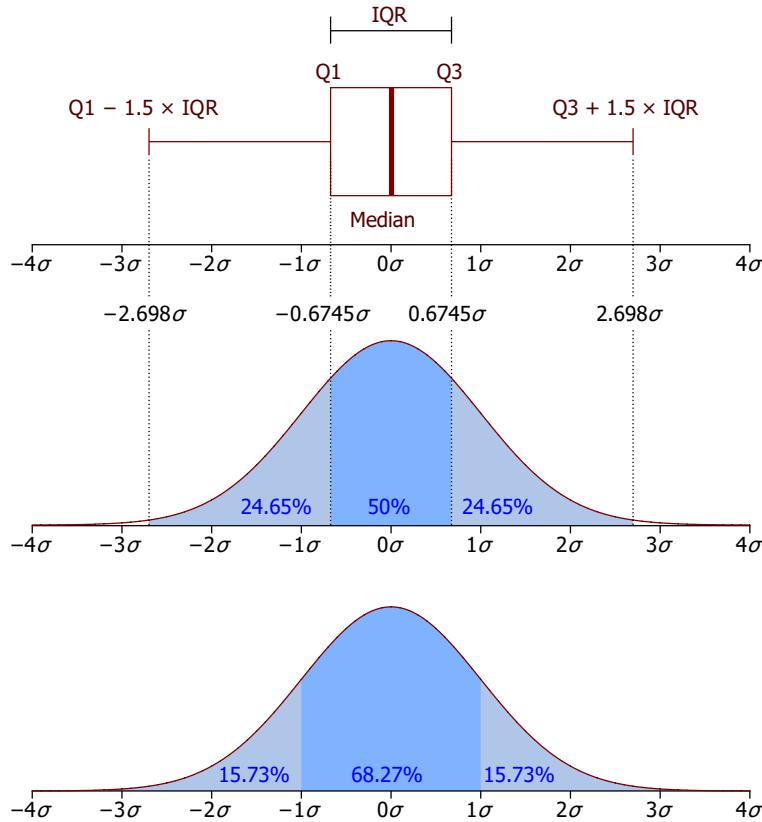
Các đại lượng thống kê khác

1. Hệ số biến thiên (*coefficient of variation*)

$$CV = \frac{s}{\bar{x}} \quad (5.6)$$

2. Tứ phân vị (*quartile*): Q_1, Q_2, Q_3 với $Q_2 = \text{median}$

5. Mẫu thống kê và ước lượng tham số



3. Khoảng tứ phân vị (*interquartile range*)

$$IQR = Q_3 - Q_1 \quad (5.7)$$

4. Giá trị chuẩn hóa (z-score)

$$z_i = \frac{x_i - \bar{x}}{s} \quad (5.8)$$

5. Hệ số bất đối xứng (*skewness*)

$$skew = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3} \quad (5.9)$$

- Nếu $|skew| \leq 0.5$ phân bố được xem đối xứng
- Nếu $|skew| > 0.5$ phân bố bị xem là bị lệch
 - Nếu $skew > 0.5$ thì phân bố lệch phải
 - Nếu $skew < -0.5$ thì phân bố lệch trái

6. Hệ số nhọn (*kurtosis*)

$$kurt = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^3} \quad (5.10)$$

excess kurtosis

$$kurt = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^3} - 3 \quad (5.11)$$

5. Mẫu thống kê và ước lượng tham số

- Phân bố chuẩn có $kurt = 0.0$
- Nếu $kurt > 0$ phân bố có phần đuôi đậm
- Nếu $kurt < 0$ phân bố có phần đuôi mỏng

7. Hiệp phương sai (*covariance*) cho hai biến x, y liên tục

$$cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (5.12)$$

8. Hệ số tương quan (*correlation coefficient - Pearson*) cho hai biến x, y liên tục

$$r(X, Y) = \frac{cov(X, Y)}{s_X s_Y} \quad (5.13)$$

9. Bảng tương quan (*contingency table*) cho hai biến x, y định tính; ví dụ

		decision		Total
gender		promoted	not promoted	
	male	21	3	24
	female	14	10	24
Total		35	13	48

5.3

Phân phối của đại lượng thống kê của mẫu thống kê

Xét một tập mẫu $\mathcal{D} = \{X_1, X_2, \dots, X_n\}$ được lấy mẫu i.i.d từ một toàn bộ

5.3.1. Phân phối của trung bình mẫu

- Thống kê trung bình mẫu

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (5.14)$$

có phân phối chuẩn $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (theo định lý giới hạn trung tâm)

- Thống kê

$$U = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \quad (5.15)$$

có phân phối chuẩn $\mathcal{N}(0, 1)$

- Thống kê

$$T = \frac{\bar{X} - \mu}{s} \sqrt{n} \quad \text{hoặc} \quad T = \frac{\bar{X} - \mu}{s} \sqrt{n-1} \quad (5.16)$$

có phân phối Student với $n - 1$ bậc tự do

5. Mẫu thống kê và ước lượng tham số

5.3.2. Phân phối của phương sai mẫu

- Thống kê

$$q = \frac{ns^2}{\sigma^2} \quad (5.17)$$

có phân phối Chi-squared với $n - 1$ bậc tự do

5.4

Ước lượng điểm cho tham số

5.4.1. Tham số

- Tham số (*parameter*) là các giá trị mô tả phân phối xác suất của biến ngẫu nhiên X trong mô hình thống kê tương ứng. Các tham số này được kí hiệu là θ và nó có thể một số hoặc một vector.

Mô hình	Ký hiệu	Tham số
Phân phối đều	$X \sim Uni(a, b)$	$\theta = [a, b]$
Phân phối Bernoulli	$X \sim Bern(p)$	$\theta = p$
Phân phối nhị thức	$X \sim Bin(n, p)$	$\theta = [n, p]$
Phân phối Poisson	$X \sim Poi(\lambda)$	$\theta = \lambda$
Phân phối hình học	$X \sim Geo(p)$	$\theta = p$
Phân phối nhị thức âm	$X \sim NegBin(r, p)$	$\theta = [r, p]$
Phân phối chuẩn	$X \sim \mathcal{N}(\mu, \sigma^2)$	$\theta = [\mu, \sigma^2]$
Phân phối mũ	$X \sim Exp(\beta)$	$\theta = \beta$

- Tham số cũng có thể là các đặc trưng (đại lượng thống kê) của biến ngẫu nhiên X như kỳ vọng $\mathbb{E}(X)$ hay phương sai $Var(X)$

Cho biến ngẫu nhiên X có thể đã biết hoặc chưa biết quy luật phân phối xác suất dạng tổng quát, nhưng chưa biết tham số θ nào đó. Hãy ước lượng θ bằng tập mẫu $X = \{X_1, \dots, X_n\}$. Vì kết quả ước lượng cho θ là một số nên ước lượng như vậy gọi là *ước lượng điểm*.

Khái niệm 5.2. Cho một mẫu ngẫu nhiên i.i.d \mathcal{D} có $\{X_1, \dots, X_n\}$, một hàm ước lượng (estimator) của tham số θ là một hàm số n biến

$$\hat{\theta} = f(X_1, \dots, X_n) \quad (5.18)$$

Như vậy, ước lượng tham số (*parameter estimate*) là quá trình đi tìm tham số để mô tả biến ngẫu nhiên hay quan hệ của các biến ngẫu nhiên theo các tiêu chuẩn: không chêch, vững, hiệu quả. Có thể có nhiều cách khác nhau để ước lượng θ , mỗi cách được xác định bởi một hàm ước lượng.

5. Mẫu thống kê và ước lượng tham số

5.4.2. Các tiêu chuẩn lựa chọn phương pháp ước lượng

Cùng một mẫu ngẫu nhiên có thể xây dựng nhiều phương pháp khác nhau để ước lượng cho tham số θ . Vì vậy ta cần lựa chọn phương pháp tốt nhất để ước lượng cho tham số θ dựa vào các tiêu chuẩn sau.

1. Ước lượng không chêch (*unbiased estimator*) nếu như

$$\theta = \mathbb{E}(\hat{\theta}) \quad (5.19)$$

2. Ước lượng vững (*consistent estimator*) nếu như

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| < \epsilon) = 1 \quad (5.20)$$

3. Ước lượng hiệu quả (*efficient estimator*) nếu như phương sai của $\hat{\theta}$ nhỏ hơn bất kỳ phương sai của hàm ước lượng không chêch nào khác

Lưu ý

- Trung bình mẫu \bar{X} là một ước lượng không chêch (vững) của μ
- Phương sai mẫu hiệu chỉnh s là một ước lượng không chêch (vững) của σ
- Tỉ lệ mẫu là một ước lượng không chêch (vững) cho xác suất biến có một tính chất A

Có nhiều phương pháp khác nhau để ước lượng các tham số, nhưng được sử dụng nhiều và phổ biến nhất là 3 phương pháp:

- ML (Maximum Likelihood): ước lượng cực đại khả năng
- MAP (Maximum A Posteriori): ước lượng cực đại xác suất hậu nghiệm
- EM (Expectation Maximization): ước lượng cực đại kỳ vọng

5.4.3. ML

Giả sử chúng ta có cần ước lượng mô hình thống kê được biểu diễn bằng tham số θ . Gọi hàm khả năng - hợp lý (*likelihood function*) của mô hình

$$f(X | \theta) \quad (5.21)$$

Giả sử tập mẫu \mathcal{D} có $\{X_1, \dots, X_n\}$ là i.i.d, giá trị likelihood của \mathcal{D} ứng với tham số θ là

$$L(\theta) = f(\mathcal{D} | \theta) = \prod_{i=1}^n f(X_i | \theta) \quad (5.22)$$

5. Mẫu thống kê và ước lượng tham số

Bài toán bây giờ trở thành là làm sao có thể tìm được tham số θ sao cho giá trị likelihood là lớn nhất có thể

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta) \quad (5.23)$$

Với argmax là hàm trả ra giá trị của tham số θ mà tại đó khiến hàm đạt được giá trị lớn nhất. Tuy nhiên, do các $f(X_i | \theta)$ là nhỏ (nhỏ hơn 1) nên với tập mẫu lớn $L(\theta)$ rất có thể sẽ rất nhỏ và dẫn đến tính toán số không ổn định. Trong thực tế, người ta sử dụng log của hàm hợp lý (*log-likelihood function*) vì hàm log là hàm tăng đơn điệu do đó $L(\theta)$ và $\log L(\theta)$ sẽ cùng các điểm cực trị như nhau. Đặt

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^n f(X_i | \theta) = \sum_{i=1}^n \log f(X_i | \theta) \quad (5.24)$$

Và ta sẽ tìm θ để tối ưu hóa hàm này

$$\hat{\theta} = \operatorname{argmax}_{\theta} LL(\theta) \quad (5.25)$$

Để tối ưu hóa hàm này ta có thể sử dụng nhiều phương pháp khác nhau, một trong các phương pháp phổ biến là sử dụng phương trình đạo hàm bậc nhất và giải phương trình này. Trong hầu hết các bài toán thực tế việc giải phương trình này cần phải dựa trên thuật toán gradient.

Tính chất

1. **Tính nhất quán:** Khi tập mẫu \mathcal{D} có kích thước lớn thì $\hat{\theta}$ tiệm cận về θ
2. **Tính “plug-in”:**
 - a) Nếu $\hat{\theta}$ là ước lượng likelihood cực đại của θ thì $g(\hat{\theta})$ là ước lượng likelihood cực đại của $g(\theta)$
 - b) Nếu se là ước lượng sai số chuẩn của ước lượng $\hat{\theta}$ thì $g'(\hat{\theta}) \times se$ là ước lượng sai số chuẩn của ước lượng $g(\hat{\theta})$ (phương pháp Delta)

Ví dụ

Giả sử ta có tập dữ liệu mẫu $X = \{X_1, X_2, \dots, X_n\}$ tuân theo luật phân phối Bernoulli $X \sim \text{Bern}(p)$ với tham số p . Giờ ta sẽ sử dụng phương pháp MLE để tìm tham số p .

Hàm hợp lý của ta lúc này sẽ có dạng:

$$L(\theta) = \prod_{i=1}^n f(X_i | \theta) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i}$$

5. Mẫu thống kê và ước lượng tham số

Phiên bản log:

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n \log(p^{X_i}(1-p)^{1-X_i}) \\ &= \sum_{i=1}^n \log(p^{X_i}) + \log((1-p)^{1-X_i}) \\ &= \sum_{i=1}^n X_i \log(p) + (1-X_i) \log(1-p) \end{aligned}$$

Đặt $Y = \sum_{i=1}^n X_i$, ta có:

$$LL(\theta) = Y \log(p) + (n - Y) \log(1 - p)$$

Giờ ta cần chọn \hat{p} sao cho hàm trên đạt giá trị lớn nhất:

$$\hat{p} = \operatorname{argmax}_p \left(Y \log(p) + (n - Y) \log(1 - p) \right)$$

Như ta đã biết hàm này đạt cực trị tại điểm có đạo hàm bằng 0, tức là:

$$\begin{aligned} LL(p)' &= 0 \\ \iff Y \frac{1}{p} + (n - Y) \frac{-1}{1-p} &= 0 \\ \iff p &= \frac{Y}{n} \end{aligned}$$

Từ đây ta sẽ có giá trị

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n} \tag{5.26}$$

là giá trị ước lượng cần tìm.

5.4.4. MAP

Khái niệm

Ý tưởng của MAP là chọn tham số θ sao chúng gần với tham số thực của dữ liệu nhất có thể bằng cách kết hợp MLE với xác suất tiên nghiệm về θ . Gọi $P(\theta)$ là xác suất hậu nghiệm

$$P(\theta) = f(\theta | X) \tag{5.27}$$

với X ở đây là vector $X = [X_1, \dots, X_n]^\top$

5. Mẫu thống kê và ước lượng tham số

Sử dụng công thức Bayes để biến đổi công thức trên một chút

$$\begin{aligned} P(\theta) &= f(\theta | X) \\ &= \frac{f(X | \theta)g(\theta)}{h(X)} \\ &= \frac{g(\theta) \prod_{i=1}^n f(X_i | \theta)}{h(X)} \end{aligned}$$

với g, h lần lượt là các xác suất tiên nghiệm và xác suất mẫu quan sát.

Bài toán là tìm được tham số θ sao cho xác suất trên là lớn nhất

$$\begin{aligned} \hat{\theta} &= \operatorname{argmax}_{\theta} P(\theta) \\ &= \operatorname{argmax}_{\theta} \frac{g(\theta) \prod_{i=1}^n f(X_i | \theta)}{h(X)} \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^n f(X_i | \theta)g(\theta) \end{aligned}$$

Ở trên ta giản lược được $h(X)$ là do xác suất của mẫu quan sát không liên quan đến tham số θ . Tiếp tục, ta lại lấy log tương tự như MLE

$$\begin{aligned} \hat{\theta} &= \operatorname{argmax}_{\theta} \prod_{i=1}^n f(X_i | \theta)g(\theta) \\ &= \operatorname{argmax}_{\theta} \log \left(\prod_{i=1}^n f(X_i | \theta)g(\theta) \right) \\ &= \operatorname{argmax}_{\theta} \left(\log g(\theta) + \sum_{i=1}^n \log f(X_i | \theta) \right) \end{aligned}$$

Từ công thức trên ta thấy rằng MAP chỉ khác MLE ở chỗ thêm $\log g(\theta)$ hay còn gọi là xác suất tiên nghiệm vào hàm mục tiêu. Nói vậy thôi nhưng vấn đề lại phát sinh rồi! Ta hiện giờ đã biết được mô hình phân phối của tập mẫu $f(X)$ nhưng lại chưa biết mô hình phân phối của tham số $g(\theta)$ nên việc tính toán có vẻ bất khả thi.

Siêu tham số

Tiếp tục với vấn đề chọn mô hình phân phối nào cho các tham số θ . Trong thực tế người ta chọn mô hình của tham số θ sao cho cùng dạng với mô hình của tham số θ có điều kiện X . Hiểu theo nghĩa của Bayes là mô hình thống kê xác suất tiền nghiệm và hậu nghiệm là cùng họ với nhau. Mô hình thống kê kiểu này được gọi là xác suất tiền nghiệm liên hợp (conjugate prior) của hàm khả năng (likelihood function).

Ví dụ, nếu mẫu của ta tuân theo phân phối Bernoulli $X \sim \text{Bern}(p)$ với tham số $\theta = p$ thì phân phối xác suất tiền nghiệm liên hợp của nó là phân phối Beta $\theta \sim \text{Beta}(\alpha, \beta)$ bởi khi kết

5. Mẫu thống kê và ước lượng tham số

hợp với $\mathcal{B}eta$ thì xác suất hậu nghiệm của nó cũng sẽ là $\mathcal{B}eta$. Ở đây tôi không đề cập sâu về dạng mô hình này nhưng bạn nên đọc thêm để hiểu về nó cũng như các ứng dụng của nó.

Do các phân phối của tham số cũng được quy định bởi các tham số của phân phối tương ứng. Tức là các tham số θ cần tìm của ta giờ phải phụ thuộc vào cả các tham số của phân phối xác suất của nó. Để phân biệt người ta gọi các tham số này là **siêu tham số** (*hyperparameters*).

Khi làm việc các siêu tham số này được thiết lập dựa vào cảm quan của người giải quyết bài toán. Việc chọn được siêu tham số hợp lý là một việc vô cùng cần thiết để thu được tham số θ tốt. Chính vì vậy mà các bài toán học máy sau này, ta thường xuyên phải xem xét nhiều bộ siêu tham số để được kết quả mong đợi là thế.

Để thuận tiện khi làm việc, ta liệt kê một số xác suất tiền nghiệm liên hợp như dưới đây:

Mô hình phân phối	Ký hiệu	Tham số	Phân phối liên hợp cho tham số
Phân phối Bernoulli	$X \sim \mathcal{B}ern(p)$	$\theta = p$	$\theta \sim \mathcal{B}eta(\alpha, \beta)$
Phân phối nhị thức	$X \sim \mathcal{B}in(n, p)$	$\theta = p$	$\theta \sim \mathcal{B}eta(\alpha, \beta)$
Phân phối Poisson	$X \sim \mathcal{P}oi(\lambda)$	$\theta = \lambda$	$\theta \sim \mathcal{G}amma(\alpha, \beta)$
Phân phối hình học	$X \sim \mathcal{G}eo(p)$	$\theta = p$	$\theta \sim \mathcal{B}eta(\alpha, \beta)$
Phân phối nhị thức âm	$X \sim \mathcal{N}eg\mathcal{B}in(r, p)$	$\theta = p$	$\theta \sim \mathcal{B}eta(\alpha, \beta)$
Phân phối chuẩn	$X \sim \mathcal{N}(\mu, _)$	$\theta = \mu$	$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$
	$X \sim \mathcal{N}(_, \sigma^2)$	$\theta = \sigma^2$	$\theta \sim \mathcal{I}nv\mathcal{G}amma(\alpha, \beta)$
Phân phối mũ	$X \sim \mathcal{E}xp(\beta)$	$\theta = \beta$	$\theta \sim \mathcal{G}amma(\alpha, \beta)$

Tới đây thì ta có thể sử dụng các phương pháp tối ưu hàm mục tiêu hét như với MLE chỉ khác là ta thêm các siêu tham số vào khi tính toán. Các siêu tham số này sẽ được thiết lập từ trước dựa vào cảm quan của người quan sát.

Ví dụ

Ta xét lại ví dụ tìm p của $X \sim \mathcal{B}ern(p)$ ở trên bằng phương pháp MAP. Vì đây là phân phối Bernoulli nên ta chọn phân phối $\mathcal{B}eta$ làm phân phối xác suất tiền nghiệm liên hợp $\theta \sim \mathcal{B}eta(\alpha, \beta)$ cho tham số $\theta = p$.

Như vậy ước lượng \hat{p} cần tìm là:

$$\begin{aligned}\hat{p} &= \operatorname{argmax}_{\theta} \left(\log \mathcal{B}eta(p) + \sum_{i=1}^n \log \mathcal{B}ern(X_i|p) \right) \\ &= \operatorname{argmax}_{\theta} \left(\log \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \right) + Y \log(p) + (n - Y) \log(1-p) \right)\end{aligned}$$

5. Mẫu thống kê và ước lượng tham số

Bỏ đi những thành phần hằng số (không phụ thuộc vào p) ta sẽ có:

$$\begin{aligned}\hat{p} &= \operatorname{argmax}_{\theta} \left(\log(p^{\alpha-1}(1-p)^{\beta-1}) + Y \log(p) + (n-Y) \log(1-p) \right) \\ &= \operatorname{argmax}_{\theta} \left((\alpha-1)\log(p) + (\beta-1)\log(1-p) + Y \log(p) + (n-Y) \log(1-p) \right) \\ &= \operatorname{argmax}_{\theta} \left((\alpha-1+Y)\log(p) + (\beta-1+n-Y)\log(1-p) \right)\end{aligned}$$

Giải bằng cách lấy đạo hàm tương tự như trên ta sẽ được:

$$\hat{p} = \frac{\alpha-1+\sum_{i=1}^n X_i}{\alpha+\beta-2+n} \quad (5.28)$$

Các siêu tham số α, β lúc này sẽ được chọn từ trước. Tuỳ vào giá trị của siêu tham số mà \hat{p} thể hiện khác nhau dẫn tới kết quả của mô hình là khác nhau.

5.4.5. EM

Trong nhiều bài toán, một mô hình xác suất có chứa các biến ẩn hoặc thiếu dữ liệu thì việc tính toán ước lượng của các tham số trở nên khó khăn hoặc không thực hiện được.

Cho mẫu ngẫu nhiên $\mathcal{D} = \{x_1, x_2, \dots, x_n\}$, ước lượng phân bố (X, Z) với lớp hàm phân bố $p(X, Z | \theta)$ trong đó Z là biến ẩn (không quan sát được). Ước lượng maximum likelihood cho log-likelihood

$$\theta = \arg \max_{\theta} \log p(x | \theta) \quad (5.29)$$

được thay bằng maximum kỳ vọng

Gọi phân bố cho biến Z là $q(z)$, ta có

$$\int_z q(z) dz = 1,$$

suy ra

$$\begin{aligned}\log p(x | \theta) &= \log p(x | \theta) \int_z q(z) dz \\ &= \int_z \log p(x | \theta) q(z) dz \\ &= \int_z \log p(x, z | \theta) q(z) dz - \int_z \log p(z | x, \theta) q(z) dz\end{aligned} \quad (5.30)$$

5. Mẫu thống kê và ước lượng tham số

trong đó,

$$\begin{aligned} \int_z \log p(x, z | \theta) q(z) dz &= \mathbb{E}_z [\log p(x, z | \theta)] \\ - \int_z \log p(z | x, \theta) q(z) dz &= \underbrace{\int_z \log \frac{q(z)}{p(z | x, \theta)} q(z) dz}_{KL[q || p(z | x, \theta)]} - \underbrace{\int_z \log q(z) q(z) dz}_{Entropy[q]} \end{aligned} \quad (5.31)$$

Như vậy, giá trị log-likelihood sẽ được phân rã thành

$$\underbrace{\log p(x | \theta)}_{\text{log-likelihood}} = \underbrace{\mathbb{E}_z [\log p(x, z | \theta)]}_{\text{kỳ vọng}} + \underbrace{KL [q || p(z | x, \theta)]}_{\text{khoảng cách KL} \geq 0} + \underbrace{Entropy[q]}_{\text{entropy}} \quad (5.32)$$

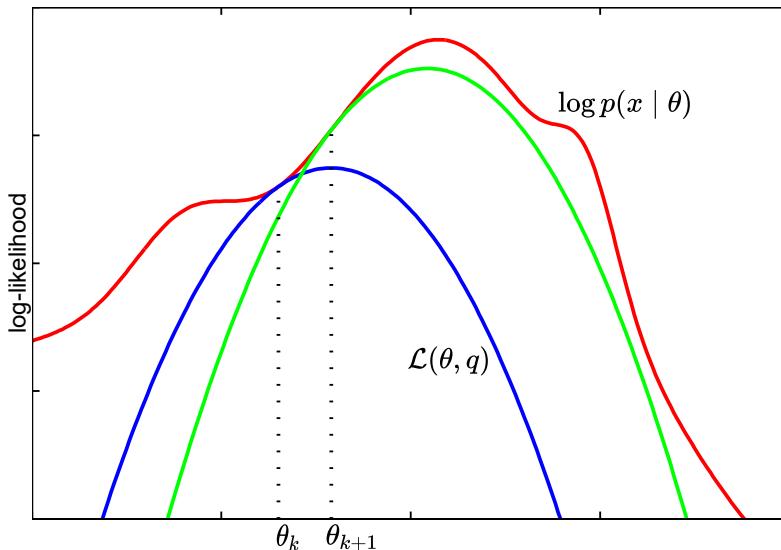
Từ biểu thức này có thể suy ra giá trị chặn dưới của log-likelihood là kỳ vọng + entropy. Ngoài ra, phân rã log-likelihood cũng có thể viết lại thành

$$\begin{aligned} \underbrace{\log p(x | \theta)}_{\text{log-likelihood}} &= \mathcal{L}(\theta, q) + \underbrace{KL [q || p(z | x, \theta)]}_{\text{khoảng cách KL} \geq 0} \\ &\geq \mathcal{L}(\theta, q) \end{aligned} \quad (5.33)$$

trong đó,

$$\mathcal{L}(\theta, q) = \int_z \log \frac{p(x, z | \theta)}{q(z)} q(z) dz \quad (5.34)$$

Tóm lại, bài toán cực đại hóa likelihood được chuyển thành bài toán cực đại hóa $\mathcal{L}(\theta, q)$.



5. Mẫu thống kê và ước lượng tham số

Thuật toán EM

Khởi tạo tham số θ^0 , phân phối q^0 và $t \leftarrow 0$

Lặp nếu chưa hội tụ

1. **Bước E:** Cực đại \mathcal{L} theo phân phối q , cố định tham số θ^t

$$q^{t+1} = \arg \max_q \mathcal{L}(\theta^t, q) \quad (5.35)$$

2. **Bước M:** Cực đại \mathcal{L} theo tham số θ , cố định phân phối q^t

$$\theta^{t+1} = \arg \max_{\theta} \mathcal{L}(\theta, q^t) \quad (5.36)$$

3. $t \leftarrow t + 1$

Lưu ý: thuật toán EM sẽ tăng giá trị likelihood sau mỗi lần lặp; tuy nhiên, nó không đảm bảo hội tụ đến ước lượng cực đại likelihood.

Ví dụ. Ước lượng phân phối Gaussian hỗn hợp (Mixture of Gaussians) cho tập dữ liệu N mẫu độc lập $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$

Xét phân phối Gaussian hỗn hợp là kết hợp của K phân phối chuẩn

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (5.37)$$

trong đó,

$$\sum_{k=1}^K \pi_k = 1 \text{ và } \pi_k \geq 0 \quad (5.38)$$

Biến ẩn: gọi $z_{ik} \in \{0, 1\}$ là biến nhị phân cho biết mẫu thứ i , \mathbf{x}_i có thuộc phân bố chuẩn thứ k hay không.

1. Khởi tạo $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ và π_k
2. **Bước E:** ước lượng phân phối cho biến ẩn

$$p(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (5.39)$$

5. Mẫu thống kê và ước lượng tham số

3. **Bước M:** ước lượng tham số cho các phân phối chuẩn

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{i=1}^N p(z_{ik}) \mathbf{x}_i \quad (5.40)$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{i=1}^N p(z_{ik}) (\mathbf{x}_i - \boldsymbol{\mu}_k^{new})(\mathbf{x}_i - \boldsymbol{\mu}_k^{new})^\top \quad (5.41)$$

$$\pi_k^{new} = \frac{N_k}{N} \quad (5.42)$$

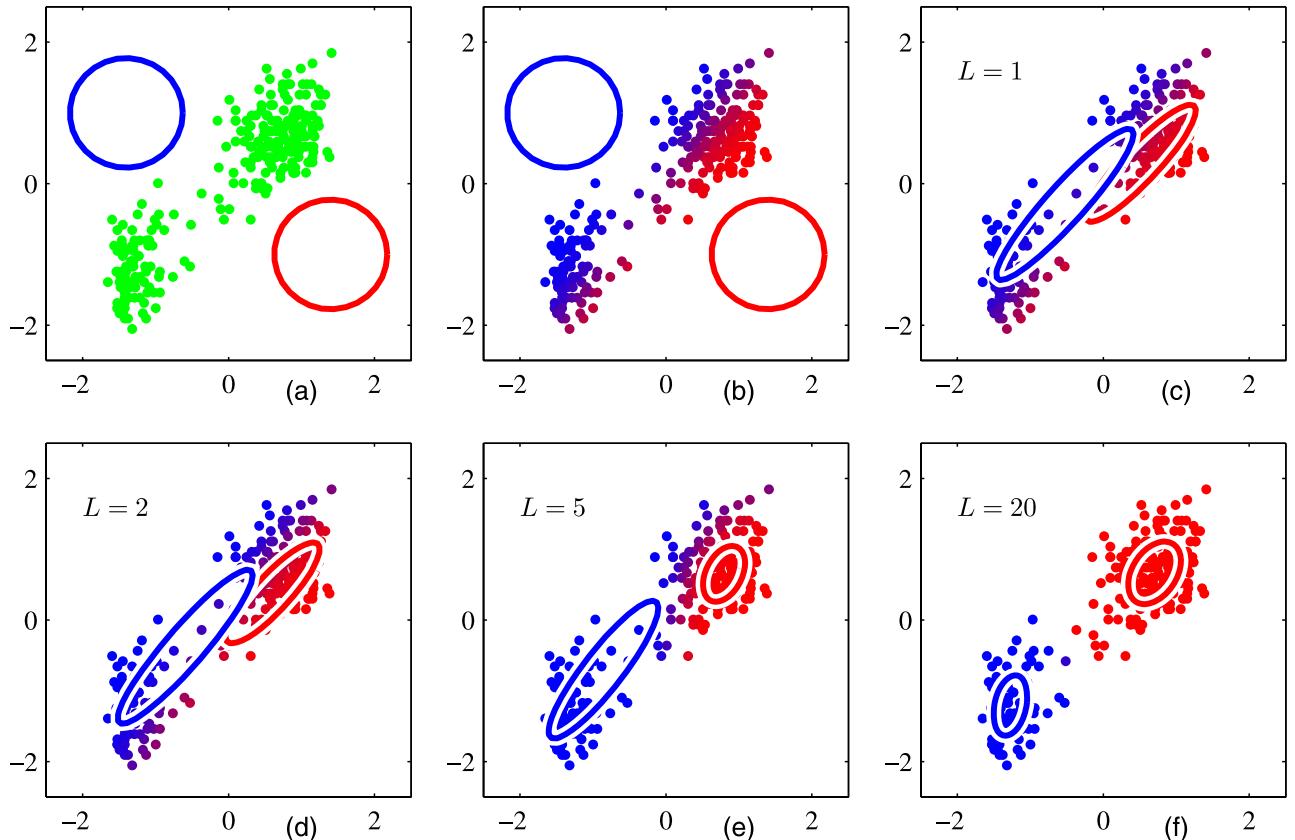
trong đó,

$$N_k = \sum_{i=1}^N p(z_{ik}) \quad (5.43)$$

4. Tính giá trị log-likelihood

$$\log p(\mathcal{D} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^N \log \sum_{k=1}^K \pi_k \cdot \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (5.44)$$

và kiểm tra sự hội tụ, nếu chưa hội tụ thì quay lại Bước 2



5.5**Ước lượng khoảng tin cậy cho tham số**

Phương pháp ước lượng điểm nói trên có nhược điểm là khi kích thước mẫu bé thì ước lượng điểm có thể sai lệch rất nhiều so với giá trị của tham số cần ước lượng. Một khía cạnh phương pháp trên cũng không thể đánh giá được khả năng mắc sai lầm khi ước lượng là bao nhiêu. Do đó khi kích thước mẫu bé người ta thường dùng phương pháp ước lượng khoảng tin cậy cho trường hợp một tham số.

5.5.1. Khái niệm ước lượng khoảng tin cậy

Giả sử chưa biết đặc trưng θ nào đó của biến ngẫu nhiên X . Ước lượng khoảng của θ là chỉ ra một khoảng số (g_1, g_2) nào đó chứa θ , tức là có thể ước lượng $g_1 < \theta < g_2$.

5.5.2. Phương pháp

Thủ tục chung ước lượng khoảng tin cậy cho tham số θ của biến ngẫu nhiên X

- Lập mẫu ngẫu nhiên $\mathcal{D} = \{x_1, x_2, \dots, x_n\}$
- Xác định quy luật phân phối $G(\theta)$ ứng với biến ngẫu nhiên X
- Xác định mức ý nghĩa $\alpha \rightarrow$ độ tin cậy $\gamma = 1 - \alpha$
- Tính khoảng tin cậy G_1, G_2 với độ tin cậy là

Khoảng tin cậy cho kỳ vọng của biến ngẫu nhiên liên tục

Bài toán. Giả sử biến ngẫu nhiên X tuân theo luật phân phối với kỳ vọng $\mathbb{E}(X) = \mu$ chưa biết. Hãy ước lượng khoảng tin cậy μ từ một tập mẫu $\{x_1, \dots, x_n\}$.

1. Trường hợp đã biết phương sai σ^2

- Chọn thống kê z (phân phối chuẩn $\mathcal{N}(0, 1)$)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- Chọn mức ý nghĩa α (có thể là 0.05, 0.01 ...) và tính các khoảng tin cậy với độ tin cậy là γ
 - a) khoảng tin cậy đối xứng

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z \left(1 - \frac{\alpha}{2} \right), \bar{x} + \frac{\sigma}{\sqrt{n}} z \left(1 - \frac{\alpha}{2} \right) \right) \quad (5.45)$$

5. Mẫu thống kê và ước lượng tham số

b) khoảng tin cây trái

$$\left(-\infty, \bar{x} + \frac{\sigma}{\sqrt{n}} z(1-\alpha) \right) \quad (5.46)$$

c) khoảng tin cây phải

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z(1-\alpha), \infty \right) \quad (5.47)$$

2. Trường hợp chưa biết phương sai, cỡ mẫu $n < 30$

- Chọn thống kê t (phân phối Student)

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- Chọn mức ý nghĩa α (có thể là 0.05, 0.01 ...) và tính khoảng tin cây

$$\left(\bar{x} - \frac{s}{\sqrt{n}} t \left(1 - \frac{\alpha}{2}; n-1 \right), \bar{x} + \frac{s}{\sqrt{n}} t \left(1 - \frac{\alpha}{2}; n-1 \right) \right) \quad (5.48)$$

với độ tin cây là γ

3. Trường hợp chưa biết phương sai, cỡ mẫu $n \geq 30$

- Chọn thống kê z (phân phối chuẩn $\mathcal{N}(0, 1)$)

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- Chọn mức ý nghĩa α (có thể là 0.05, 0.01 ...) và tính khoảng tin cây

$$\left(\bar{x} - \frac{s}{\sqrt{n}} z \left(1 - \frac{\alpha}{2} \right), \bar{x} + \frac{s}{\sqrt{n}} z \left(1 - \frac{\alpha}{2} \right) \right) \quad (5.49)$$

với độ tin cây là γ

Khoảng tin cây cho tỷ lệ của biến ngẫu nhiên rời rạc

Bài toán. Trong một tập mẫu \mathcal{D} có tổng số là n mẫu, một biến ngẫu nhiên rời rạc X có giá trị x xuất hiện m lần. Vậy tần suất xuất hiện của x là $\hat{p} = m/n$ là ước lượng điểm không chênh p . Với độ tin cây $\gamma = 1 - \alpha$ hãy ước lượng khoảng cho p .

- Chọn thống kê z (phân phối chuẩn $\mathcal{N}(0, 1)$)

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

5. Mẫu thống kê và ước lượng tham số

- Chọn mức ý nghĩa α (có thể là 0.05, 0.01 ...) và tính khoảng tin cậy

$$\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z \left(1 - \frac{\alpha}{2} \right), \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z \left(1 - \frac{\alpha}{2} \right) \right) \quad (5.50)$$

với độ tin cậy là γ

5.6

Ước lượng kích thước mẫu

Bài toán ước lượng trung bình tổng thể

Bài toán. Với độ tin cậy $\gamma = 1 - \alpha$, nếu ta muốn sai số ϵ của ước lượng trung bình tổng thể đạt được ở một mức nào đó cho trước ϵ_0 thì kích thước mẫu n tối thiểu là bao nhiêu?

- Nếu biết $\text{Var}(X) = \sigma^2$, từ công thức

$$\epsilon = \frac{\sigma}{\sqrt{n}} z \left(1 - \frac{\alpha}{2} \right),$$

để $\epsilon \leq \epsilon_0$ ta cần chọn

$$n \geq \left[z \left(1 - \frac{\alpha}{2} \right) \frac{\sigma}{\epsilon_0} \right]^2 \quad (5.51)$$

- Nếu chưa biết σ^2 , ta căn cứ vào mẫu đã cho để tính s^2 . Từ đó ta xác định được kích thước mẫu tối thiểu

$$n \geq \left[z \left(1 - \frac{\alpha}{2} \right) \frac{s}{\epsilon_0} \right]^2 \quad (5.52)$$

Bài toán ước lượng tỉ lệ tổng thể

Bài toán. Với độ tin cậy $\gamma = 1 - \alpha$, nếu ta muốn sai số ϵ của ước lượng tỉ lệ tổng thể đạt được ở một mức nào đó cho trước ϵ_0 thì kích thước mẫu n tối thiểu là bao nhiêu?

- Ta có

$$\epsilon = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z \left(1 - \frac{\alpha}{2} \right),$$

do $\hat{p}(1-\hat{p})$ đạt giá trị cực đại $\frac{1}{4}$ khi $\hat{p} = 0.5$ nên

$$\epsilon \leq \sqrt{\frac{1}{4n}} z \left(1 - \frac{\alpha}{2} \right).$$

Do đó, để $\epsilon \leq \epsilon_0$ ta cần chọn

$$n \geq \frac{1}{4\epsilon_0^2} z \left(1 - \frac{\alpha}{2} \right)^2 \quad (5.53)$$

Bài tập

Mẫu thống kê

B 5.1. Số liệu về chiều cao của các sinh viên nữ (Đơn vị: inch) trong một lớp học như sau:

62 64 66 67 65 68 61 65 67 65 64 63 67
68 64 66 68 69 65 67 62 66 68 67 66 65
69 65 70 65 67 68 65 63 64 67 67

1. Tính chiều cao trung bình và độ lệch tiêu chuẩn.
2. Trung vị của chiều cao sinh viên lớp này là bao nhiêu?

B 5.2. Cho bộ dữ liệu sau:

4.2 4.7 4.7 5.0 3.8 3.6 3.0 5.1 3.1 3.8
4.8 4.0 5.2 4.3 2.8 2.0 2.8 3.3 4.8 5.0

Tính trung bình mẫu, phương sai mẫu và độ lệch tiêu chuẩn.

B 5.3. Cho bộ dữ liệu sau:

43 47 51 48 52 50 46 49
45 52 46 51 44 49 46 51
49 45 44 50 48 50 49 50

Tính trung bình mẫu, phương sai mẫu và độ lệch tiêu chuẩn.

B 5.4. Xét biểu thức $y = \sum_{i=1}^n (x_i - a)^2$. Với a nào thì y đạt giá trị nhỏ nhất?

B 5.5. Xét $y_i = a + bx_i$, $i = 1, \dots, n$ và a, b là các hằng số khác 0. Hãy tìm mối liên hệ giữa \bar{x} và \bar{y} , s_x và s_y .

B 5.6. Giả sử ta có mẫu cỡ n gồm các giá trị quan trắc x_1, x_2, \dots, x_n và đã tính được trung bình mẫu \bar{x}_n và phương sai mẫu s_n^2 . Quan trắc thêm giá trị thứ $(n+1)$ là x_{n+1} , gọi \bar{x}_{n+1} và s_{n+1}^2 lần lượt là trung bình mẫu và phương sai mẫu ứng với mẫu có $(n+1)$ quan trắc.

1. Tính \bar{x}_{n+1} theo \bar{x}_n và x_{n+1} .
2. Chứng tỏ rằng

$$ns_{n+1}^2 = (n-1)s_n^2 + \frac{n(x_{n+1} - \bar{x}_n)^2}{n+1}$$

B 5.7. Từ bảng các số ngẫu nhiên người ta lấy ra 150 số. Các số đó được phân thành 10 khoảng như sau:

5. Mẫu thống kê và ước lượng tham số

x_i	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
n_i	16	15	19	13	14	19	14	11	13	16

Xác định trung bình mẫu và phương sai mẫu.

B 5.8. Khảo sát thu nhập của công nhân ở một công ty, cho bởi bảng sau (đơn vị ngàn đồng).

Thu nhập	[500-600]	[600-700]	[700-800]	[800-900]	[900-1000]	[1000-1100]	[1100-1200]
Số người	2	10	15	30	25	14	4

Xác định thu nhập trung bình, độ lệch chuẩn.

B 5.9. Đo lượng huyết tương của 8 người mạnh khoẻ, ta có

2,863,372,752,623,503,253,123,15

Hãy xác định các đặc trưng mẫu.

B 5.10. Quan sát thời gian cần thiết để sản xuất một chi tiết máy, ta thu được số liệu cho bảng sau:

Khoảng thời gian (phút)	Số lần quan sát
20-25	2
25-30	14
30-35	26
35-40	32
40-45	14
45-50	8
50-55	4

Tính trung bình mẫu \bar{x} , phương sai mẫu s^2 .

B 5.11. Đo độ dài của một loại trực xe, ta có kết quả

Nhóm	18.4-18.6	18.6-18.8	18.8-19	19-19.2	19.2-19.4	19.4-19.6	19.6-19.8
n_i	1	4	20	41	19	8	4

Hãy tính độ dài trung bình và phương sai mẫu

B 5.12. Số liệu về tiền lương (USD/tháng) của 2 nhóm công nhân như sau:

- Nhóm 1: 300, 400, 500, 600, 700
- Nhóm 2: 400, 450, 500, 550, 600

1. So sánh số trung bình mẫu về tiền lương giữa 2 nhóm công nhân.

2. So sánh độ lệch chuẩn mẫu hiệu chỉnh về tiền lương giữa 2 nhóm công nhân. Nhận xét.

5. Mẫu thống kê và ước lượng tham số

Ước lượng trung bình tổng thể

B 5.13. Trên tập mẫu gồm 100 số liệu, người ta tính được $\bar{x} = 0.1$, $s = 0.014$. Xác định khoảng tin cậy 95% cho giá trị trung bình thật.

B 5.14. Chọn ngẫu nhiên 36 công nhân của xí nghiệp thì thấy lương trung bình là 7600 ngàn đồng/tháng. Giả sử lương công nhân tuân theo phân phối chuẩn với $\sigma = 140$ ngàn đồng. Với độ tin cậy 95%, hãy ước lượng mức lương trung bình của công nhân trong toàn xí nghiệp.

B 5.15. Đo sức bền chịu lực của một loại ống thí nghiệm, người ta thu được bộ số liệu sau

$$4500, 6500, 5200, 4800, 4900, 5125, 6200, 5375$$

Từ kinh nghiệm nghề nghiệp, người ta cũng biết rằng sức bền đó có phân phối chuẩn với độ lệch chuẩn $\sigma = 300$. Hãy xây dựng khoảng tin cậy 90% cho sức bền trung bình của loại ống trên.

B 5.16. Sản lượng mỗi ngày của một phân xưởng là biến ngẫu nhiên tuân theo luật chuẩn. Kết quả thống kê của 9 ngày cho ta:

$$27, 26, 21, 28, 25, 30, 26, 23, 26$$

Hãy xác định các khoảng tin cậy 95% cho sản lượng trung bình.

B 5.17. Quan sát chiều cao X (cm) của một số người, ta ghi nhận

x (cm)	140-145	145-150	150-155	155-160	160-165	165-170
Số người	1	3	7	9	5	2

1. Tính \bar{x} và s^2
2. Ước lượng μ ở độ tin cậy 0.95

B 5.18. Điểm trung bình môn toán của 100 thí sinh dự thi vào trường A là 5 với độ lệch chuẩn là 2.5.

1. Ước lượng điểm trung bình môn toán của toàn thể thí sinh với độ tin cậy là 95%.
2. Với sai số ước lượng điểm trung bình ở câu (1) là 0.25 điểm, hãy xác định độ tin cậy.

B 5.19. Tuổi thọ của một loại bóng đèn được biết theo quy luật chuẩn với độ lệch chuẩn 100 giờ.

1. Chọn ngẫu nhiên 100 bóng đèn để thử nghiệm, thấy mỗi bóng tuổi thọ trung bình là 1000 giờ. Hãy ước lượng tuổi thọ trung bình của bóng đèn xí nghiệp A sản xuất với độ tin cậy là 95%.
2. Với dung sai của ước lượng tuổi thọ trung bình là 15 giờ, hãy xác định độ tin cậy.

5. Mẫu thống kê và ước lượng tham số

3. Để dung sai của ước lượng tuổi thọ trung bình không quá 25 giờ với độ tin cậy là 95% thì cần phải thử nghiệm ít nhất bao nhiêu bóng.

B 5.20. Khối lượng các bao bột mì tại một cửa hàng lương thực tuân theo phân phối chuẩn. Kiểm tra 20 bao, thấy khối lượng trung bình của mỗi bao bột mì là 48kg, và phương sai mẫu $s^2 = (0.5kg)^2$.

1. Với độ tin cậy 95% hãy ước lượng khối lượng trung bình của một bao bột mì thuộc cửa hàng.
2. Với dung sai của ước lượng ở câu (1) là 0.284kg, hãy xác định độ tin cậy.
3. Để dung sai của ước lượng ở câu (1) không quá 160g với độ tin cậy là 95%, cần phải kiểm tra ít nhất bao nhiêu bao?

B 5.21. Đo đường kính của một chi tiết máy do một máy tiện tự động sản xuất, ta ghi nhận được số liệu như sau: với n chỉ số trường hợp tính theo từng giá trị của X (mm).

x	12.00	12.05	12.10	12.15	12.20	12.25	12.30	12.35	12.40
n	2	3	7	9	10	8	6	5	3

1. Tính trung bình mẫu \bar{x} và độ lệch chuẩn s của mẫu.
2. Ước lượng đường kính trung bình μ ở độ tin cậy 0.95.
3. Nếu muốn sai số ước lượng không quá $\epsilon = 0.02mm$ ở độ tin cậy 0.95 thì phải quan sát ít nhất mấy trường hợp.

B 5.22. Người ta đo ion Na^+ trên một số người và ghi nhận lại được kết quả như sau

$$129, 132, 140, 141, 138, 143, 133, 137, 140, 143, 138, 140$$

1. Tính trung bình mẫu \bar{x} và phương sai mẫu s^2 .
2. Ước lượng trung bình μ của tổng thể ở độ tin cậy 0.95.
3. Nếu muốn sai số ước lượng trung bình không quá $\epsilon = 1$ với độ tin cậy 0.95 thì phải quan sát mẫu gồm ít nhất mấy người?

B 5.23. Quan sát tuổi thọ x (giờ) của một số bóng đèn do xí nghiệp A sản xuất, ta ghi nhận với n chỉ số trường hợp theo từng giá trị của x .

x	1000	1100	1200	1300	1400	1500	1600	1700	1800
n	10	14	16	17	18	16	16	12	9

1. Tính trung bình mẫu \bar{x} và độ lệch chuẩn mẫu s .

5. Mẫu thống kê và ước lượng tham số

2. Ước lượng tuổi thọ trung bình của bóng đèn ở độ tin cậy 0.95.
3. Nếu muốn sai số ước lượng không quá $\epsilon = 30$ giờ với độ tin cậy 0.99 thì phải quan sát mẫu gồm ít nhất mấy bóng đèn?

B 5.24. Chiều dài của một loại sản phẩm được xuất khẩu hàng loạt là biến ngẫu nhiên phân phối chuẩn với $\mu = 100mm$ và $\sigma^2 = 4^2 mm^2$. Kiểm tra ngẫu nhiên 25 sản phẩm. Khả năng chiều dài trung bình của số sản phẩm kiểm tra nằm trong khoảng từ $98mm$ đến $101mm$ là bao nhiêu?

Ước lượng tỉ lệ tổng thể

B 5.25. Trước bầu cử, người ta phỏng vấn ngẫu nhiên 2000 cử tri thì thấy có 1380 người ủng hộ một ứng cử viên K. Với độ tin cậy 95%, hỏi ứng cử viên đó thu được tối thiểu bao nhiêu phần trăm phiếu bầu?

B 5.26. Một loại bệnh có tỷ lệ tử vong là 0.01. Muốn chứng tỏ một loại thuốc có hiệu nghiệm (nghĩa là hạ thấp được tỷ lệ tử vong nhỏ hơn 0.005) ở độ tin cậy 0.95 thì phải thử thuốc đó trên ít nhất bao nhiêu người?

B 5.27. Để ước lượng xác suất mắc bệnh gan với độ tin cậy 90% và sai số không vượt quá 2% thì cần phải khám ít nhất bao nhiêu người, biết rằng tỷ lệ mắc bệnh gan thực nghiệm đã cho bằng 0.9.

B 5.28. Giả sử quan sát 100 người thấy có 20 người bị bệnh sốt xuất huyết. Hãy ước lượng tỷ lệ bệnh sốt xuất huyết ở độ tin cậy 97%. Nếu muốn sai số ước lượng không quá 3% ở độ tin cậy 95% thì phải quan sát ít nhất bao nhiêu người?

B 5.29. Một loại thuốc mới đem điều trị cho 50 người bị bệnh B, kết quả có 40 người khỏi bệnh.

1. Ước lượng tỷ lệ khỏi bệnh p nếu dùng thuốc đó điều trị với độ tin cậy 0.95 và 0.99.
2. Nếu muốn sai số ước lượng không quá 0.02 ở độ tin cậy 0.95 thì phải quan sát ít nhất mấy trường hợp?

B 5.30. Ta muốn ước lượng tỷ lệ viên thuốc bị sứt mẻ p trong một lô thuốc lớn.

1. Nếu muốn sai số ước lượng không quá 0.01 với độ tin cậy 0.95 thì phải quan sát ít nhất mấy viên?
2. Quan sát ngẫu nhiên 200 viên, thấy có 18 viên bị sứt mẻ. Hãy ước lượng p ở độ tin cậy 0.95.
3. Khi đó, nếu muốn sai số ước lượng không quá 0.01 với độ tin cậy 0.95 thì phải quan sát ít nhất mấy viên?

5. Mẫu thống kê và ước lượng tham số

B 5.31. Muốn biết trong ao có bao nhiêu cá, người ta bắt lên 2000 con, đánh dấu xong lại thả xuống hồ. Sau một thời gian, người ta bắt lên 500 con và thấy có 20 con cá có đánh dấu của lần bắt trước. Dựa vào kết quả đó hãy ước lượng số cá có trong hồ với độ tin cậy 95%.

B 5.32. Để có thể dự đoán được số lượng chim thường nghỉ tại vườn nhà mình, người chủ bắt 89 con, đeo khoen cho chúng rồi thả đi. Sau một thời gian, ông bắt ngẫu nhiên được 120 con và thấy có 7 con có đeo khoen. Hãy dự đoán số chim giúp ông chủ vườn ở độ tin cậy 99%.

Tổng hợp

B 5.33. Cân thử 100 quả cam, ta có bộ số liệu sau:

Khối lượng (g)	32	33	34	35	36	37	38	39	40
Số quả	2	3	15	26	28	6	8	8	4

1. Hãy ước lượng khối lượng trung bình các quả cam ở độ tin cậy 95%.
2. Cam có khối lượng dưới 34g được coi là cam loại 2. Tìm khoảng ước lượng cho tỷ lệ loại 2 với độ tin cậy 90%.

B 5.34. Đem cân một số trái cây vừa thu hoạch, ta được kết quả sau:

Khối lượng (gam)	200-210	210-220	220-230	230-240	240-250
Số trái	12	17	20	18	15

1. Tìm khoảng ước lượng của trọng lượng trung bình μ của trái cây với độ tin cậy 0.95 và 0.99.
2. Nếu muốn sai số ước lượng không quá $\epsilon = 2\text{ gam}$ ở độ tin cậy 99% thì phải quan sát ít nhất bao nhiêu trái?
3. Trái cây có khối lượng $X \geq 230\text{ gam}$ được xếp vào loại A. Hãy tìm khoảng ước lượng cho tỷ lệ p của trái cây loại A ở độ tin cậy 0.95 và 0.99. Nếu muốn sai số ước lượng không quá 0.04 ở độ tin cậy 0.99 thì phải quan sát ít nhất mấy trường hợp?

Kiểm định giả thuyết thống kê là một dạng phân tích thống kê. Đây là một phương pháp quan trọng cho phép giải quyết nhiều bài toán trong thực tế. Nội dung của kiểm định giả thuyết thống kê là dựa vào mẫu cụ thể và các quy tắc hay thủ tục quyết định dẫn đến bác bỏ hay chấp nhận giả thuyết của tổng thể.

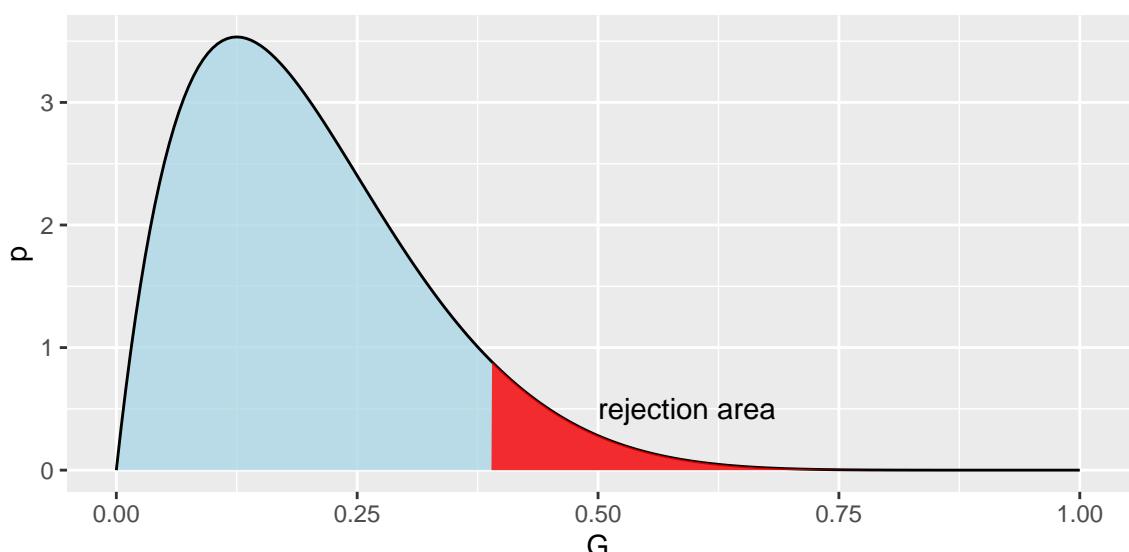
6.1

Thủ tục kiểm định giả thuyết thống kê

Một thủ tục kiểm định giả thuyết thống kê bao gồm:

1. Phát biểu giả thuyết H_0 và đối thuyết H_a .
2. Từ tổng thể nghiên cứu lập mẫu ngẫu nhiên kích thước n . Chọn tiêu chuẩn kiểm định G và xác định quy luật phân phối xác suất của G với điều kiện giả thuyết H_0 đúng.
3. Với mức ý nghĩa α , xác định miền bác bỏ giả thuyết H_0 (ký hiệu là W_α) tốt nhất tùy thuộc vào đối thuyết H_a .
4. Bác bỏ H_0 nếu giá trị kiểm định $\in W_\alpha$.

Tiêu chuẩn kiểm định G thường là các phân phối như: z (phân phối chuẩn), t (phân phối Student), χ^2 (phân phối Chi squared) và F (phân phối Fisher).



6. Kiểm định thống kê

Lưu ý

Trong các tính toán thống kê trên máy tính ngày nay, người ta hay sử dụng thủ tục sau

1. Tính giá trị z (hoặc t, q)
2. Tính giá trị xác suất p -value tương ứng với z
3. Bác bỏ H_0 nếu p -value $< \alpha$

6.1.1. Sai lầm loại I và II

Khi đưa ra quyết định về chấp nhận hay bác bỏ một giả thuyết có thể chúng ta đúng hoặc sai. Bảng sau tóm tắt các tình huống

		Quyết định	
		chấp nhận H_0	bác bỏ H_0
Chân lý	H_0 đúng	Quyết định đúng (xác suất $1 - \alpha$)	Sai lầm loại I (xác suất α)
	H_a đúng	Sai lầm loại II (xác suất β)	Quyết định đúng (xác suất $1 - \beta$)

6.2 Kiểm định tham số

6.2.1. Kiểm định một tham số, một tổng thể, một mẫu

Trung bình tổng thể (biết phương sai tổng thể)

1. Kiểm định hai phía

<ul style="list-style-type: none"> • Giả thuyết $H_0 : \mu = \mu_0$ • Đối thuyết $H_a : \mu \neq \mu_0$ • Biết σ • Mức ý nghĩa α 	<ul style="list-style-type: none"> • Tính z $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ <ul style="list-style-type: none"> • Miền bác bỏ H_0 $ z > z\left(\frac{\alpha}{2}\right)$
--	---

2. Kiểm định một phía

<ul style="list-style-type: none"> • Giả thuyết $H_0 : \mu = \mu_0$ • Đối thuyết $H_a : \mu > \mu_0$ • Biết σ 	<ul style="list-style-type: none"> • Tính z <ul style="list-style-type: none"> • Miền bác bỏ H_0 $z > z(\alpha)$
---	--

6. Kiểm định thống kê

3. Kiểm định một phía

- | | |
|---|---|
| <ul style="list-style-type: none">Giả thuyết $H_0 : \mu = \mu_0$Đối thuyết $H_a : \mu < \mu_0$Biết σ | <ul style="list-style-type: none">Tính zMiền bác bỏ H_0 $z < -z(\alpha)$ |
|---|---|

Trung bình tổng thể (không biết phương sai tổng thể)

1. Kiểm định hai phía

- | | |
|--|---|
| <ul style="list-style-type: none">Giả thuyết $H_0 : \mu = \mu_0$Đối thuyết $H_a : \mu \neq \mu_0$Mức ý nghĩa α | <ul style="list-style-type: none">Tính tMiền bác bỏ H_0 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ $ t > t\left(\frac{\alpha}{2}; n - 1\right)$ |
|--|---|

2. Kiểm định một phía

- | | |
|--|---|
| <ul style="list-style-type: none">Giả thuyết $H_0 : \mu = \mu_0$Đối thuyết $H_a : \mu > \mu_0$ | <ul style="list-style-type: none">Tính tMiền bác bỏ H_0 $t > t(\alpha; n - 1)$ |
|--|---|

3. Kiểm định một phía

- | | |
|--|--|
| <ul style="list-style-type: none">Giả thuyết $H_0 : \mu = \mu_0$Đối thuyết $H_a : \mu < \mu_0$ | <ul style="list-style-type: none">Tính tMiền bác bỏ H_0 $t < -t(\alpha; n - 1)$ |
|--|--|

Phương sai tổng thể

1. Kiểm định hai phía

6. Kiểm định thống kê

<ul style="list-style-type: none"> Giả thuyết $H_0 : \sigma = \sigma_0$ Đối thuyết $H_a : \sigma \neq \sigma_0$ Mức ý nghĩa α 	<ul style="list-style-type: none"> Tính q $q = \frac{(n - 1)s^2}{\sigma_0^2}$ <ul style="list-style-type: none"> Miền bác bỏ H_0 $q > \chi^2 \left(\frac{\alpha}{2}; n - 1 \right)$ <p>hoặc</p> $q < \chi^2 \left(1 - \frac{\alpha}{2}; n - 1 \right)$
--	--

2. Kiểm định một phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \sigma = \sigma_0$ Đối thuyết $H_a : \sigma > \sigma_0$ 	<ul style="list-style-type: none"> Tính q Miền bác bỏ H_0 $q > \chi^2 (\alpha; n - 1)$
---	--

3. Kiểm định một phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \sigma = \sigma_0$ Đối thuyết $H_a : \sigma < \sigma_0$ 	<ul style="list-style-type: none"> Tính q Miền bác bỏ H_0 $q < \chi^2 (1 - \alpha; n - 1)$
---	--

Tỉ lệ tổng thể

1. Kiểm định hai phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : p = p_0$ Đối thuyết $H_a : p \neq p_0$ Mức ý nghĩa α 	<ul style="list-style-type: none"> Tính z $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$ <ul style="list-style-type: none"> Miền bác bỏ H_0 $ z > z \left(\frac{\alpha}{2} \right)$
--	--

6. Kiểm định thống kê

2. Kiểm định một phía

- | | |
|---|--|
| <ul style="list-style-type: none"> • Giả thuyết $H_0 : p = p_0$ • Đối thuyết $H_a : p > p_0$ | <ul style="list-style-type: none"> • Tính z • Miền bác bỏ H_0 <p style="text-align: center;">$z > z(\alpha)$</p> |
|---|--|

3. Kiểm định một phía

- | | |
|---|---|
| <ul style="list-style-type: none"> • Giả thuyết $H_0 : p = p_0$ • Đối thuyết $H_a : p < p_0$ | <ul style="list-style-type: none"> • Tính z • Miền bác bỏ H_0 <p style="text-align: center;">$z < -z(\alpha)$</p> |
|---|---|

6.2.2. Kiểm định hai tham số, hai tổng thể, hai mẫu

Hai trung bình tổng thể (giả sử phương sai bằng nhau)

1. Kiểm định hai phía

- | | |
|--|--|
| <ul style="list-style-type: none"> • Giả thuyết $H_0 : \mu_1 = \mu_2$ • Đối thuyết $H_a : \mu_1 \neq \mu_2$ • Mức ý nghĩa α | <ul style="list-style-type: none"> • Tính t $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ <p>với</p> $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ <ul style="list-style-type: none"> • Miền bác bỏ H_0 $ t > t \left(\frac{\alpha}{2}; n_1 + n_2 - 2 \right)$ |
|--|--|

2. Kiểm định một phía

6. Kiểm định thống kê

<ul style="list-style-type: none"> Giả thuyết $H_0 : \mu_1 = \mu_2$ Đối thuyết $H_a : \mu_1 > \mu_2$ 	<ul style="list-style-type: none"> Tính t Miền bác bỏ H_0 $t > t(\alpha; n_1 + n_2 - 2)$
---	--

3. Kiểm định một phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \mu_1 = \mu_2$ Đối thuyết $H_a : \mu_1 < \mu_2$ 	<ul style="list-style-type: none"> Tính t Miền bác bỏ H_0 $t < -t(\alpha; n_1 + n_2 - 2)$
---	---

Hai trung bình tổng thể (giả sử phương sai khác nhau)

1. Kiểm định hai phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \mu_1 = \mu_2$ Đối thuyết $H_a : \mu_1 \neq \mu_2$ Mức ý nghĩa α $n_1 \geq 30, n_2 \geq 30$ 	<ul style="list-style-type: none"> Tính z $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ Miền bác bỏ H_0 $ z > z\left(\frac{\alpha}{2}\right)$
---	---

2. Kiểm định một phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \mu_1 = \mu_2$ Đối thuyết $H_a : \mu_1 > \mu_2$ 	<ul style="list-style-type: none"> Tính z Miền bác bỏ H_0 $z > z(\alpha)$
---	---

3. Kiểm định một phía

6. Kiểm định thống kê

<ul style="list-style-type: none"> Giả thuyết $H_0 : \mu_1 = \mu_2$ Đối thuyết $H_a : \mu_1 < \mu_2$ 	<ul style="list-style-type: none"> Tính z Miền bác bỏ H_0 $z < -z(\alpha)$
---	--

Hai phương sai tổng thể

1. Kiểm định hai phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \sigma_1 = \sigma_2$ Đối thuyết $H_a : \sigma_1 \neq \sigma_2$ Mức ý nghĩa α 	<ul style="list-style-type: none"> Tính f $f = \frac{s_1^2}{s_2^2}$ <ul style="list-style-type: none"> Miền bác bỏ H_0 $f > F\left(\frac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$ <p>hoặc</p> $f < F\left(1 - \frac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$
--	--

2. Kiểm định một phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \sigma_1 = \sigma_2$ Đối thuyết $H_a : \sigma_1 > \sigma_2$ 	<ul style="list-style-type: none"> Tính f Miền bác bỏ H_0 $f > F(\alpha; n_1 - 1, n_2 - 1)$
---	---

3. Kiểm định một phía

<ul style="list-style-type: none"> Giả thuyết $H_0 : \sigma_1 = \sigma_2$ Đối thuyết $H_a : \sigma_1 < \sigma_2$ 	<ul style="list-style-type: none"> Tính f Miền bác bỏ H_0 $f < F(1 - \alpha; n_1 - 1, n_2 - 1)$
---	---

Hai tỉ lệ tổng thể

1. Kiểm định hai phía

6. Kiểm định thống kê

- Giả thuyết $H_0 : p_1 = p_2$
- Đối thuyết $H_a : p_1 \neq p_2$
- Mức ý nghĩa α

- Tính z

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

với

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

- Miền bác bỏ H_0

$$|z| > z\left(\frac{\alpha}{2}\right)$$

2. Kiểm định một phía

- Giả thuyết $H_0 : p_1 = p_2$
- Đối thuyết $H_a : p_1 > p_2$

- Tính z
- Miền bác bỏ H_0

$$z > z(\alpha)$$

3. Kiểm định một phía

- Giả thuyết $H_0 : p_1 = p_2$
- Đối thuyết $H_a : p_1 < p_2$

- Tính z
- Miền bác bỏ H_0

$$z < -z(\alpha)$$

6.3**Kiểm định phi tham số****6.3.1. Kiểm định tính độc lập của hai dấu hiệu định tính**

<ul style="list-style-type: none"> Giả thuyết H_0: hai biến độc lập Đối thuyết H_a: hai biến không độc lập Mức ý nghĩa α 	<ul style="list-style-type: none"> Tính q từ bảng tương quan $q = n \left[\sum_{i=1}^h \sum_{j=1}^k \frac{n_i m_j}{n_{ij}^2} - 1 \right]$ <ul style="list-style-type: none"> Miền bác bỏ H_0 $q > \chi^2(\alpha; (h-1)(k-1))$
---	--

6.3.2. Kiểm định tính phân phối chuẩn - Jacque-Berra

<ul style="list-style-type: none"> Giả thuyết H_0: biến phân phối chuẩn Đối thuyết H_a: biến không phân phối chuẩn Mức ý nghĩa α 	<ul style="list-style-type: none"> Tính q $q = n \left[\frac{skew^2}{6} + \frac{kurt^2}{24} \right]$ <ul style="list-style-type: none"> Miền bác bỏ H_0 $q > \chi^2(\alpha; 2)$
---	--

6.3.3. Kiểm định phân phối - Kolmogorov-Smirnov

<ul style="list-style-type: none"> Giả thuyết H_0: phân phối thực nghiệm F_o và phân phối mô hình F_m là tương đồng Đối thuyết H_a: phân phối thực nghiệm F_o và phân phối mô hình F_m là khác nhau Mức ý nghĩa α 	<ul style="list-style-type: none"> Tính d $d = \max(F_o(x) - F_m(x))$ <ul style="list-style-type: none"> Bác bỏ H_0 $d > D(\alpha)$
--	---

 **Bài tập**

So sánh kì vọng với một số cho trước

B 6.1. Giám đốc một xí nghiệp cho biết lương trung bình của 1 công nhân thuộc xí nghiệp là 380 ngàn đ/tháng. Chọn ngẫu nhiên 36 công nhân thấy lương trung bình là 350 ngàn đ/tháng, với độ lệch chuẩn $s = 40$. Lời báo cáo của giám đốc có tin cậy được không, với mức có ý nghĩa là $\alpha = 5\%$.

B 6.2. Trong thập niên 80, trọng lượng trung bình của thanh niên là 48kg. Nay để xác định lại trọng lượng ấy, người ta chọn ngẫu nhiên 100 thanh niên đo trọng lượng trung bình là 50kg và phương sai mẫu $s^2 = (10\text{kg})^2$. Thủ xem trọng lượng thanh niên hiện nay phải chăng có thay đổi, với mức có ý nghĩa là 1%?

B 6.3. Một cửa hàng thực phẩm nhận thấy thời gian vừa qua trung bình một khách hàng mua 250 ngàn đồng thực phẩm trong ngày. Nay cửa hàng chọn ngẫu nhiên 15 khách hàng thấy trung bình một khách hàng mua 240 ngàn đồng trong ngày và phương sai mẫu là $s^2 = (20 \text{ ngàn đồng})^2$. Với mức ý nghĩa là 5%, kiểm định xem có phải sức mua của khách hàng hiện nay thực sự giảm sút hay không. Biết rằng sức mua của khách hàng có phân phối chuẩn.

B 6.4. Đối với người Việt Nam, lượng huyết sắc tố trung bình là 138.3 g/l. Khám cho 80 công nhân ở nhà máy có tiếp xúc hóa chất, thấy huyết sắc tố trung bình $\bar{x} = 120 \text{ g/l}$; $s = 15 \text{ g/l}$. Từ kết quả trên, có thể kết luận lượng huyết sắc tố trung bình của công nhân nhà máy hóa chất này thấp hơn mức chung hay không? Kết luận với $\alpha = 0.05$.

B 6.5. Trong điều kiện chăn nuôi bình thường, lượng sữa trung bình của 1 con bò là 14 kg/ngày. Nghi ngờ điều kiện chăn nuôi kém đi làm cho lượng sữa giảm xuống, người ta điều tra ngẫu nhiên 25 con và tính được lượng sữa trung bình của 1 con trong 1 ngày là 12.5 và độ lệch chuẩn $s = 2.5$. Với mức ý nghĩa $\alpha = 0.05$, hãy kết luận điều nghi ngờ nói trên. Giả thiết lượng sữa bò là 1 biến ngẫu nhiên chuẩn.

B 6.6. Tiền lương trung bình của công nhân trước đây là 400 ngàn đ/tháng. Để xét xem tiền lương hiện nay so với mức trước đây thế nào, người ta điều tra 100 công nhân và tính được $\bar{x} = 404.8 \text{ ngàn đ/tháng}$ và $s = 20 \text{ ngàn đ/tháng}$. Với $\alpha = 1\%$

1. Nếu lập giả thiết 2 phía và giả thiết 1 phía thì kết quả kiểm định như thế nào?
2. Giống câu (1), với $\bar{x} = 406 \text{ ngàn đ/tháng}$ và $s = 20 \text{ ngàn đ/tháng}$.

B 6.7. Một máy đóng gói các sản phẩm có khối lượng 1 kg. Nghi ngờ máy hoạt động không bình thường, người ta chọn ra một mẫu ngẫu nhiên gồm 100 sản phẩm thì thấy như sau:

Khối lượng	0.95	0.97	0.99	1.01	1.03	1.05
Số gói	9	31	40	15	3	2

6. Kiểm định thống kê

Với mức ý nghĩa 0.05, hãy kết luận về nghi ngờ trên.

B 6.8. Trọng lượng trung bình khi xuất chuồng ở một trại chăn nuôi trước là 3.3 kg/con. Năm nay người ta sử dụng một loại thức ăn mới, cân thử 15 con khi xuất chuồng ta được các số liệu như sau:

$3.25, 2.50, 4.00, 3.75, 3.80, 3.90, 4.02, 3.60, 3.80, 3.20, 3.82, 3.40, 3.75, 4.00, 3.50$

Giả thiết trọng lượng gà là đại lượng ngẫu nhiên phân phối theo quy luật chuẩn.

1. Với mức ý nghĩa $\alpha = 0.05$. Hãy cho kết luận về tác dụng của loại thức ăn này?
2. Nếu trại chăn nuôi báo cáo trọng lượng trung bình khi xuất chuồng là 3.5 kg/con thì có chấp nhận được không? ($\alpha = 0.05$).

B 6.9. Đo cholesterol (đơn vị mg%) cho một nhóm người, ta ghi nhận lại được

Chol.	150–160	160–170	170–180	180–190	190–200	200–210
Số người	3	9	11	3	2	1

Cho rằng độ cholesterol tuân theo phân phối chuẩn.

1. Tính trung bình mẫu \bar{x} và phương sai mẫu s^2 .
2. Tìm khoảng ước lượng cho trung bình cholesterol trong dân số ở độ tin cậy 0.95.
3. Có tài liệu cho biết lượng cholesterol trung bình là $\mu_0 = 175$ mg%. Giá trị này có phù hợp với mẫu quan sát không? (kết luận với $\alpha = 0.05$).

B 6.10. Quan sát số hoa hồng bán ra trong một ngày của một cửa hàng bán hoa sau một thời gian, người ta ghi được số liệu sau:

Số hoa hồng(đoá)	12	13	15	16	17	18	19
Số ngày	3	2	7	7	3	2	1

Giả thiết rằng số hoa bán ra trong ngày có phân phối chuẩn.

1. Tính trung bình mẫu \bar{x} và phương sai mẫu s^2 .
2. Sau khi tính toán, ông chủ cửa hàng nói rằng nếu trung bình một ngày không bán được 15 đoá hoa thì chẳng thà đóng cửa còn hơn. Dựa vào số liệu trên, anh (chị) hãy kết luận giúp ông chủ cửa hàng xem có nên tiếp tục bán hay không ở mức ý nghĩa $\alpha = 0.05$.
3. Giả sử những ngày bán được từ 13 đến 17 đoá hồng là những ngày “bình thường”. Hãy ước lượng tỉ lệ của những ngày bình thường của cửa hàng ở độ tin cậy 90%.

B 6.11. Một xí nghiệp đúc một số rất lớn các sản phẩm bằng thép với số khuyết tật trung bình ở mỗi sản phẩm là 3. Người ta cải tiến cách sản xuất và kiểm tra 36 sản phẩm. Kết quả như sau:

6. Kiểm định thống kê

Số khuyết tật trên sản phẩm	0	1	2	3	4	5	6
Số sản phẩm tương ứng	7	4	5	7	6	6	1

Giả sử số khuyết tật của các sản phẩm có phân phối chuẩn.

- Hãy ước lượng số khuyết tật trung bình ở mỗi sản phẩm sau khi cải tiến, với độ tin cậy 90 %.
- Hãy cho kết luận về hiệu quả của việc cải tiến sản xuất ở mức ý nghĩa 0.05.

B 6.12. Đánh giá tác dụng của một chế độ ăn bồi dưỡng mà dấu hiệu quan sát là số hồng cầu. Người ta đếm số hồng cầu của 20 người trước và sau khi ăn bồi dưỡng:

x_i	32	40	38	42	41	35	36	47	50	30
y_i	40	45	42	50	52	43	48	45	55	34

x_i	38	45	43	36	50	38	42	41	45	44
y_i	32	54	58	30	60	35	50	48	40	50

Với mức ý nghĩa $\alpha = 0.05$, có thể kết luận gì về tác dụng của chế độ ăn bồi dưỡng này?

B 6.13. Giả sử ta muốn xác định xem hiệu quả của chế độ ăn kiêng đối với việc giảm trọng lượng như thế nào. 20 người quá béo đã thực hiện chế độ ăn kiêng. Trọng lượng của từng người trước khi ăn kiêng (X kg) và sau khi ăn kiêng (Y kg) được cho như sau:

X	80	78	85	70	90	78	92	88	75	75
Y	75	77	80	70	84	74	85	82	80	65

X	63	72	89	76	77	71	83	78	82	90
Y	62	71	83	72	82	71	79	76	83	81

Kiểm tra xem chế độ ăn kiêng có tác dụng làm thay đổi trọng lượng hay không ($\alpha = 0.05$).

So sánh hai kì vọng

B 6.14. Một nhà phát triển sản phẩm quan tâm đến việc giảm thời gian khô của sơn. Vì vậy hai công thức sơn được đếm thử nghiệm. Công thức 1 là công thức có các thành phần chuẩn và công thức 2 có thêm một thành phần làm khô mới được cho rằng sẽ làm giảm thời gian khô của sơn. Từ các thí nghiệm người ta thấy rằng $\sigma_1 = \sigma_2 = 8$ phút. 10 đồ vật được sơn với công thức 1 và 10 đồ vật khác được sơn với công thức 2. Thời gian khô trung bình của từng mẫu là $\bar{x}_1 = 121$ phút và $\bar{x}_2 = 112$ phút. Nhà phát triển sản phẩm có thể rút ra kết luận gì về ảnh hưởng của thành phần làm khô mới? Với mức ý nghĩa 5%.

6. Kiểm định thống kê

B 6.15. Tốc độ cháy của hai loại chất nổ lỏng được dùng làm nhiên liệu trong tàu vũ trụ được nghiên cứu. Người ta biết rằng độ lệch chuẩn của tốc độ cháy của hai loại nhiên liệu bằng nhau và bằng 3 cm/s . Hai mẫu ngẫu nhiên kích thước $n_1 = 20$ và $n_2 = 20$ được thử nghiệm; trung bình mẫu tốc độ cháy là $\bar{x}_1 = 18 \text{ cm/s}$ và $\bar{x}_2 = 24 \text{ cm/s}$. Với mức ý nghĩa $\alpha = 0.05$ hãy kiểm định giả thuyết hai loại chất nổ lỏng này có cùng tốc độ đốt cháy.

B 6.16. Theo dõi giá cổ phiếu của 2 công ty A và B trong vòng 31 ngày người ta tính được các giá trị sau

	\bar{x}	s
Công ty A	37.58	1.50
Công ty B	38.24	2.20

Giả thiết rằng giá cổ phiếu của hai công ty A và B là hai biến ngẫu nhiên phân phối theo quy luật chuẩn. Hãy cho biết ý nghĩa kì vọng của các biến ngẫu nhiên nói trên? Hãy cho biết có sự khác biệt thực sự về giá cổ phiếu trung bình của hai công ty A và B không? Với mức ý nghĩa $\alpha = 5\%$

B 6.17. Hàm lượng đường trong máu của công nhân sau 5 giờ làm việc với máy siêu cao tần đã đo được ở hai thời điểm trước và sau 5 giờ làm việc. Ta có kết quả sau:

- Trước: $n_1 = 50$, $\bar{x} = 60 \text{ mg\%}$, $s_x = 7$
- Sau: $n_2 = 40$, $\bar{y} = 52 \text{ mg\%}$, $s_y = 9.2$

Với mức ý nghĩa $\alpha = 0.05$, có thể khẳng định hàm lượng đường trong máu sau 5 giờ làm việc đã giảm đi hay không?

B 6.18. Trồng cùng một giống lúa trên hai thửa ruộng như nhau và bón hai loại phân khác nhau. Đến ngày thu hoạch ta có kết quả như sau:

- Thửa thứ nhất lấy mẫu 1000 bông lúa thấy số hạt trung bình của mỗi bông là $\bar{x} = 70$ hạt và $s_x = 10$.
- Thửa thứ hai lấy mẫu 500 bông thấy số hạt trung bình mỗi bông là $\bar{y} = 72$ hạt và $s_y = 20$.

Hỏi sự khác nhau giữa X và Y là ngẫu nhiên hay bản chất, với $\alpha = 0.05$?

B 6.19. Để so sánh trọng lượng trung bình của trẻ sơ sinh ở thành thị và nông thôn, người ta thử cân trọng lượng của 10000 cháu và thu được kết quả sau đây:

Vùng	Số cháu được cân	Trọng lượng trung bình	Độ lệch chuẩn mẫu
Nông thôn	8000	3.0 kg	0.3 kg
Thành thị	2000	3.2 kg	0.2 kg

Với mức ý nghĩa $\alpha = 0.05$ có thể coi trọng lượng trung bình của trẻ sơ sinh ở thành thị cao hơn ở nông thôn hay không? (Giả thiết trọng lượng trẻ sơ sinh là biến ngẫu nhiên chuẩn).

6. Kiểm định thống kê

B 6.20. Để so sánh năng lực học toán và vật lý của học sinh, người ta kiểm tra ngẫu nhiên 8 em bằng hai bài toán và vật lý. Kết quả cho bởi bảng dưới đây (X là điểm toán, Y là điểm lý):

X	15	20	16	22	24	18	20	14
Y	15	22	14	25	19	20	24	16

Giả sử X và Y đều có phân phối chuẩn. Hãy so sánh điểm trung bình giữa X và Y , mức ý nghĩa 5%.

B 6.21. Hai máy được sử dụng để rót nước vào các bình. Người ta lấy mẫu ngẫu nhiên 10 bình do máy thứ nhất và 10 bình do máy thứ hai thì được kết quả sau:

Máy 1	16.03	16.01	16.04	15.96	16.05	15.98	16.05	16.02	16.02	15.99
Máy 2	16.02	16.03	15.97	16.04	15.96	16.02	16.01	16.01	15.99	16.00

Với mức ý nghĩa $\alpha = 0.05$ có thể nói rằng hai máy rót nước vào bình như nhau không?

B 6.22. Để nghiên cứu ảnh hưởng của một loại thuốc, người ta cho 10 bệnh nhân uống thuốc. Lần khác họ cũng cho bệnh nhân uống thuốc nhưng là thuốc giả. Kết quả thí nghiệm thu được như sau:

Bệnh nhân	1	2	3	4	5	6	7	8	9	10
Số giờ ngủ có thuốc	6.1	7.0	8.2	7.6	6.5	8.4	6.9	6.7	7.4	5.8
Số giờ ngủ với thuốc giả	5.2	7.9	3.9	4.7	5.3	5.4	4.2	6.1	3.8	6.3

Giả sử số giờ ngủ của bệnh nhân tuân theo phân phối chuẩn. Với mức ý nghĩa 5%, hãy kết luận về ảnh hưởng của loại thuốc trên.

B 6.23. Quan sát sức nặng của bé trai (X) và bé gái (Y) lúc sơ sinh (đơn vị gam), ta có kết quả

Trọng lượng	3000-3200	3200-3400	3400-3600	3600-3800	3800-4000
Số bé trai	1	3	8	10	3
Số bé gái	2	10	10	5	1

- Tính \bar{x} , \bar{y} , s_x^2 , s_y^2 .
- So sánh các kì vọng μ_X , μ_Y (kết luận với $\alpha = 5\%$).
- Nhập hai mẫu lại. Tính trung bình và độ lệch chuẩn của mẫu nhập. Dùng mẫu nhập để ước lượng sức nặng trung bình của trẻ sơ sinh ở độ tin cậy 95%.

6. Kiểm định thống kê

So sánh tỉ lệ với một số cho trước

B 6.24. Trong một vùng dân cư có 18 bé trai và 28 bé gái mắc bệnh B. Hỏi rằng tỷ lệ nhiễm bệnh của bé trai và bé gái có như nhau không? (kết luận với $\alpha = 0.05$ và giả sử rằng số lượng bé trai và bé gái trong vùng tương đương nhau, và rất nhiều).

B 6.25. Một máy sản xuất tự động với tỷ lệ chính phẩm là 98%. Sau một thời gian hoạt động, người ta nghi ngờ tỷ lệ trên đã bị giảm. Kiểm tra ngẫu nhiên 500 sản phẩm thấy có 28 phế phẩm, với $\alpha = 0.05$ hãy kiểm tra xem chất lượng làm việc của máy có còn được như trước hay không?

B 6.26. Đo huyết sắc tố cho 50 công nhân nông trường thấy có 60% ở mức dưới 110 g/l. Số liệu chung của khu vực này là 30% ở mức dưới 110 g/l. Với mức ý nghĩa $\alpha = 0.05$, có thể kết luận công nhân nông trường có tỷ lệ huyết sắc tố dưới 110 g/l cao hơn mức chung hay không?

B 6.27. Theo một nguồn tin thì tỉ lệ hộ dân thích xem dân ca trên Tivi là 80%. Thăm dò 36 hộ dân thấy có 25 hộ thích xem dân ca. Với mức có ý nghĩa là 5%. Kiểm định xem nguồn tin này có đáng tin cậy không?

B 6.28. Một máy sản suất tự động, lúc đầu tỷ lệ sản phẩm loại A là 20%. Sau khi áp dụng một phương pháp cải tiến sản xuất mới, người ta lấy 40 mẫu, mỗi mẫu gồm 10 sản phẩm để kiểm tra. Kết quả kiểm tra cho ở bảng sau:

Số sản phẩm loại A trong mẫu	1	2	3	4	5	6	7	8	9	10
Số mẫu	2	0	4	6	8	10	4	5	1	0

Với mức ý nghĩa 5%. Hãy cho kết luận về phương pháp sản suất này.

B 6.29. Tỷ lệ phế phẩm của một nhà máy trước đây là 5%. Năm nay nhà máy áp dụng một biện pháp kỹ thuật mới. Để nghiên cứu tác dụng của biện pháp kỹ thuật mới, người ta lấy một mẫu gồm 800 sản phẩm để kiểm tra và thấy có 24 phế phẩm.

1. Với $\alpha = 0.01$. Hãy cho kết luận về biện pháp kỹ thuật mới này?
2. Nếu nhà máy báo cáo tỷ lệ phế phẩm sau khi áp dụng biện pháp kỹ thuật mới là 2% thì có chấp nhận được không? ($\alpha = 0.01$).

So sánh hai tỉ lệ

B 6.30. Trong 90 người dùng DDT để ngừa bệnh ngoài da thì có 10 người nhiễm bệnh; trong 100 người không dùng DDT thì có 26 người mắc bệnh. Hỏi rằng DDT có tác dụng ngừa bệnh ngoài da không? (kết luận với $\alpha = 0.05$)

B 6.31. Người ta điều tra 250 người ở xã A thấy có 140 nữ và điều tra 160 người ở xã B thấy có 80 nữ. Hãy so sánh tỉ lệ nữ ở hai xã với mức ý nghĩa 5%.

6. Kiểm định thống kê

B 6.32. Áp dụng hai phương pháp gieo hạt. Theo phương pháp A gieo 180 hạt thì có 150 hạt nảy mầm; theo phương pháp B gieo 256 hạt thì thấy có 160 hạt nảy mầm. Hãy so sánh hiệu quả của hai phương pháp với mức ý nghĩa $\alpha = 5\%$.

B 6.33. Theo dõi trọng lượng của một số trẻ sơ sinh tại một số nhà hộ sinh thành phố và nông thôn, người ta thấy rằng trong số 150 trẻ sơ sinh ở thành phố có 100 cháu nặng hơn 3000 gam, và trong 200 trẻ sơ sinh ở nông thôn có 98 cháu nặng hơn 3000 gam. Từ kết quả đó hãy so sánh tỉ lệ trẻ sơ sinh có trọng lượng trên 3000 gam ở thành phố và nông thôn với mức ý nghĩa 5%.

Phân tích hồi quy tuyến tính đơn

Ứng dụng chính của hồi quy là để dự báo 1 biến phụ thuộc (*dependent variable, response variable*) dựa vào 1 hay nhiều biến độc lập (*independent variables, predictor variables, explanatory variables*)

- **Phân tích khi tăng 1 nhân viên thì lợi nhuận ngân hàng tăng hay giảm bao nhiêu tiền.** Biến phụ thuộc là lợi nhuận ngân hàng, biến độc lập là số lượng nhân viên
- **Tính toán xem khi tăng 1 cây ATM thì lợi nhuận ngân hàng tăng hay giảm bao nhiêu phần trăm.** Biến phụ thuộc là log(lợi nhuận), biến độc lập có thể là số lượng máy ATM, số lượng máy ATM bình phương
- **Dự báo khả năng phát sinh nợ xấu của khách hàng.** Biến phụ thuộc là khả năng phát sinh nợ xấu, biến độc lập ví dụ: tuổi, giới tính, trình độ học vấn ...
- **Dự báo giá cổ phiếu của ngân hàng tại các thời điểm trong tương lai.** Biến phụ thuộc là giá cổ phiếu, biến độc lập có thể có là trễ của biến giá, hoặc 1 số yếu tố như GDP, lạm phát ...

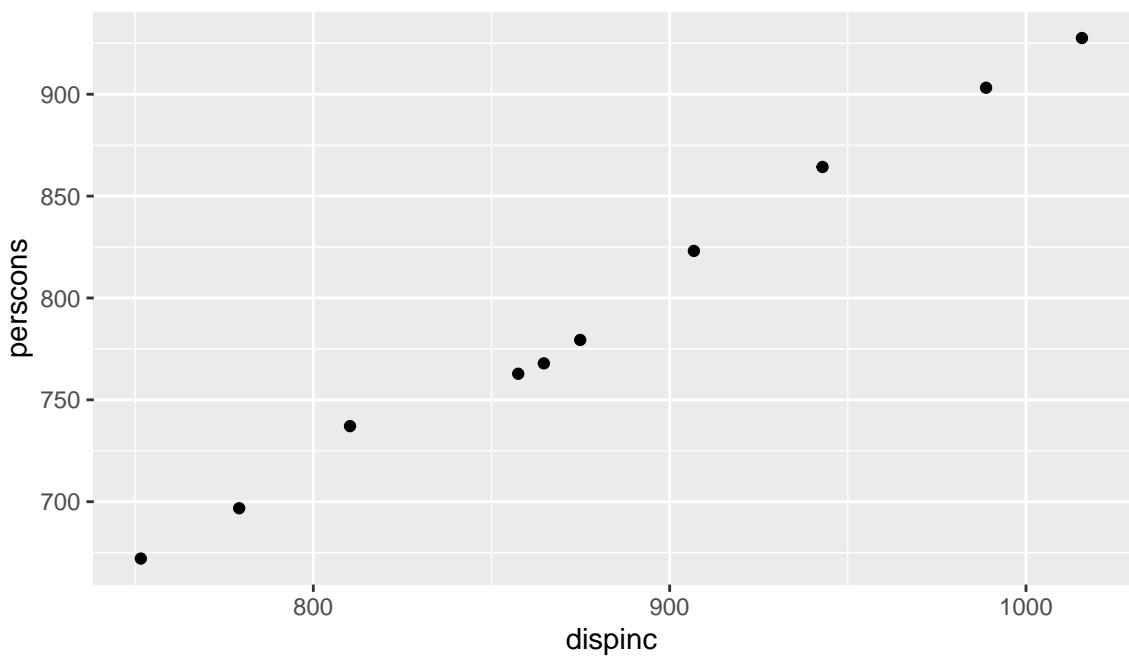
7.1

Hồi quy tuyến tính đơn

Ví dụ. Số liệu về *tiêu dùng trung bình (percons)* và *thu nhập khả dụng (dispinc)* theo giá cố định theo năm 1972 của nền kinh tế Mỹ trong 10 năm 1970 - 1979

7. Phân tích hồi quy tuyến tính đơn

Năm	dispinc (x)	perscons (y)	(ĐVT: tỷ dollars)
1970	751.6	672.1	
1971	779.2	696.8	
1972	810.3	737.1	
1973	864.7	767.9	
1974	857.5	762.8	
1975	874.9	779.4	
1976	906.8	823.1	
1977	942.9	864.3	
1978	988.8	903.2	
1979	1015.7	927.6	



Một cách tổng quát, dạng hàm mô tả tốt nhất khuynh hướng tiêu dùng theo thu nhập có dạng tuyến tính

$$y = w_0 + w_1 x \quad (7.1)$$

Mặc dù dữ liệu xem ra thể hiện khá tốt qui luật tuyến tính nêu ở trên nhưng rõ ràng mối quan hệ có tính xác định đó là không đủ để mô tả thực tiễn, vì còn rất nhiều yếu tố khác ảnh hưởng đến tiêu dùng (giới tính, tuổi tác, tâm lý,...).

Nói chung, chúng ta không có tham vọng đưa hết tất cả mọi yếu tố ảnh hưởng tới tiêu dùng vào mô hình, mà chỉ những yếu tố quan trọng, thiết yếu nhất.

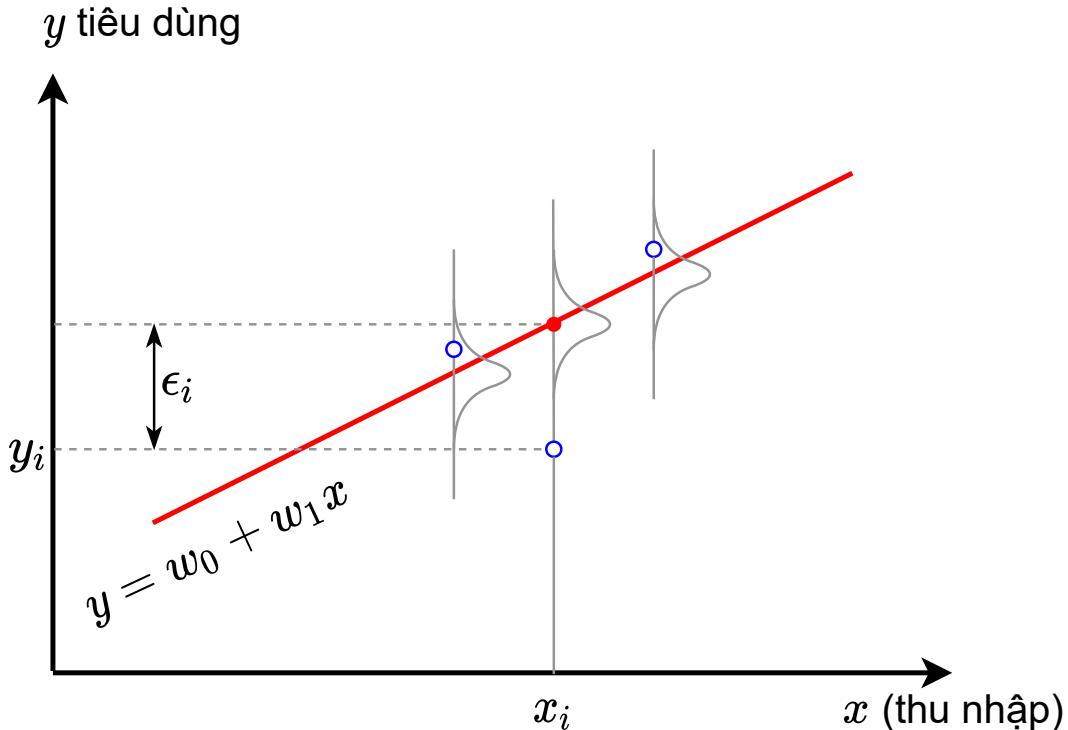
Vì vậy, để có thể biểu diễn qui luật tiêu dùng trên thế giới thực, ta cần đưa thêm vào mô hình tuyến tính (6.1) một thành phần khác nữa, mang tính ngẫu nhiên, thể hiện sự tác động tổng hợp của các nhân tố nhỏ, không ổn định, tới tiêu dùng. Tức là, những yếu tố làm cho

7. Phân tích hồi quy tuyến tính đơn

quan sát thật về tiêu dùng và thu nhập bị lệch khỏi xu thế ổn định, tuyến tính nêu trên.

$$y = \mathbf{w}_0 + \mathbf{w}_1 x + \epsilon \quad (7.2)$$

Tức là, ta muốn biểu diễn mối quan hệ thực giữa các cặp dữ liệu quan sát được về thu nhập và tiêu dùng $\{(x_i, y_i)\}, i = 1, \dots, n$



$$y_i = \mathbf{w}_0 + \mathbf{w}_1 x_i + \epsilon_i \quad (7.3)$$

Trong đó,

- (x_i, y_i) là thu nhập và tiêu dùng thực tế của mẫu quan sát thứ n
- $\mathbf{w}_0 + x_i \mathbf{w}_1$ là quy luật xác định hay mô hình (*deterministic part*), $\mathbf{w}_0, \mathbf{w}_1$ được gọi là các tham số, hệ số hoặc trọng số
- ϵ_i là yếu tố nhiễu (*random part*), còn được gọi là phần dư (*residual*)

Lưu ý, phương trình (7.2) được gọi là hàm hồi quy tổng thể (*Population Regression Function - PRF*) và phương trình (7.3) được gọi là hàm hồi quy mẫu (*Sample Regression Function - SRF*)

7.2 Ước lượng qui luật bằng OLS

Bài toán. Tìm một quy luật phù hợp tốt nhất với tập dữ liệu mẫu quan sát được

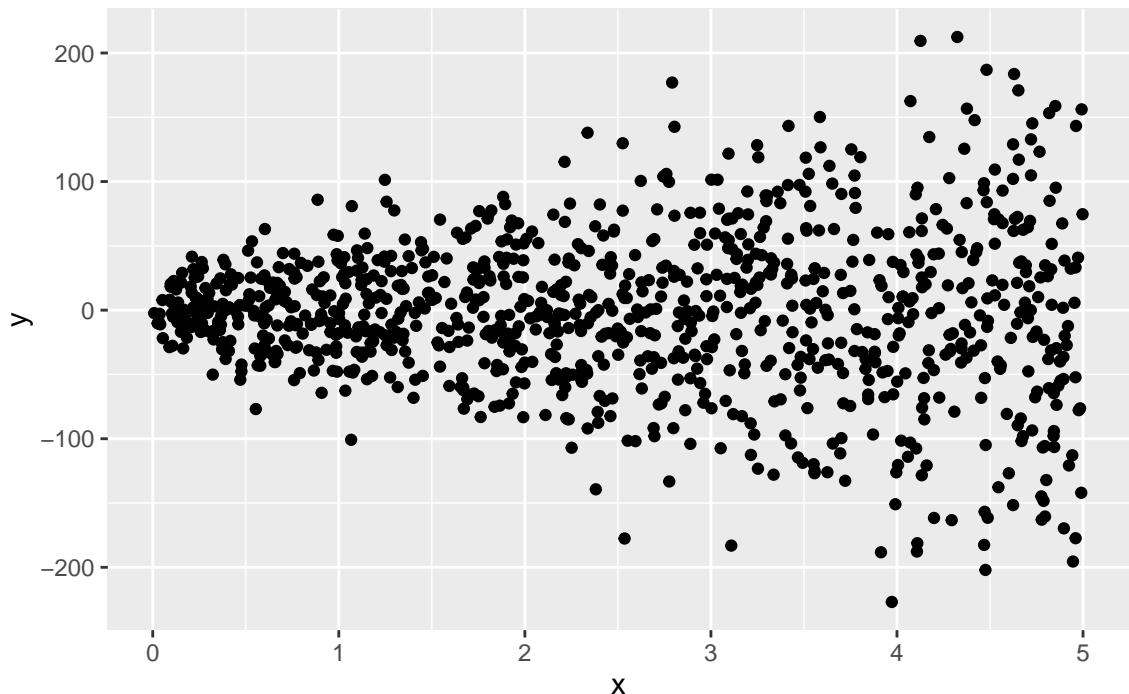
7. Phân tích hồi quy tuyến tính đơn

Các giả định cần thiết để có được một quy luật tốt

Giả định Gauss-Markov

1. $\{x_i, y_i\}$ là tập mẫu ngẫu nhiên và độc lập.
 2. Sai số ϵ_i có trung bình là 0 và phương sai giống nhau (*homoskedasticity*); nghĩa là không có phương sai thay đổi (*heteroskedasticity*).
- $$\forall i, \mathbb{E}(\epsilon_i) = 0 \text{ và } \text{Var}(\epsilon_i) = \sigma^2 \quad (7.4)$$
3. Không có tương quan giữa ϵ_i và ϵ_j ($i \neq j$).
 4. Không có tương quan giữa x_i và ϵ_i .
 5. Sai số ϵ_i tuân theo luật phân phối chuẩn $\mathcal{N}(0, \sigma^2)$ (không bắt buộc)

Heteroskedasticity



Chúng ta định nghĩa lỗi của quy luật là tổng bình phương của tất cả các giá trị lỗi RSS

$$RSS = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \mathbf{w}_0 - \mathbf{w}_1 x_i)^2 \quad (7.5)$$

Một cách tự nhiên, chúng ta có thể xem một quy luật tốt nhất là quy luật cực tiểu hóa giá trị RSS (phương pháp bình phương nhỏ nhất - *ordinary least square*). Vậy bài toán đặt ra là

7. Phân tích hồi quy tuyến tính đơn

tìm (ước lượng) tham số w_0, w_1 sao cho cực tiểu hóa RSS

$$\arg \min_{w_0, w_1} RSS \quad (7.6)$$

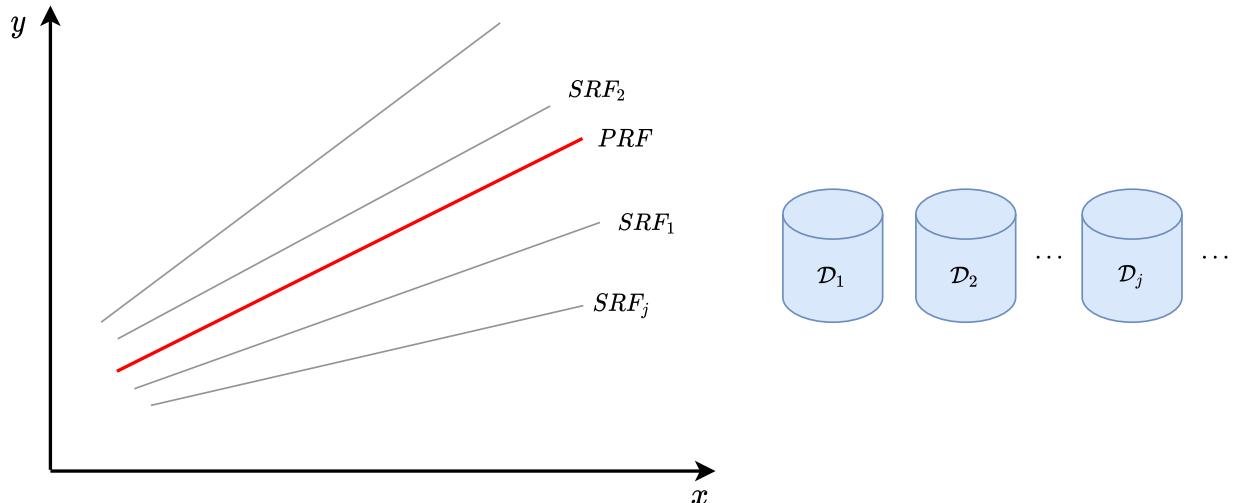
Kết quả ước lượng ta có

$$\begin{aligned}\hat{w}_1 &= \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x}^2 - \bar{x}^2} \\ \hat{w}_0 &= \bar{y} - \hat{w}_1\bar{x}\end{aligned} \quad (7.7)$$

Chúng ta cũng có thể tính \hat{w}_1 bằng

$$\hat{w}_1 = \sum \left[\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right] y_i \quad (7.8)$$

Lưu ý, ước lượng từ các mẫu khác nhau cho kết quả khác nhau dù cùng một tổng thể, do đó có thể khác biệt với hệ số hồi quy tổng thể



Một số các đại lượng khác

- Tổng bình phương toàn phần TSS

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (7.9)$$

- Tổng bình phương hồi quy ESS

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\hat{w}_0 + x_i \hat{w}_1 - \bar{y})^2 \quad (7.10)$$

Lưu ý, ta có

$$TSS = RSS + ESS \quad (7.11)$$

7. Phân tích hồi quy tuyến tính đơn

- Hệ số xác định R^2

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} \quad (7.12)$$

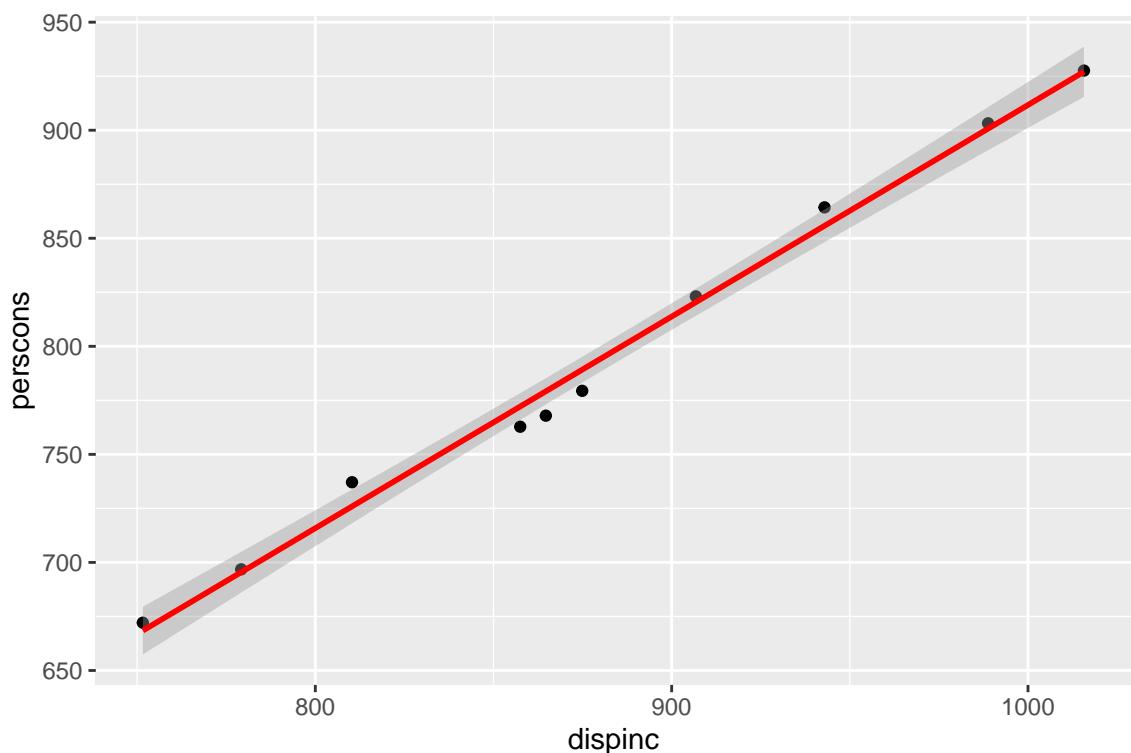
- Hệ số xác định hiệu chỉnh R_{adj}^2

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1} = 1 - \frac{RSS/(n - k - 1)}{TSS/(n - 1)} \quad (7.13)$$

với k là số lượng biến số độc lập trong phương trình hồi quy tuyến tính bội tổng quát

$$y = \mathbf{w}_0 + x_1 \mathbf{w}_1 + \dots + x_k \mathbf{w}_k + \epsilon \quad (7.14)$$

Lưu ý, ta có $R^2, R_{adj}^2 \in [0, 1]$



7.3

Tính chất của hàm hồi quy mẫu

Tính chất.

1. Tổng các phần dư bằng 0

$$\sum_{i=1}^n \epsilon_i = 0$$

2. Hiệp phương sai mẫu giữa biến độc lập và phần dư bằng 0

$$Cov(x_i, \epsilon_i) = 0$$

7. Phân tích hồi quy tuyến tính đơn

3. Đường hồi quy mẫu luôn đi qua giá trị trung bình mẫu (\bar{x}, \bar{y})

4. Trung bình của giá trị ước lượng của biến phụ thuộc bằng trung bình mẫu của nó

- Nếu giả định Gauss-Markov là đúng thì ước lượng bằng OLS là không chêch

$$\begin{aligned}\mathbb{E}(\hat{w}_0) &= \mathbb{E}(\hat{w}_0 | \mathbf{x}) = w_0 \\ \mathbb{E}(\hat{w}_1) &= \mathbb{E}(\hat{w}_1 | \mathbf{x}) = w_1\end{aligned}\quad (7.15)$$

và độ chính xác của ước lượng

$$\begin{aligned}\text{Var}(\hat{w}_0 | \mathbf{x}) &= \sigma^2 \left[\frac{1}{n} \frac{\sum x_i^2}{\sum(x_i - \bar{x})^2} \right] \\ \text{Var}(\hat{w}_1 | \mathbf{x}) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\end{aligned}\quad (7.16)$$

với $\mathbf{x} = \{x_1, \dots, x_n\}$

- Các ước lượng bằng OLS cũng là những ước lượng vững và hiệu quả

Ma trận hiệp phương sai của hệ số ước lượng

Chúng ta có công thức tổng quát

$$\begin{pmatrix} \text{Var}(\hat{w}_0 | \mathbf{x}) & \text{cov}(\hat{w}_0, \hat{w}_1 | \mathbf{x}) \\ \text{cov}(\hat{w}_1, \hat{w}_0 | \mathbf{x}) & \text{Var}(\hat{w}_1 | \mathbf{x}) \end{pmatrix} = \sigma^2 \begin{pmatrix} \frac{1}{n} \frac{\sum x_i^2}{\sum(x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum(x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum(x_i - \bar{x})^2} & \frac{1}{\sum(x_i - \bar{x})^2} \end{pmatrix} \quad (7.17)$$

Ước lượng của phương sai sai số ngẫu nhiên

- Ước lượng không chêch của σ^2 là

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \epsilon_i^2 \quad (7.18)$$

với

$$\mathbb{E}(\hat{\sigma}^2 | \mathbf{x}) = \sigma^2 \quad (7.19)$$

- Ước lượng $\hat{\sigma}^2$ được gọi là sai số chuẩn của hàm hồi quy (*standard error of regression*)
- Sai số chuẩn của hệ số ước lượng

$$\begin{aligned}se^2(\hat{w}_0) &= \hat{\sigma}^2 \left[\frac{1}{n} \frac{\sum x_i^2}{\sum(x_i - \bar{x})^2} \right] \\ se^2(\hat{w}_1) &= \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}\end{aligned}\quad (7.20)$$

7.4**Vấn đề về thay đổi đơn vị của biến**

Xét hàm hồi quy mẫu

$$y_i = \hat{w}_0 + x_i \hat{w}_1 + \epsilon_i$$

- Nếu gia tăng đơn vị của X lên k lần (ví dụ: 10 lần từ kg lên yên) và giữ nguyên đơn vị của Y thì giá trị của X giảm đi k lần và hệ số góc \hat{w}_1 tăng lên k lần còn hệ số chặn \hat{w}_0 không đổi.
- Nếu gia tăng đơn vị của Y lên k lần (ví dụ: 1000 lần từ triệu lên tỷ) và giữ nguyên đơn vị của X thì giá trị của Y giảm đi k lần và cả hệ số góc \hat{w}_1 và hệ số chặn \hat{w}_0 giảm k lần.
- Thay đổi đơn vị không làm thay đổi R^2 và ý nghĩa kinh tế của các hệ số

7.5**Vấn đề tương quan**

Một trong các giả thiết của mô hình hồi quy tuyến tính là không có (tự) tương quan giữa các sai số ngẫu nhiên ϵ_i . Về bản chất thì giả thiết này muốn ngụ ý rằng quan sát của biến phụ thuộc ở thời điểm này sẽ không có quan hệ với quan sát của biến phụ thuộc ở thời điểm khác. Nhưng trong thực tế giả thiết này có thể vi phạm.

Hậu quả:

- Các ước lượng OLS vẫn là các ước lượng không chêch, nhưng không còn hiệu quả nữa.
- Phương sai ước lượng của các ước lượng OLS thường là chêch. Khi tính phương sai và sai số chuẩn của các ước lượng OLS thường cho những giá trị thấp hơn các giá trị thực và do đó làm cho giá trị của t lớn, dẫn đến kết luận sai khi kiểm định. Do đó kiểm định t và F không còn tin cậy nữa.
- Các giá trị dự báo không đáng tin cậy (không hiệu quả).

7.5.1. Nguyên nhân của tự tương quan**Nguyên nhân khách quan**

- **Quán tính:** Nét nổi bật của hầu hết các chuỗi thời gian trong kinh tế là quán tính. Chúng ta đều biết các chuỗi thời gian như: Tổng sản phẩm, chỉ số giá, thất nghiệp... mang tính chu kỳ. Chẳng hạn ở giai đoạn đầu của thời kỳ khôi phục kinh tế, tổng sản phẩm có xu hướng đi lên, do đó giá trị của chuỗi ở điểm sau thường cao hơn điểm trước và khi hồi quy chuỗi thời gian, các quan sát kế tiếp có nhiều khả năng phụ thuộc vào nhau.

7. Phân tích hồi quy tuyến tính đơn

- **Hiện tượng mạng nhện:** Người ta thấy rằng việc cung cấp nhiều mặt hàng nông sản biểu hiện hiện tượng “mạng nhện”, trong đó lượng cung phản ứng lại với giá cả trễ một khoảng thời gian, vì các quyết định cung cần phải mất một khoảng thời gian để thực hiện.
- **Các độ trễ:** Trong phân tích chuỗi thời gian, chúng ta có thể gặp hiện tượng biến phụ thuộc ở thời kỳ t phụ thuộc vào chính biến đó ở thời kỳ $t - 1$ và các biến khác.

Nguyên nhân chủ quan

- **Xử lý số liệu:** Trong phân tích thực nghiệm, số liệu thô thường được xử lý. Chẳng hạn trong hồi quy chuỗi thời gian gắn với các số liệu quý, các số liệu này thường được suy ra từ số liệu tháng bằng cách cộng 3 quan sát theo tháng rồi chia cho 3. Việc lấy trung bình làm trên các số liệu và làm giảm sự dao động trong số liệu tháng. Do vậy đồ thị số liệu quý trơn tru hơn nhiều so với số liệu tháng. Chính sự làm trơn này có thể dẫn tới sai số hệ thống trong các sai số ngẫu nhiên và gây ra sự tương quan.
- **Sai lệch do lập mô hình:** (a) không đưa đủ các biến ảnh hưởng cơ bản vào mô hình (b) dạng hàm sai

7.5.2. Phát hiện hiện tượng tự tương quan

1. Kiểm định Durbin-Watson: Kiểm tra hiện tượng tự tương quan bậc 1, nghĩa là liệu có mối quan hệ hồi quy sau hay không?

$$\epsilon_t = \rho\epsilon_{t-1} + \eta_t, \quad |\rho| < 1 \quad (7.21)$$

2. Kiểm định Breusch-Godfrey: Kiểm tra hiện tượng tự tương quan bậc p , nghĩa là liệu có mối quan hệ hồi quy sau hay không?

$$\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + \dots + \rho_p\epsilon_{t-p} + \eta_t, \quad |\rho| < 1 \quad (7.22)$$

7.5.3. Các biện pháp khắc phục

Ý tưởng chung của các biện pháp khắc phục là biến đổi các biến để chuyển thành mô hình không có tự tương quan. Chúng ta chỉ xem xét trường hợp biết tự tương quan bậc 1

1. Biết ρ

Phương trình hồi quy mẫu

$$y_i = \textcolor{red}{w_0} + \textcolor{red}{w_1}x_i + \epsilon_i \quad (7.23)$$

có sai số ngẫu nhiên tương quan bậc 1

$$\epsilon_i = \rho\epsilon_{i-1} + \eta_i, \quad |\rho| < 1$$

7. Phân tích hồi quy tuyến tính đơn

Sử dụng biến đổi sai phân

$$\begin{aligned}y_i^* &= y_i - \rho y_{i-1} \\x_i^* &= x_i - \rho x_{i-1} \\\epsilon_i^* &= \epsilon_i - \rho \epsilon_{i-1} = \eta_i\end{aligned}\tag{7.24}$$

phương trình mẫu sẽ trở thành

$$y_i^* = \textcolor{red}{w}_0^* + \textcolor{red}{w}_1^* x_i^* + \epsilon_i^*\tag{7.25}$$

Vì ϵ_i^* thỏa mãn các giả thiết của OLS nên các ước lượng là không chêch và hiệu quả.

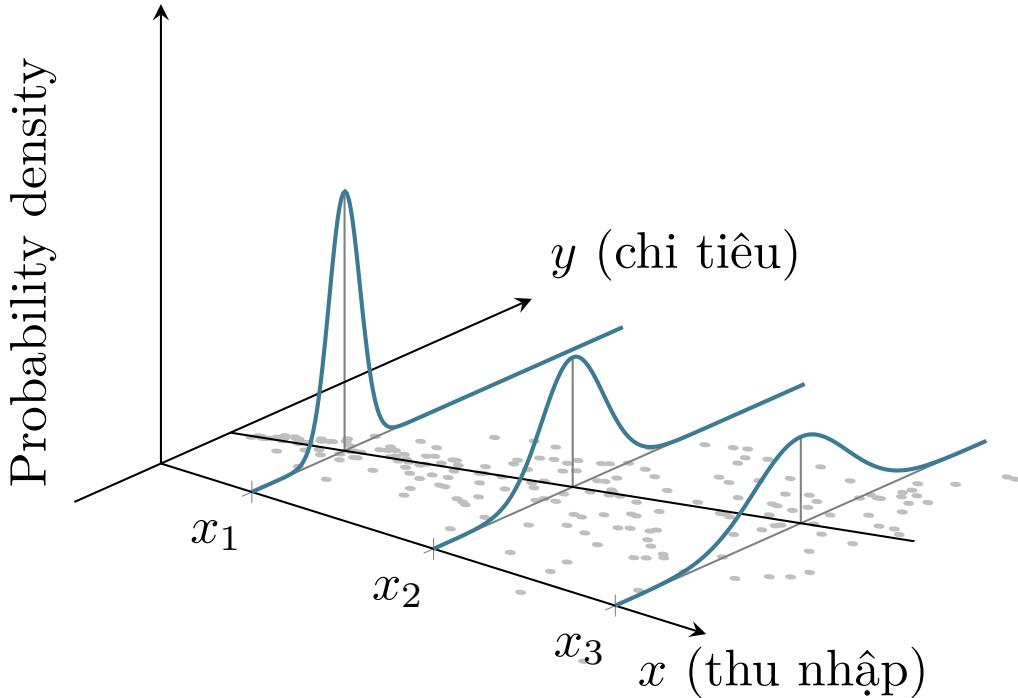
2. Chưa biết ρ : ước lượng giá trị ρ bằng giá trị thống kê Durbin-Watson

7.6

Vấn đề phương sai thay đổi

Giả sử như chúng ta tiến hành điều tra một mẫu ngẫu nhiên các hộ gia đình và thu được thông tin về *chi phí tiêu dùng* của từng hộ gia đình và *thu nhập* của họ trong một năm cho trước. Những hộ gia đình với mức thu nhập thấp không có nhiều linh động trong chi tiêu. Phần lớn thu nhập sẽ tập trung vào các nhu cầu căn bản chẳng hạn như thức ăn, chỗ ở, quần áo, và đi lại. Tuy nhiên, những gia đình giàu có có sự linh động rất lớn trong chi tiêu. Một vài gia đình là những người tiêu dùng lớn; những người khác có thể là những người tiết kiệm nhiều và đầu tư nhiều vào bất động sản, thị trường chứng khoán. Điều này hàm ý rằng tiêu dùng thực có thể khác nhiều so với mức thu nhập trung bình. Hay nói cách khác, rất có khả năng những hộ gia đình có thu nhập cao có mức độ phân tán xung quanh giá trị tiêu dùng trung bình lớn hơn những hộ gia đình có thu nhập thấp. Trong trường hợp như thế, biểu đồ phân tán giữa tiêu dùng và thu nhập sẽ chỉ ra những điểm của mẫu gần với đường hồi qui hơn cho những hộ gia đình thu nhập thấp nhưng những điểm phân tán rộng hơn cho những hộ gia đình thu nhập cao.

7. Phân tích hồi quy tuyến tính đơn



Hậu quả: Các ước lượng tham số vẫn không chênh $\mathbb{E}(\hat{w}_i) = w_i$. Tuy nhiên, các ước lượng này không còn hiệu quả nữa; nghĩa là $\text{Var}(\hat{w}_i | \mathbf{x})$ không còn là nhỏ nhất nữa.

7.6.1. Giải pháp khắc phục WLS

Phương pháp bình phương trọng số nhỏ nhất (*weighted least square - WLS*) sẽ gắn mỗi quan sát (x_i, y_i) một trọng số k_i không âm. Như đã biết phương pháp OLS sẽ cực tiểu tổng bình phương các phần dư

$$\sum_{i=1}^n \epsilon_i^2,$$

còn phương pháp WLS sẽ cực tiểu tổng bình phương các phần dư có trọng số

$$\sum_{i=1}^n k_i \epsilon_i^2 = \sum_{i=1}^n k_i (y_i - w_0 - x_i w_1)^2 \quad (7.26)$$

nghĩa là, chúng ta sẽ tìm các tham số sao cho biểu thức (7.20) là nhỏ nhất.

Kết quả ước lượng ta có

$$\begin{aligned} \hat{w}_1 &= \frac{(\sum k_i)(\sum k_i x_i y_i) - (\sum k_i x_i)(\sum k_i y_i)}{(\sum k_i)(\sum k_i x_i^2) - (\sum k_i x_i)^2}, \\ \hat{w}_0 &= \bar{y} - \hat{w}_1 \bar{x} \end{aligned} \quad (7.27)$$

trong đó,

$$\bar{x} = \frac{\sum k_i x_i}{\sum k_i} \quad \text{và} \quad \bar{y} = \frac{\sum k_i y_i}{\sum k_i} \quad (7.28)$$

7. Phân tích hồi quy tuyến tính đơn

Nếu ta chọn $k_i = \frac{1}{\sigma_i^2}$, thì các ước lượng của WLS sẽ là không chêch và hiệu quả. Tuy nhiên, trong các bài toán thực tế chúng ta thường không biết được σ_i . Vậy chúng ta có thể giả định là

$$\sigma_i = x_i \quad \text{hoặc} \quad \sigma_i = \sqrt{x_i} \quad (7.29)$$

7.7 Kiểm định mô hình ước lượng

1. Kiểm định tham số

- | | |
|--|--|
| <ul style="list-style-type: none"> Giả thuyết $H_0 : w_i = 0$ Đối thuyết $H_a : w_i \neq 0$ Các mức ý nghĩa $\alpha = 0.001, 0.01, 0.05$ | <ul style="list-style-type: none"> Tính t Tính p-value tương ứng $\Pr(> t)$ |
|--|--|

2. Kiểm định độ tốt của mô hình quy luật (*good of fitness*)

- | | |
|---|--|
| <ul style="list-style-type: none"> Giả thuyết $H_0 : w_i = 0, \forall i$ Đối thuyết $H_a : w_i \neq 0, \exists i$ | <ul style="list-style-type: none"> Tính f Tính p-value tương ứng |
|---|--|

Khả năng giải thích của mô hình

$$R^2 \quad \text{hoặc} \quad R_{adj}^2$$

- Kết quả phân tích hồi quy tuyến tính cho ví dụ

Residuals:

Min	1Q	Median	3Q	Max
-11.291	-6.871	1.909	3.418	11.181

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	-67.58065	27.91071	-2.421	0.0418 *							
dispinc	0.97927	0.03161	30.983	1.28e-09 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

Residual standard error: 8.193 on 8 degrees of freedom

Multiple R-squared: 0.9917, Adjusted R-squared: 0.9907

F-statistic: 959.9 on 1 and 8 DF, p-value: 1.28e-09

7. Phân tích hồi quy tuyến tính đơn

Kết luận:

- Biến **dispinc** có ý nghĩa đối với mô hình về mặt thống kê ($p\text{-value} = 1.28e-09$)
- Mô hình phù hợp tốt với dữ liệu quan sát về mặt thống kê ($p\text{-value} = 1.28e-09$)
- Biến **dispinc** có thể giải thích được 99.07% sự thay đổi của biến **perscons**
- Phương trình hồi quy

$$percons = -67.58 + 0.98 \times dispinc$$

7.8

Ứng dụng mô hình hồi qui

Dự báo giá trị trung bình của biến phụ thuộc

Giả sử ta biết giá trị của biến độc lập $X = x$ và ta cần đưa ra dự báo giá trị trung bình của biến phụ thuộc Y , thì ta có:

$$y = \mathbb{E}(Y | X = x) = \color{red}w_0 + w_1x \quad (7.30)$$

Sử dụng hàm hồi quy mău ta có ước lượng điểm cho $y = \mathbb{E}(Y | X = x)$

$$\hat{y} = \color{red}\hat{w}_0 + \hat{w}_1x \quad (7.31)$$

Ước lượng \hat{y} này là ước lượng không chêch

$$\mathbb{E}(\hat{y}) = y, \quad (7.32)$$

và có phuơng sai

$$\text{Var}(\hat{y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \quad (7.33)$$

Do chưa biết σ^2 , nên ta có thể sử dụng ước lượng không chêch $\hat{\sigma}^2$ để tính độ lệch chuẩn cho \hat{y}

$$se^2(\hat{y}) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \quad (7.34)$$

Ước lượng khoảng tin cậy cho giá trị trung bình với mức ý nghĩa α

$$\left[\hat{y} - se(\hat{y}) \times t \left(1 - \frac{\alpha}{2}; n - 2 \right), \hat{y} + se(\hat{y}) \times t \left(1 - \frac{\alpha}{2}; n - 2 \right) \right] \quad (7.35)$$

Kiểm định giá trị của biến phụ thuộc

Giả sử chúng ta cần kiểm định giá trị biến phụ thuộc $y = y_0$ ứng với giá trị biến độc lập $x = x_0$

- Giả thuyết $H_0 : y = y_0$

7. Phân tích hồi quy tuyến tính đơn

- Đồi thuyết $H_a : y \neq y_0$ (hai phía)
- Đồi thuyết $H_a : y > y_0$ (một phía)
- Đồi thuyết $H_a : y < y_0$ (một phía)

Case Study

Công ty dầu ăn Cái Lân đang xem xét việc giảm giá bán sản phẩm (loại bình 5 lít) để tăng lượng hàng bán ra, đồng thời quảng bá sản phẩm của mình đến khách hàng. Người quản lí của công ty muốn tính toán xem nếu sản phẩm này được giảm giá đi 1000 đồng/lít thì lượng hàng trung bình bán ra sẽ thay đổi thế nào. Đồng thời, nếu như giảm giá 1000 đồng/lít mà lượng hàng bán thêm được là nhiều hơn 50000 sản phẩm thì công ty sẽ tiến hành 1 chiến dịch khuyến mại trong 1 tháng với giá giảm đi là 10000 đồng/lít. Để tiến hành nghiên cứu này, phòng marketing của công ty đã dựa vào các số liệu bán hàng của công ty trong vòng 15 tháng qua ($n = 15$ quan sát) để thu thập số liệu về *giá bán* (X) và *lượng bán* (Y) cho loại dầu ăn này. Nghiên cứu viên sau khi tiến hành các thống kê mô tả đã quyết định dùng hàm tuyến tính để xem xét ảnh hưởng của giá đến lượng bán:

$$Y = w_0 + w_1 X + \epsilon_i$$

Dùng số liệu của mẫu, ước lượng được hàm hồi quy mẫu có dạng.

$$\hat{Y} = 6225 - 30.5X$$

Câu hỏi

- Theo kết quả của mô hình, khi giá giảm 1 đơn vị, lượng hàng bán ra thay đổi thế nào?
- Liệu khi giá giảm đi 1000 đồng 1 lít thì lượng hàng bán thêm lớn hơn được 50000 sản phẩm như các nhà nghiên cứu muốn kiểm tra không?
- Giá bán quyết định bao nhiêu % trong sự thay đổi của lượng bán?
- Nếu giá bán là 150000 đồng 1 bình thì lượng bán dự báo là bao nhiêu?

Bài tập

B 7.1. Bảng dưới đây cho các cặp biến phụ thuộc và độc lập. Trong mỗi trường hợp hãy cho biết quan hệ giữa hai biến là: *cùng chiều, ngược chiều hay không xác định?* Hãy giải thích?

7. Phân tích hồi quy tuyến tính đơn

Biến phụ thuộc	Biến độc lập
Vốn đầu tư	Lãi suất
Tiết kiệm cá nhân	Lãi suất
Cầu về tiền	GDP
Sản lượng	Vốn cơ bản (hoặc lao động)
Lượng cầu về xe máy	Giá xăng
Lượng điện tiêu thụ của hộ gia đình	Giá gas

B 7.2. Quan sát về *thu nhập* (X USD/tuần) và *chi tiêu* (Y USD/tuần) của 10 người, ta thu được số liệu sau

<i>thu nhập</i>	31	50	47	45	39	50	35	40	45	50
<i>chi tiêu</i>	29	42	38	30	29	41	23	36	42	48

1. Ước lượng hàm hồi quy tuyến tính:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

2. Nêu ý nghĩa kinh tế của các hệ số hồi quy đã ước lượng được. Các giá trị có phù hợp với lý thuyết kinh tế hay không?
3. Tìm khoảng tin cậy của β_0, β_1 với độ tin cậy 95%?
4. Kiểm định giả thiết $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$ với mức ý nghĩa 5%?
5. Tính R^2 và đánh giá mức độ phù hợp của mô hình?
6. Dự báo chi tiêu của một người có mức thu nhập 40USD/tuần

B 7.3. Một nghiên cứu về mối quan hệ giữa số đơn vị sản phẩm và các yếu tố đầu vào của quá trình sản xuất ở một số cơ sở sản xuất đã đưa ra những mô hình hồi quy. Lúc đầu nhà nghiên cứu chú trọng vào quan lý nguồn nhân lực nên đưa ra mô hình sau:

$$S = \beta_0 + \beta_1 L + \epsilon$$

S là sản lượng, L là lao động (người). Kết quả hồi quy như sau

7. Phân tích hồi quy tuyến tính đơn

Method: Ordinary Least Squares Estimation
Dependent variable: S
20 observations
Variable Coefficient Standard Error t-statistic Prob
C 34.4438 29.0219 1.1868 .25
L 19.2371 6.8786 27.967 .000
R-Squared: .30290 F-statistic: 7.8213
Adj-R-Squared: .26417 SE. of Regression: 49.5267
Residual Sum of Squares: 44152.1 Mean of dependent Variable: 109.4666
SD. of dependent Variable: 57.7367 Maximum of Log-likelihood: -105.3754
DW-statistic: 0.7151

1. Viết hàm hồi quy tổng thể, hàm số đó và các tham số có ý nghĩa như thế nào?
2. Viết hàm hồi quy mẫu. Các hệ số của hồi quy mẫu có phù hợp với lý thuyết kinh tế không?
3. Theo lý thuyết thì không có lao động sẽ không có sản lượng, nhưng trong hàm hồi quy mẫu thì không có lao động ước lượng điểm mức sản lượng lại không bằng không. Trên thực tế giá trị đó có thể coi là bằng không hay không?
4. Hệ số góc của mô hình có ý nghĩa thống kê không?
5. Hệ số xác định bằng bao nhiêu? Giá trị đó có ý nghĩa như thế nào?
6. Có thể coi hàm hồi quy phù hợp không?
7. Tìm ước lượng điểm cho phương sai của yếu tố ngẫu nhiên?
8. RSS, ESS, TSS bằng bao nhiêu?
9. Tìm khoảng tin cậy cho hệ số chặn của mô hình với độ chính xác 95%?
10. Khi doanh nghiệp thêm một lao động thì sản lượng tăng trong khoảng nào?
11. Khi giảm bớt một lao động thì sản lượng giảm tối đa bao nhiêu đơn vị?
12. Có thể nói rằng khi bớt một lao động thì sản lượng giảm 30 đơn vị, đúng hay sai?
13. Nếu giảm một lao động thì sản lượng giảm nhiều hơn hay ít hơn 22 đơn vị?
14. Nếu tăng một lao động thì sản lượng tăng nhiều hơn 20 đơn vị, đúng hay sai?
15. Tìm ước lượng điểm mức sản lượng với doanh nghiệp có 30 lao động?
16. Tìm mức sản lượng trung bình và cá biệt khi doanh nghiệp có 30 lao động?

7. Phân tích hồi quy tuyến tính đơn

B 7.4. Có số liệu thống kê về *lãi suất ngân hàng* (X , % năm) và *tổng vốn đầu tư* (Y , tỉ đồng) trên địa bàn tỉnh A qua 10 năm liên tiếp như sau:

Năm	1	2	3	4	5	6	7	8	9	10
X	7.0	6.5	6.5	6.0	6.0	6.0	5.5	5.5	5.0	4.5
Y	29	32	31	34	32	35	40	43	48	50

1. Hãy lập mô hình hồi quy tuyến tính mô tả quan hệ giữa tổng vốn đầu tư và lãi suất ngân hàng (mô hình hồi quy đơn)? Nêu ý nghĩa của các hệ số hồi quy ước lượng được? Đánh giá mức độ phù hợp của mô hình?
2. Kiểm định giả thiết “Hệ số hồi quy của X trong hàm hồi quy tổng thể bằng 0 với mức ý nghĩa 2%” và nêu ý nghĩa của kết quả?
3. Dự báo tổng vốn đầu tư trung bình khi lãi suất là 4.8% năm với độ tin cậy 98%?

B 7.5. Cho QA là lượng bán (đơn vị; nghìn lít), PA là giá bán (đơn vị nghìn đồng/lít) của hãng nước giải khát A, thời gian từ quý 1 năm 2001 đến quý 4 năm 2006 và kết quả hồi quy mô hình như sau:

Dependent Variable: QA
Method: Least squares
Sample: 2001Q1 2006Q4
Included observations: 24
Variable Coefficient Std.error t-statistic Prob
C 1814.139 174.1613 10.41643 0.000
PA -51.75140 9.840903 -5.258806 0.000
R-Squared: 0.556943 Mean dependent var: 923.5833
Adjusted Squared: 0.536804 S.D. dependent var: 292.7673
S.E. of regression: 199.2530 F- statistic: 27.65504
Sum squared resid: 873438.5 Prob(F- statistic): 0.000028

1. Viết hàm hồi quy tổng thể, tổng thể ngẫu nhiên, hàm hồi quy mẫu và mẫu ngẫu nhiên?
Giải thích ý nghĩa kết quả ước lượng?
2. Tìm ước lượng điểm lượng bán trung bình khi giá bán là 20 nghìn đồng/lít?
3. Lượng bán có thực sự phụ thuộc vào giá bán không?
4. Giảm giá có làm tăng lượng bán không?
5. Giảm giá một nghìn thì lượng bán thay đổi trong khoảng nào?
6. Giá tăng một nghìn đồng thì lượng bán giảm tối đa bao nhiêu?

7. Phân tích hồi quy tuyến tính đơn

7. Có thể cho rằng giá tăng một nghìn đồng thì lượng bán giảm hơn 50 nghìn lít hay không?
8. Tính các đại lượng TSS, ESS?
9. Hệ số xác định của mô hình bằng bao nhiêu? Giải thích ý nghĩa của hệ số đó?
10. Tìm ước lượng điểm và khoảng cho phương sai sai số ngẫu nhiên?
11. Dự báo giá trị trung bình và cá biệt của lượng bán khi giá bán là 18 nghìn đồng/lít?

B 7.6. Số liệu về *diện tích nhà (sqft)* và *giá bán (price)* của một tập mẫu gồm 14 mẫu nhà bán trong khu vực San Diego. Hãy xây dựng mô hình hồi quy phân tích cho dữ liệu này

#	sqft (x)	price (y)	(ĐVT: ngàn dollars)
1	1.065	199	
2	1.254	288	
3	1.300	235	
4	1.577	285	
5	1.600	239	
6	1.750	293	
7	1.800	285	
8	1.870	365	
9	1.935	295	
10	1.948	290	
11	2.254	385	
12	2.600	505	
13	2.800	425	
14	3.000	415	

B 7.7. Các dữ liệu sau đây là các số đo *chiều cao* và *trọng lượng* của 10 người đàn ông:

chiều cao	63	71	72	68	75	66	68	76	71	70
trọng lượng	158	156	148	163	155	153	158	150	154	145

1. Tìm hồi quy tuyến tính *chiều cao* đối với *trọng lượng*.
2. Tìm hồi quy tuyến tính của *trọng lượng* đối với *chiều cao*.
3. Giải thích tại sao hai phương trình (trong (1) và (2)) là khác nhau?

B 7.8. Các dữ liệu sau đây là về số lượng các ống bị lỗi (Y) trong lô hàng và tổng số các ống (X) trong một lô hàng cho 12 lô hàng.

X	5	10	4	10	7	8	8	5	10	5	12	6
Y	30	51	26	52	40	43	45	31	52	30	59	36

7. Phân tích hồi quy tuyến tính đơn

Thử tìm một phương trình hồi quy bậc hai và giải thích dữ liệu.

B 7.9. Theo dõi doanh thu X và tiền lời Y của 10 đại lý thức ăn chăn nuôi trong một tháng tại tỉnh A ta có kết quả sau: (đơn vị triệu đồng/tháng)

X	32.0	34.0	36.0	38.0
Y	4.2	4.4	4.6	5.0
n_i	2	3	3	2

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.10. Để nghiên cứu ảnh hưởng của lượng phân bón X (tạ/ha) và năng suất lúa Y (tấn/ha). Người ta thí nghiệm trên 10 thửa ruộng. Sau thu hoạch ta có kết quả sau:

X	1.2	1.4	1.5	1.6
Y				
3.9	1	2		
4.1		2	2	
4.3			2	1

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.11. Theo dõi doanh thu X và tiền lời Y của một cửa hàng bán giống cây trong 12 tháng ta có kết quả sau: (đơn vị triệu đồng/tháng)

X	14.0	16.0	18.0	20.0
Y	2.8	3.2	3.0	3.4
n_i	2	4	4	2

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.12. Điều tra tổng sản phẩm nông nghiệp X và tổng tài sản cố định Y (đơn vị triệu đồng) tại 10 nông trại ta thu được các số liệu sau

X	11.3	12.9	13.6	16.8	18.8	22.0	22.2	23.7	26.6	27.5
Y	13.2	15.6	17.2	18.8	20.2	21.9	22.4	23.0	24.4	24.5

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X và của X theo Y .

7. Phân tích hồi quy tuyến tính đơn

B 7.13. Theo dõi vi lượng A trong đất trồng X (mg/kg) và năng suất một loại rau Y (tấn/ha) ta có kết quả sau

X	83	80	90	83	85	95	90	85	93	88
Y	5	4	7.5	5.5	5.3	5.6	6.8	6.9	7.3	6.5

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.14. Người ta xét trên 10 mảnh ruộng được kết quả sau giữa tỷ lệ phần trăm hạt chắc X và năng suất lúa Y (tấn/ha):

X	16	17	18	19	20	21	22	23	24	25	26	27
Y	10	9.3	8.7	9.7	9	8.1	8	8.2	7.7	7.6	7.9	7.8

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.15. Số vi khuẩn Y (triệu con) sinh sản sau X (giờ) được ghi lại trong bảng sau qua một thí nghiệm:

X	0	1	2	3	4	5
Y	30	32	35	40	48	52

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .
- Dự báo số vi khuẩn sau 10 giờ

B 7.16. Để nghiên cứu tác dụng của phân vi sinh X (tạ/ha) tới năng suất cà chua Y (tấn/ha), người ta thí nghiệm trên 20 thửa ruộng. Sau thu hoạch ta có kết quả sau

X	0.15	0.17	0.19	0.21	0.23
Y	3	2	3		
20					
22		4	1	2	
24			2	2	1

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.17. Để thực hiện một công trình nghiên cứu về mối quan hệ giữa chiều cao Y (m) và đường kính X (cm) của một loại cây, người ta quan sát trên một mẫu ngẫu nhiên và có kết quả

7. Phân tích hồi quy tuyến tính đơn

X	28	28	24	30	60	30	32	42	43	49
Y	5	6	5	6	10	5	7	8	9	10

- Tìm hệ số tương quan mẫu và cho nhận xét.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .
- Hãy dự báo chiều cao của cây có đường kính 45 (cm)

B 7.18. Chiều dài xương đùi X (cm) và chiều cao Y (cm) của những người đàn ông độ tuổi 20-30 là các biến ngẫu nhiên tuân theo quy luật phân phối chuẩn. Đo chiều dài xương đùi và chiều cao của 10 người đàn ông được chọn ngẫu nhiên ở độ tuổi trên. Kết quả được cho trong bảng sau:

X	44	46	47	47	48	49	50	50	51	52
Y	155	169	163	166	169	172	174	176	176	179

- Tìm hệ số tương quan mẫu và cho nhận xét về mức độ tương quan giữa X và Y .
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .
- Hãy dự báo xem nếu giá trị của X giảm bớt 1 (cm) thì giá trị tương ứng của Y biến thiên như thế nào? Tại sao?

B 7.19. Cho một mẫu ngẫu nhiên cỡ 18 được chọn từ tổng thể (X, Y) có phân phối chuẩn 2 chiều, người ta tính được giá trị hệ số tương quan mẫu $r = 0.32$. Với mức ý nghĩa 5%, có sự tương quan tuyến tính giữa X và Y không?

B 7.20. Nghiên cứu về lượng phân bón X (kg) được dùng để bón cho ruộng trong một vụ Y ($\text{kg}/1000\text{m}^2$) là năng suất lúa. Thông kê ở 30 hộ gia đình, kết quả như sau

X	40	40	50	50	50	60	60	60
Y	270	280	280	290	300	300	310	320
số hộ	3	5	2	6	4	3	5	2

- Tìm hệ số tương quan giữa X và Y .
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.21. Theo các nghiên cứu cho thấy giữa lượng đạm (N) và carbon (C) trong đất mùn có mối quan hệ tuyến tính. Hãy kiểm tra lại qua dữ liệu đất mùn ở Quảng Ninh

C	1.79	4.39	3.07	4.40	3.10	5.60	7.81	3.95	4.71
N	0.06	0.42	0.18	0.30	0.22	0.38	0.46	0.23	0.42

- Tìm hệ số tương quan mẫu.

7. Phân tích hồi quy tuyến tính đơn

- Viết phương trình hồi quy tuyến tính thực nghiệm của N theo C .
- Kiểm tra lại nhận xét N và C có quan hệ tuyến tính

B 7.22. Nghiên cứu mối liên hệ giữa X (VND) là số tiền đầu tư cho việc phòng bệnh tính trên đầu người và Y là tỷ lệ người mắc bệnh ở 50 địa phương thu được bảng kết quả sau

X	2.0	2.5	3.0	3.5	4.0
Y					
100				2	3
200			3	6	2
300		4	6	3	
400	1	6	4	1	
500	6	3			

- Tìm hệ số tương quan mẫu.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .
- Nếu năm sau đầu tư cho phòng bệnh là 600 VND/người thì tỷ lệ mắc bệnh khoảng bao nhiêu phần trăm?

B 7.23. X (%) và Y (kg/mm^2) là hai chỉ tiêu chất lượng của một loại sản phẩm. Điều tra ở một số sản phẩm, ta thu được bảng kết quả sau (x_i, y_i)

$$\begin{array}{ccccccc}
 (8, 15) & (4, 15) & (2, 5) & (2, 10) & (8, 25) & (6, 10) \\
 (4, 10) & (4, 10) & (8, 20) & (6, 15) & (2, 5) & (6, 10) \\
 (6, 20) & (8, 15) & (6, 15) & (8, 20) & (8, 20) & (6, 10) \\
 (6, 15) & (4, 15) & (6, 20) & (6, 15) & (8, 15) & (6, 25) \\
 (6, 15) & (8, 25) & (8, 15) & (6, 10) & (6, 15) & (6, 20)
 \end{array}$$

- Tìm hệ số tương quan mẫu.
- Viết phương trình hồi quy tuyến tính thực nghiệm của Y theo X .

B 7.24. 16. Người ta nghiên cứ về số lượng protein chứa trong hạt lúa mì Y và năng suất lúa trên 10 thửa ruộng cùng kích thước, kết quả đo đạc như sau:

năng suất X	9.9	10.2	11	11.6	11.8	12.5	12.8	13.5	14.3	14.4
protein Y	10.7	10.8	12.1	12.5	12.2	12.8	12.4	11.8	11.8	12.6

- Tính hệ số tương quan mẫu và cho nhận xét.
- Xác định đường hồi quy tuyến tính thực nghiệm của Y theo X

Phân tích hồi quy bội

Rất nhiều các nghiên cứu trên thế giới quan tâm tới mối quan hệ giữa *thu nhập* và trình độ *học vấn*. Chúng ta kỳ vọng rằng, ít ra về trung bình mà nói, học vấn càng cao, thì thu nhập càng cao. Vì vậy, chúng ta có thể lập phương trình hồi quy sau:

$$\text{thu nhập} = \color{red}{w_0} + \color{red}{w_1} \text{học vấn} + \epsilon \quad (8.1)$$

Tuy nhiên, mô hình này đã bỏ qua một yếu tố khá quan trọng là mọi người thường có mức thu nhập cao hơn khi họ làm việc lâu năm hơn, bất kể trình độ học vấn của họ thế nào. Vậy nên, mô hình tốt hơn cho mục đích nghiên cứu của chúng ta sẽ là:

$$\text{thu nhập} = \color{red}{w_0} + \color{red}{w_1} \text{học vấn} + \color{red}{w_2} \text{tuổi} + \epsilon \quad (8.2)$$

Nhưng người ta cũng thường quan sát thấy, thu nhập có xu hướng tăng chậm dần khi người ta càng nhiều tuổi hơn so với thời trẻ. Để thể hiện điều đó, chúng ta mở rộng mô hình như sau:

$$\text{thu nhập} = \color{red}{w_0} + \color{red}{w_1} \text{học vấn} + \color{red}{w_2} \text{tuổi} + \color{red}{w_3} \text{tuổi}^2 + \epsilon \quad (8.3)$$

Và chúng ta sẽ kỳ vọng rằng, w_2 mang dấu dương, và w_3 mang dấu âm.

8.1

Biểu diễn đại số của mô hình hồi quy

Phương trình hồi quy tuyến tính bội tổng quát

$$y = \color{red}{w_0} + \color{red}{w_1}x_1 + \dots + \color{red}{w_k}x_k + \epsilon \quad (8.4)$$

Biểu diễn các biến và tham số thành vector

$$\mathbf{x} = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_k \end{bmatrix}^\top \quad (8.5)$$

$$\mathbf{w} = \begin{bmatrix} \color{red}{w_0} & \color{red}{w_1} & \color{red}{w_2} & \dots & \color{red}{w_k} \end{bmatrix}^\top \quad (8.6)$$

8. Phân tích hồi quy bội

Mô hình hồi quy tuyến tính bội tổng quát trở thành

$$y = \mathbf{w}^\top \mathbf{x} + \epsilon \quad (8.7)$$

Bài toán. Tìm một quy luật phù hợp tốt nhất với tập dữ liệu mẫu quan sát \mathcal{D}

	x_1	x_2	\dots	x_k	
i	\mathbf{x}_i				y_i
1					
2					
\dots					
n					

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i \quad (8.8)$$

Giả định Gauss-Markov và Mô hình xác suất

1. $\{\mathbf{x}_i, y_i\}$ là tập mẫu ngẫu nhiên và độc lập.
2. Sai số ϵ_i có trung bình là 0 và phương sai giống nhau (*homoskedasticity*); nghĩa là không có phương sai thay đổi (*heteroskedasticity*).
3. Không có tương quan giữa ϵ_i và ϵ_j ($i \neq j$).
4. Không có tương quan giữa \mathbf{x}_i và ϵ_i .
5. Không có hiện tượng đa cộng tuyến hoàn hảo giữa các biến độc của \mathbf{x}
6. Sai số ϵ_i phân phối theo luật phân phối chuẩn $\mathcal{N}(0, \sigma^2)$. (không bắt buộc)

Mô hình xác suất tương ứng

$$p(y | \mathbf{x}; \mathbf{w}) = \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2) \quad (8.10)$$

và phương trình hồi quy tổng thể

$$\hat{y} = \mathbb{E}(y | \mathbf{x}; \mathbf{w}) = \mathbf{w}^\top \mathbf{x} \quad (8.11)$$

- Tổng bình phương của tất cả các giá trị lỗi RSS

$$RSS = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \quad (8.12)$$

8. Phân tích hồi quy bội

- Tổng bình phương toàn phần TSS

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (8.13)$$

- Tổng bình phương hồi quy ESS

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\mathbf{w}^\top \mathbf{x}_i - \bar{y})^2 \quad (8.14)$$

Bài toán. Chúng ta tìm quy luật tốt nhất là quy luật cực tiểu hóa giá trị RSS

Lời giải.

- Tính đạo hàm riêng hay gradient theo \mathbf{w}_j ($j = 0, \dots, k$)

$$\nabla_{\mathbf{w}}(RSS)$$

- Cho gradient bằng 0 và giải phương trình

$$\nabla_{\mathbf{w}}(RSS) = 0$$

- Kết quả

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (8.15)$$

với

$$\mathbf{X} = \underbrace{\begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}}_{\text{input data matrix}}, \quad \mathbf{y} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\text{target vector}} \quad (8.16)$$

■

- Ước lượng của phương sai sai số ngẫu nhiên

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^n \epsilon_i^2 \quad (8.17)$$

- Ma trận hiệp phương sai của hệ số ước lượng

$$\begin{pmatrix} \text{Var}(\hat{w}_0) & \text{cov}(\hat{w}_0, \hat{w}_1) & \cdots & \text{cov}(\hat{w}_0, \hat{w}_k) \\ \text{cov}(\hat{w}_1, \hat{w}_0) & \text{Var}(\hat{w}_1) & \cdots & \text{cov}(\hat{w}_1, \hat{w}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{w}_k, \hat{w}_0) & \text{cov}(\hat{w}_k, \hat{w}_1) & \cdots & \text{Var}(\hat{w}_k) \end{pmatrix} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1} \quad (8.18)$$

8. Phân tích hồi quy bội

	Population	Illiteracy	Income	Frost	Murder
Alabama	3615	2.1	3624	20	15.1
Alaska	365	1.5	6315	152	11.3
Arizona	2212	1.8	4530	15	7.8
Arkansas	2110	1.9	3378	65	10.1
California	21198	1.1	5114	20	10.3

- Phương sai của hệ số ước lượng

$$\begin{aligned}\text{Var}(\hat{\mathbf{w}}) &= \sigma^2 \text{diag} [(\mathbf{X}^\top \mathbf{X})^{-1}] \\ &= \hat{\sigma}^2 \text{diag} [(\mathbf{X}^\top \mathbf{X})^{-1}]\end{aligned}\quad (8.19)$$

Tính chất.

- Tổng các phần dư bằng 0
- Các biến độc lập và phần dư không tương quan
- Đường hồi quy mẫu luôn đi qua giá trị trung bình mẫu (\bar{x}, \bar{y})
- Trung bình của giá trị ước lượng của biến phụ thuộc bằng trung bình mẫu của nó
- Các ước lượng hệ số không chêch và hiệu quả

8.2

Ví dụ hồi quy

Ví dụ. Thống kê số vụ giết người (Murder) trên 100000 người của các bang của nước Mỹ trong năm 1977, và bốn biến

- Population:** tổng dân số
- Illiteracy:** tỉ lệ thất học
- Income:** thu nhập bình quân
- Frost:** số ngày giá rét trung bình

5 dòng đầu tiên của bảng dữ liệu

- Kết quả phân tích hồi quy tuyến tính cho ví dụ

Residuals :

Min	1Q	Median	3Q	Max
-4.7960	-1.6495	-0.0811	1.4815	7.6210

8. Phân tích hồi quy bội

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	1.235e+00	3.866e+00	0.319	0.7510		
Population	2.237e-04	9.052e-05	2.471	0.0173 *		
Illiteracy	4.143e+00	8.744e-01	4.738	2.19e-05 ***		
Income	6.442e-05	6.837e-04	0.094	0.9253		
Frost	5.813e-04	1.005e-02	0.058	0.9541		

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

Residual standard error: 2.535 on 45 degrees of freedom

Multiple R-squared: 0.567, Adjusted R-squared: 0.5285

F-statistic: 14.73 on 4 and 45 DF, p-value: 9.133e-08

- **Kết luận:**

- Biến **Population** có ý nghĩa đối với mô hình về mặt thống kê (với mức ý nghĩa *)
- Biến **Illiteracy** có ý nghĩa đối với mô hình về mặt thống kê (với mức ý nghĩa ***)
- Biến **Income** không có ý nghĩa (có thể loại bỏ)
- Biến **Frost** không có ý nghĩa (có thể loại bỏ)
- Mô hình có thể giải thích được 52.85% sự thay đổi của biến **Murder**
- Mô hình tương đối tốt (*p*-value = 9.133e-08)

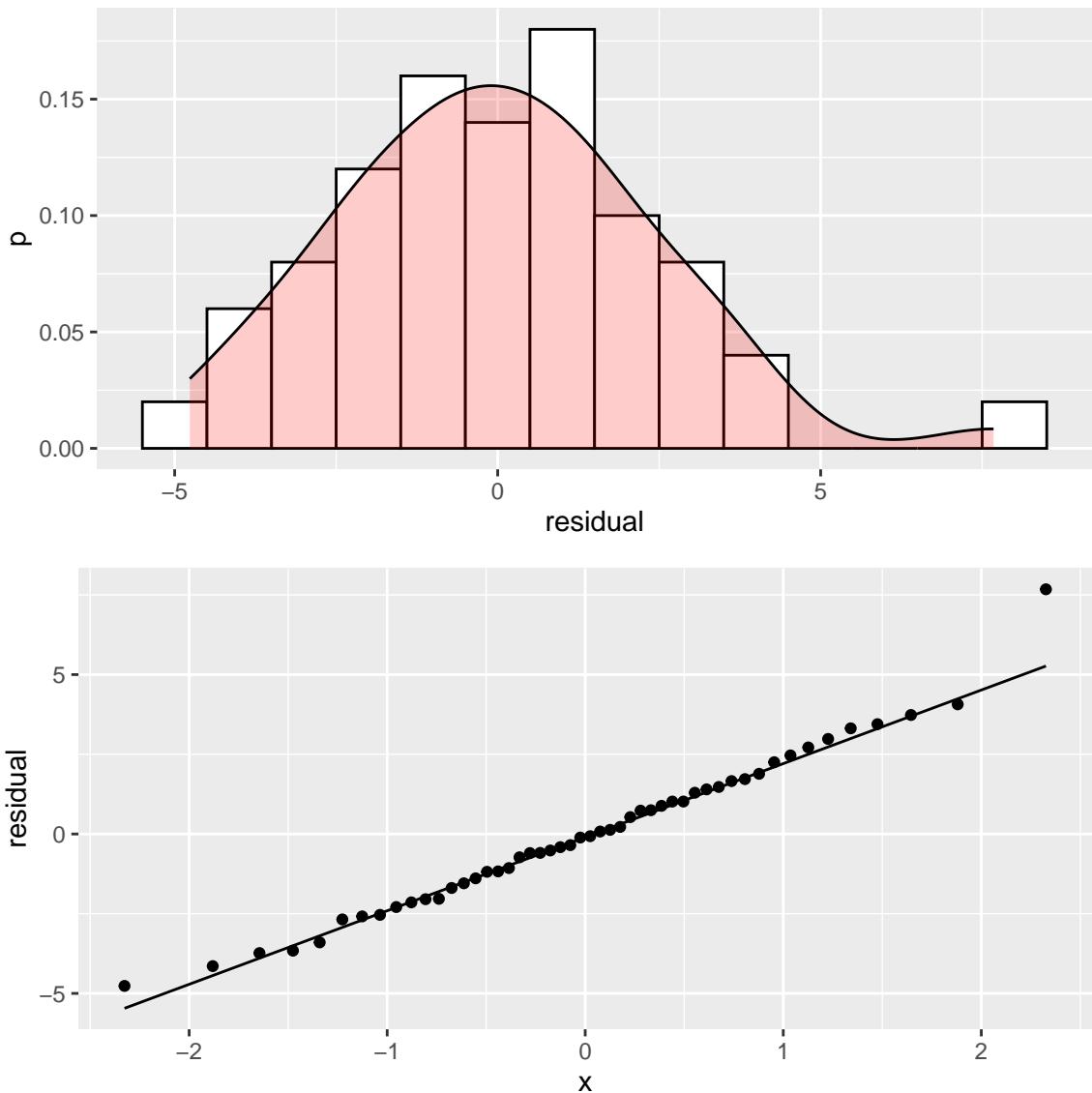
8.3

Kiểm định mô hình

8.3.1. Phân tích phần dư

Kiểm tra phần dư có thỏa điều kiện phân phối chuẩn bằng cách vẽ các biểu đồ histogram và Q-Q plot

8. Phân tích hồi quy bội



8.3.2. Phân tích phương sai

Kiểm định phát hiện

Để phát hiện hiện tượng phương sai sai số thay đổi, ta có thể sử dụng một trong ba mô hình *hồi quy mẫu phụ* để kiểm định

- Kiểm định Breusch-Pagan

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_{1,i} + \dots + \alpha_p z_{p,i} + \eta_i \quad (8.20)$$

- Kiểm định Glejser

$$|\sigma_i| = \alpha_0 + \alpha_1 z_{1,i} + \dots + \alpha_p z_{p,i} + \eta_i \quad (8.21)$$

8. Phân tích hồi quy bội

- Kiểm định Harvey-Godfrey

$$\ln(\sigma_i^2) = \alpha_0 + \alpha_1 z_{1,i} + \dots + \alpha_p z_{p,i} + \eta_i \quad (8.22)$$

trong đó các biến $\{z_1, \dots, z_p\} \subseteq \{x_1, \dots, x_k\}$

Bài toán

- Giả thuyết $H_0 : \alpha_i = 0, \forall i$
- Đốii thuyết $H_a : \alpha_i \neq 0, \exists i$

Nếu giả thuyết H_0 chấp nhận thì có nghĩa là phương sai sai số trong mô hình không thay đổi; ngược lại là có hiện tượng phương sai thay đổi.

Giải pháp khắc phục

- Thay đổi mô hình
- Sử dụng phương pháp phân tích Bayesian tổng quát

8.3.3. Kiểm định đa cộng tuyến

Định nghĩa. Nếu các biến *giải thích* có quan hệ chặt chẽ với nhau, có mối quan hệ tuyến tính, nghĩa là chúng không còn độc lập lẫn nhau.

Có hai loại **đa cộng tuyến** (*multicollinearity*):

1. Đa cộng tuyến nhân tạo: một biến mới được tạo ra từ một biến khác
2. Đa cộng tuyến tự nhiên: các biến có mối quan hệ lẩn nhau

Nguyên nhân

- Do dữ liệu không đầy đủ
- Do chọn biến độc lập
 - có độ biến thiên nhỏ
 - có mối quan hệ nhân quả
 - đồng thời phụ thuộc vào một điều kiện khác

Hậu quả

- Sai số chuẩn của các tham số tăng
- Khoảng tin cậy tăng

8. Phân tích hồi quy bội

- Các ước lượng trong phân tích hồi quy giảm sự chính xác
- Có những biến không có ý nghĩa thống kê

Dấu hiệu nhận biết

- Dựa vào ma trận hệ số tương quan
- Dựa vào hệ số VIF từ chạy các hồi quy phụ

Kiểm định hệ số tương quan

Định nghĩa. Hệ số tương quan đo lường mức độ quan hệ tuyến tính giữa hai biến, không phân biệt biến này phụ thuộc vào biến kia

giá trị	tương quan
[0.9...1]	hoàn hảo
[0.7...0.9)	mạnh
[0.5...0.7)	vừa
[0.1...0.5)	yếu
[0...0.1)	không có

- Hệ số tương quan Pearson r :

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (8.23)$$

- Hệ số tương quan Spearman ρ : Khi phân phối của x, y không phải là phân phối chuẩn hoặc có các giá trị quan sát bất thường (lớn quá hoặc nhỏ quá)

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (8.24)$$

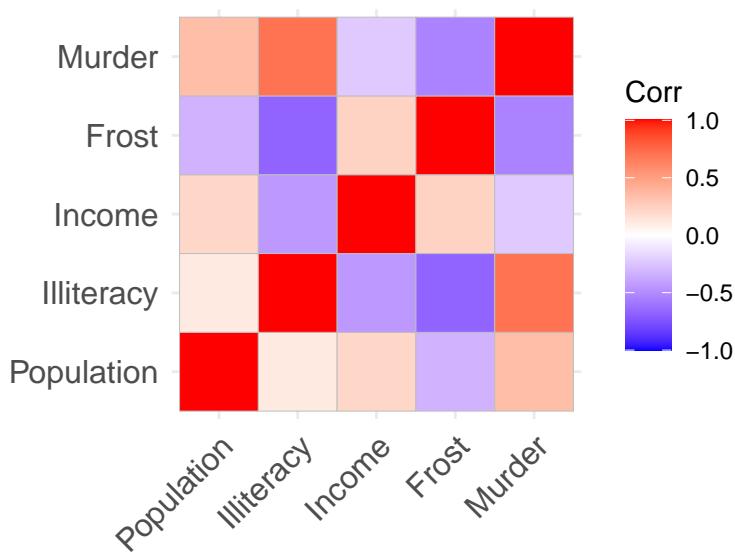
với $d_i = rank(x_i) - rank(y_i)$ và $rank(\cdot)$ là hàm thứ bậc cho một giá trị của biến

x	$rank(x)$
4.2	2
5.3	3
3.5	1
6.7	4

- Hệ số tương quan Kendall τ :

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} sign(x_i - x_j) sign(y_i - y_j) \quad (8.25)$$

8. Phân tích hồi quy bội



Kiểm định VIF

- Hệ số VIF (*variance inflation factor*) có thể dùng để kiểm tra đa cộng tuyến

Giá trị VIF	tương quan
[1, 2)	không có mối tương quan giữa biến độc lập này và bất kỳ biến nào khác
[2, 5)	có một mối tương quan vừa phải, nhưng không nghiêm trọng
[5, 10)	có một mối tương quan và khá nghiêm trọng
[10, ∞)	chắc chắn có đa cộng tuyến

- Xác định hệ số VIF cho các biến: chúng ta chạy mô hình hồi quy tuyến tính với x_i như là hàm của tất cả các biến còn lại

$$x_i = f_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k) \quad (8.26)$$

gọi R_i là hệ số xác định của kết quả hồi quy thì hệ số VIF cho biến x_i được tính

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \quad (8.27)$$

Giải pháp khắc phục

- Loại bỏ một số biến độc lập có tương quan cao.
- Bổ sung dữ liệu hoặc tìm thêm những dữ liệu mới.

8. Phân tích hồi quy bội

VIF	
Population	1.245282
Illiteracy	2.165848
Income	1.345822
Frost	2.082547

- Thay đổi mô hình

8.4 Lựa chọn mô hình

Một nghiên cứu thường có nhiều biến, do đó số mô hình có thể rất nhiều

- Nếu $n = 2$ biến thì số mô hình tối thiểu là $2^2 - 1$
- Nói chung, nếu $n = k$ biến thì số mô hình tối thiểu là $2^k - 1$

Tuy nhiên, không phải biến nào cũng thực sự có ý nghĩa do đó việc lựa chọn biến thực sự có ý nghĩa là một việc rất quan trọng

Định nghĩa. Nếu mô hình có quá nhiều biến (tham số) so với dữ liệu quan sát thì có thể dẫn đến tình trạng *quá khớp* (*overfitting*)

- Hậu quả là mô hình dự đoán không chính xác

Định nghĩa. Nếu mô hình có quá ít biến (tham số) so với dữ liệu quan sát thì có thể dẫn đến tình trạng *chưa khớp* (*underfitting*)

- Hậu quả là mô hình dự đoán không chính xác
- Tham số ước lượng bị lệch và phương sai tăng

8.4.1. Phương pháp stepwise regression

Là phương pháp phân tích hồi quy bội trong đó các biến độc lập được bổ sung dần dần (từng biến một) vào phương trình hồi quy và ảnh hưởng của chúng tính bằng mức bổ sung và khả năng giải thích của phương trình hồi quy.

- Lựa chọn tiến (*forward selection*):
 1. Bắt đầu mô hình rỗng không có biến
 2. Bổ sung biến vào mô hình nếu có sự cải thiện có ý nghĩa thống kê (dựa trên một tiêu chí nào đó)

Lặp lại quá trình này cho đến khi không có gì cải thiện mô hình.

8. Phân tích hồi quy bội

- Lựa chọn lùi (*backward elimination*):

1. Bắt đầu mô hình đầy đủ các biến
2. Xóa biến ra khỏi mô hình nếu có sự cải thiện có ý nghĩa thống kê (dựa trên một tiêu chí nào đó)

Lặp lại quá trình này cho đến không có gì cải thiện mô hình.

- Lựa chọn từng bước (*bidirectional stepwise*): sự kết hợp của hai phương pháp trên, kiểm tra ở mỗi bước để biết các biến được thêm vào hoặc loại ra.

Giả sử ta có 4 biến độc lập $\{X_1, X_2, X_3, X_4\}$ và 1 biến phụ thuộc Y

Số bước	Lùi	Tiến	Từng bước
1	$\{X_1, X_2, X_3, X_4\}$	$\{X_4\}$	$\{X_4\}$
2	$\{X_1, X_2, X_4\}$	$\{X_1, X_4\}$	$\{X_1, X_4\}$
3	$\{X_1, X_2\}$	$\{X_1, X_2, X_4\}$	$\{X_1, X_2, X_4\}$
4	$\{X_2\}$	$\{X_1, X_2, X_3, X_4\}$	$\{X_1, X_2\}$
5			$\{X_1, X_2, X_4\}$

Các tiêu chí đánh giá mô hình

- Hệ số xác định R_{adj}^2 (càng lớn càng tốt)
- Giá trị kiểm định f (càng lớn càng tốt)
- Hệ số AIC (Akaike Information Criterion) (càng nhỏ càng tốt)

$$AIC = n \log(RSS) + 2k \quad (8.28)$$

- Hệ số BIC (Bayesian Information Criterion) (càng nhỏ càng tốt)

$$BIC = n \log(RSS) + k \log n \quad (8.29)$$

8.5

Hồi quy với biến chuẩn hóa

Định nghĩa. Mô hình hồi quy trên các biến đều đã được chuẩn hóa trước khi đưa vào mô hình hồi quy

- Sau khi chuẩn hóa các biến sẽ có *trung bình* là 0 và *độ lệch chuẩn* là 1.
- Hồi quy dựa trên biến chuẩn hóa giúp loại bỏ các vấn đề có thể phát sinh do đơn vị đo lường khác nhau giữa các biến.

8. Phân tích hồi quy bội

- Trong kinh tế, các nhà kinh tế học có thể dựa vào phân tích hồi quy với biến chuẩn hóa để xác định rằng
 - biến nào quan trọng nhất (ứng với tham số có ý nghĩa thống kê và càng lớn càng quan trọng)
 - biến nào ít quan trọngđể đưa ra những quyết định hợp lý.

8.6

Hồi quy với biến định tính

Trong kinh tế xã hội có nhiều biến được đặc trưng bởi các trạng thái, tính chất hay phạm trù mà ta gọi là các biến định tính. Chẳng hạn

- Biến *giới tính* gồm hai trạng thái là nam và nữ,
- Biến *quê quán* đặc trưng bởi hai trạng thái thành thị và nông thôn,
- Biến *vùng miền* chia ba trạng thái là miền bắc, miền trung và miền nam

8.6.1. Biến độc lập là biến định tính

Biến định tính nhị phân

Định nghĩa. Trong mô hình tuyến tính có biến định tính, biến định tính sẽ được mã hóa thành các biến định lượng (trong kinh tế gọi là **biến giả**)

Thông tin về biến định tính *giới tính* nhân viên (với hai trạng thái là nam và nữ) có thể được thể hiện bởi biến giả D như sau

$$D = \begin{cases} 1 & \text{nếu là nam} \\ 0 & \text{nếu là nữ} \end{cases}$$

Ví dụ. Giả sử một công ty sử dụng hai quá trình sản xuất (kí hiệu quá trình sản xuất A và quá trình sản xuất B) để sản xuất ra một loại sản phẩm. Giả sử sản phẩm thu được từ mỗi một quá trình sản xuất là đại lượng ngẫu nhiên có phân phối chuẩn và có kỳ vọng khác nhau nhưng phương sai như nhau. Để xem xét kết quả sản lượng do 2 quá trình sản xuất A và B có khác nhau hay không người ta tiến hành lấy một mẫu được cho trong bảng dưới đây.

X (quá trình)	A	B	B	A	B	A	A	B	B	A
Y (sản lượng)	22.0	19.0	18.0	21.0	18.5	21.0	20.5	17.0	17.5	21.2

8. Phân tích hồi quy bội

Chúng ta có thể biểu thị phương trình hồi quy sản lượng sản phẩm

$$y_i = \textcolor{red}{w_0} + \textcolor{red}{w_1}d_i + \epsilon_i \quad (8.30)$$

trong d_i là biến giả nhận 1 trong 2 giá trị:

$$d_i = \begin{cases} 1 & \text{nếu sản lượng sản phẩm thu được từ quá trình sản xuất A} \\ 0 & \text{nếu sản lượng sản phẩm thu được từ quá trình sản xuất B} \end{cases}$$

X	A	B	B	A	B	A	A	B	B	A
D	1	0	0	1	0	1	1	0	0	1
Y	22.0	19.0	18.0	21.0	18.5	21.0	20.5	17.0	17.5	21.2

Kết quả hồi quy

Biến	Hệ số	Sai số chuẩn	t-statistic
Const	18	0.32	57.74
D	3.28	0.44	7.439
$R^2 = 0.8737$			

Phân tích và kết luận

- Hệ số chặn có ý nghĩa thống kê
- Biến D có ý nghĩa đối với mô hình về mặt thống kê
- Mô hình giải thích được 87.37% sự thay đổi của biến *sản lượng*
- Phương trình hồi quy chung

$$\hat{y} = \mathbb{E}(y | d) = 18 + 3.28d$$

phương trình cho quá trình A

$$\hat{y} = \mathbb{E}(y | d = 1) = 18 + 3.28 \cdot 1 = 21.28$$

phương trình cho quá trình B

$$\hat{y} = \mathbb{E}(y | d = 0) = 18$$

Biến định tính tổng quát

Nếu một biến định tính có m giá trị, thì chúng ta chỉ đưa $(m - 1)$ biến giả vào mô hình. Ví dụ, biến trình độ giáo dục *Education* có 3 giá trị $\{\text{low}, \text{medium}, \text{high}\}$ sẽ được chuyển thành 2 biến giả D_1 và D_2

8. Phân tích hồi quy bội

Giá trị	D_1	D_2	ý nghĩa
<i>low</i>	0	0	giá trị tham chiếu
<i>medium</i>	1	0	giá trị so sánh
<i>high</i>	0	1	giá trị so sánh

Quá trình mã hóa như sau

#	...	<i>Education</i>
1	...	<i>low</i>
2	...	<i>high</i>
3	...	<i>medium</i>
4	...	<i>medium</i>
5	...	<i>low</i>
...	...	

→

#	...	<i>Education</i>	D_1	D_2
1	...	<i>low</i>	0	0
2	...	<i>high</i>	0	1
3	...	<i>medium</i>	1	0
4	...	<i>medium</i>	1	0
5	...	<i>low</i>	0	0
...	...			

Cách mã hóa này khá giống với *one-hot encoding* nhưng bớt đi 1 biến. One-hot encoding không được sử dụng bởi nó sẽ làm chúng ta rơi vào *bẫy biến giả* (*dummy variable trap*) nghĩa là các biến giả tạo ra một đa cộng tuyến hoàn hảo.

8.6.2. Biến độc lập là biến định tính + biến định lượng

Xét bài toán hồi quy *lương* (*Salary*) theo biến *tuổi* (*Age*) và biến *giới tính* (*Gender*). Giả sử biến định tính và biến định lượng không tương tác với nhau ta có phương trình hồi quy tổng thể

$$\text{Salary} = \mathbf{w}_0 + \mathbf{w}_1 D + \mathbf{w}_2 \text{Age} + \epsilon, \quad (8.31)$$

và phương trình hồi quy mẫu

$$\text{Salary}_i = \mathbf{w}_0 + \mathbf{w}_1 D_i + \mathbf{w}_2 \text{Age}_i + \epsilon_i, \quad (8.32)$$

trong đó

- Salary_i là tiền lương của người thứ i
- D_i là biến giả (nam = 1, nữ = 0) của người thứ i
- Age_i là tuổi của người thứ i

Ý nghĩa của phương trình hồi quy

- Tiền lương trung bình của đàn ông theo tuổi

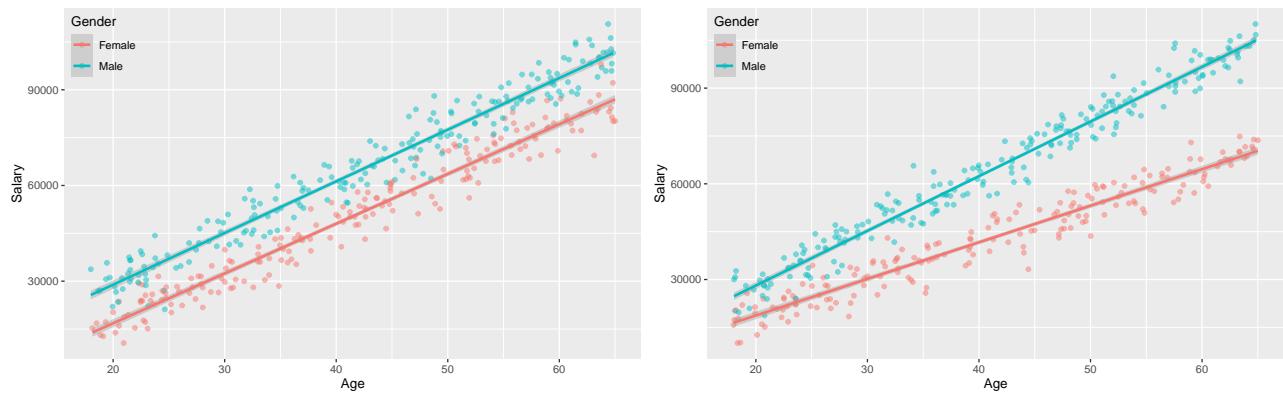
$$\mathbb{E}(\text{Salary} | \text{Age}, D = 1) = \mathbf{w}_0 + \mathbf{w}_1 + \mathbf{w}_2 \text{Age} \quad (8.33)$$

- Tiền lương trung bình của phụ nữ theo tuổi

$$\mathbb{E}(\text{Salary} | \text{Age}, D = 0) = \mathbf{w}_0 + \mathbf{w}_2 \text{Age} \quad (8.34)$$

8. Phân tích hồi quy bội

8.6.3. Ảnh hưởng tương tác



Mô hình chúng ta vừa đề cập giả thiết là không có sự tương tác giữa biến định tính và biến định lượng. Nghĩa là, biến *Gender* không ảnh hưởng tới hệ số góc của biến *Age*. Tuy nhiên, trong thực tế tốc độ tăng thu nhập của đàn ông cao hơn so với phụ nữ. Để kiểm tra liệu giữa hai biến có tương tác với nhau hay không chúng ta có thể sử dụng

- Kiểm định Chow
- hoặc thủ tục biến giả

Thủ tục biến giả

Trong thủ tục này chúng ta giới thiệu thêm một thành phần mới trong phương trình hồi quy mẫu

$$Salary = w_0 + w_1 D + w_2 Age + \underbrace{w_3 D \times Age}_{\text{tương tác}} + \epsilon, \quad (8.35)$$

Ý nghĩa của phương trình hồi quy

- Tiền lương trung bình của đàn ông theo tuổi

$$\mathbb{E}(Salary | Age, D = 1) = w_0 + w_1 + (w_2 + w_3)Age \quad (8.36)$$

- Tiền lương trung bình của phụ nữ theo tuổi

$$\mathbb{E}(Salary | Age, D = 0) = w_0 + w_2 Age \quad (8.37)$$

8.7

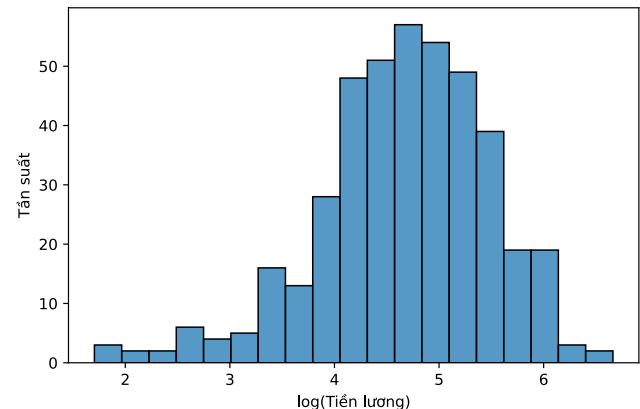
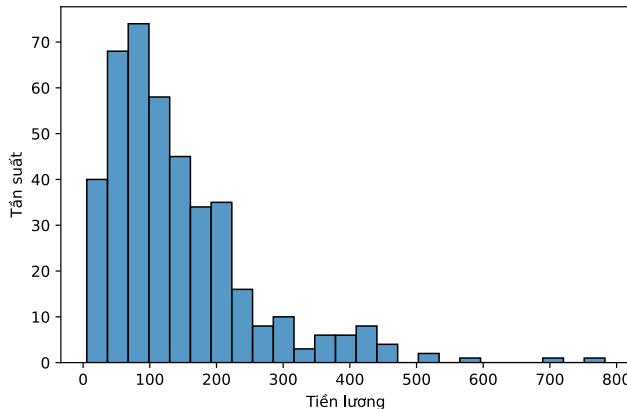
Hồi quy với hàm cơ sở

Dạng của hàm hồi quy là một vấn đề quan trọng, một trong những nhân tố có tính chất quyết định đối với kết quả nghiên cứu. Tuy vậy, vấn đề “dạng của hàm hồi quy” lại không có một cơ

8. Phân tích hồi quy bội

sở lý thuyết đủ mạnh để có thể khẳng định dạng của hàm hồi quy là dạng này mà không phải là dạng khác. Dạng của mô hình hồi quy là một vấn đề thực nghiệm.

Ý tưởng chung là sử dụng các phép biến đổi (hàm cơ sở) để tạo ra các biến mới; và giả định rằng với các biến mới đó có tồn tại một mối quan hệ tuyến tính giữa chúng. Ví dụ, trong kinh tế học lao động, người ta thường sử dụng dạng log của tiền lương thay vì tiền lương làm biến bởi vì phân phối của tiền lương có xu hướng bị lệch khá nhiều.



- dạng log-lin hoặc lin-log

$$\log y = w_0 + w_1 x + \epsilon \quad \text{hoặc} \quad y = w_0 + w_1 \log x + \epsilon \quad (8.38)$$

- dạng log-log

$$\log y = w_0 + w_1 \log x + \epsilon \quad (8.39)$$

- dạng lin-hyper

$$y = w_0 + w_1 \frac{1}{x} + \epsilon \quad (8.40)$$

- dạng đa thức

$$y = w_0 + w_1 x + w_2 x^2 + \epsilon \quad (8.41)$$

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$

- dạng kết hợp

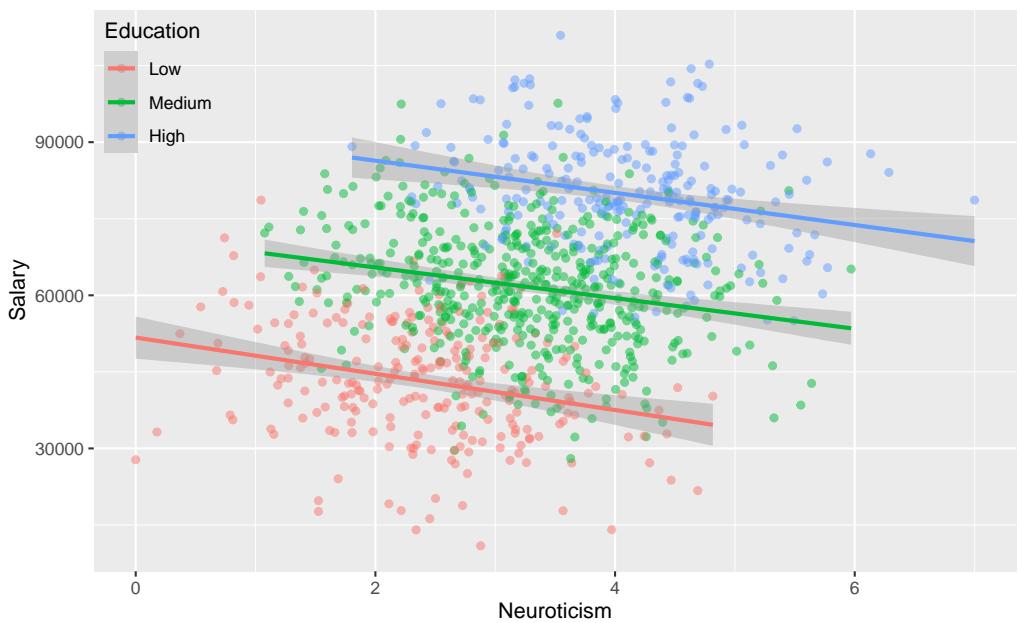
$$y = w_0 + w_1 x + w_2 \log x + \epsilon \quad (8.42)$$

8.8

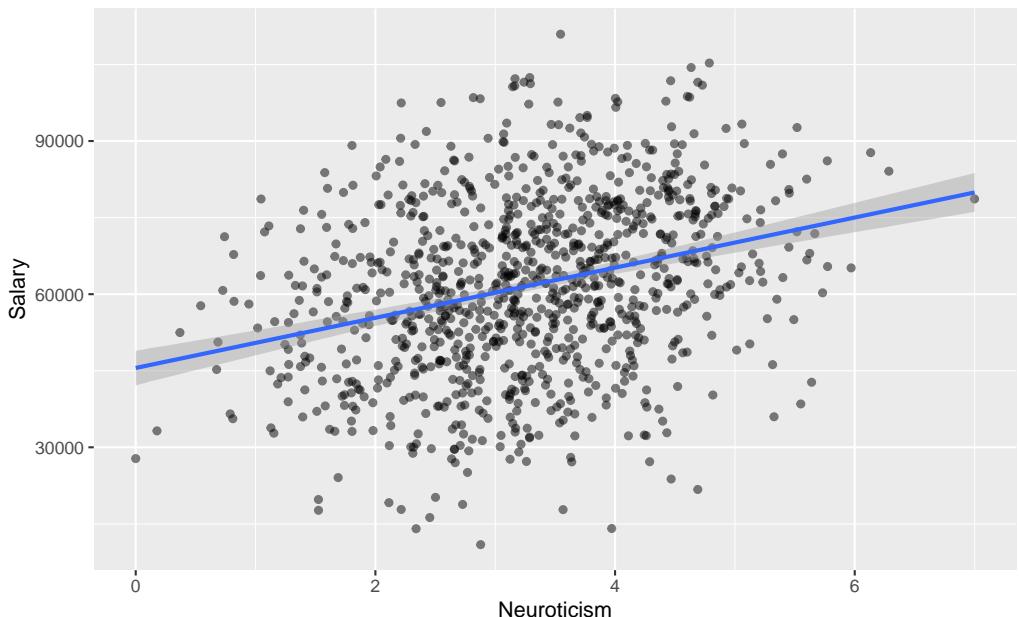
Nghịch lý Simpson

- Biến *Neuroticism* có tác động âm với biến *Salary* trong từng nhóm *Education*

8. Phân tích hồi quy bội



- Biên *Neuroticism* có tác động dương với biến *Salary*



Bài tập

B 8.1. Bảng dữ liệu sau đây với Y (thu nhập/đầu người tính bằng USD), X_1 (Tỷ lệ phần trăm lao động nông nghiệp), X_2 (Số năm trung bình được đào tạo đối với những người trên 25 tuổi). Dựa vào dữ liệu mẫu hãy tìm hàm hồi quy mẫu?

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

8. Phân tích hồi quy bội

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	6	8	8	7	7	12	9	8	9	10	10	11	9	10	11
X_1	9	10	8	7	10	4	5	5	6	8	7	4	9	5	8
X_2	8	13	11	10	12	16	10	10	12	14	12	16	14	10	12

1. Tìm ước lượng phương sai của yếu tố ngẫu nhiên.
2. Tìm ước lượng phương sai của các hệ số hồi quy mẫu.
3. Xác định hệ số xác định R^2 và hệ số hiệu chỉnh.
4. Tìm khoảng tin cậy của các hệ số hồi quy với mức ý nghĩa $\alpha = 5\%$
5. Kiểm định giả thiết đồng thời, $H_0 : \beta_1 = \beta_2 = 0$. Cho biết ý nghĩa của kết quả?

B 8.2. Khi hồi quy biến sản lượng theo *lao động* (L người), và thấy hệ số xác định ($R^2 = .3029$) khá nhỏ của mô hình S phụ thuộc vào L , nên người ta đưa thêm biến K là vốn (triệu đồng) vào và hồi quy được kết quả dưới đây

Method: Ordinary Least Squares Estimation				
Dependent variable: S				
20 observations				
Variable	Coefficient	Standard Error	t-statistic	Prob
C	-20.6583	22.0029	-0.93889	.36
K	10.7720	2.1599	4.9874	.000
L	17.2232	4.5279	3.8038	.001
R-Squared:		F-statistic: 21.5343		
Adj-R-Squared: .68369		SE. of Regression: 32.4717		
Residual Sum of Squares: 17925.5		Mean of dependent Variable: 109.4666		
SD. of dependent Variable: 57.7367		Maximum of Log-likelihood: -96.3610		
DW-statistic: 2.3574				

1. Viết hàm hồi quy tổng thể, hồi quy mẫu?
2. Các ước lượng nhận được có phù hợp về lý thuyết không?
3. Tìm ước lượng điểm mức sản lượng khi doanh nghiệp có 20 lao động, nguồn vốn 300 triệu đồng?
4. Các giá trị ước lượng có ý nghĩa thống kê không?
5. Tính hệ số xác định bội bằng các cách?
6. Phải chăng các biến giải thích không giải thích được sự biến động của sản lượng?

8. Phân tích hồi quy bội

7. Có thể nói vốn, lao động cùng tác động thuận chiều đến sản lượng?
8. Khi lao động không đổi, nếu thêm vốn một triệu đồng thì sản lượng tăng trong khoảng nào?
9. Có thể nói khi lao động không đổi, tăng vốn thêm một triệu đồng thì sản lượng tăng ít hơn 10 đơn vị được không?
10. Nguồn vốn không đổi, thêm một lao động thì sản lượng tăng có bằng 20 đơn vị không?

B 8.3. Mô hình hồi quy số lao động ngành CNTT (L) phụ thuộc lương bình quân/tháng của các ngành khác (RW), lương bình quân/tháng ngành CNTT (IW), chi phí đào tạo (C), mức tăng trưởng (G) và số lao động của thời kỳ trước (L_{t-1}). Kết quả hàm hồi quy mẫu với $R^2 = 0.984062, n = 26$

$$\begin{array}{l}
 L_t = 11.53214 - 1.822161RW_t + 2.332665IW_t - 1.4414C_t + 1.8843G_t + 0.86L_{t-1} \\
 se \quad \quad \quad (6.502) \quad \quad \quad (0.6006) \quad \quad \quad (1.555862) \quad \quad \quad (0.683097) \quad \quad \quad (1.25178) \quad \quad \quad (0.086365) \\
 t \quad \quad \quad (1.7734) \quad \quad \quad (-3.033898) \quad \quad \quad (1.499275) \quad \quad \quad (-2.553246) \quad \quad \quad (1.4973) \quad \quad \quad (10.01914)
 \end{array}$$

1. Viết hàm hồi quy tổng thể (PRF) và hàm hồi quy tổng thể ngẫu nhiên.
2. Viết hàm hồi quy mẫu (SRF) và hàm hồi quy mẫu ngẫu nhiên.
3. Hàm hồi quy có phù hợp không?
4. Giải thích ý nghĩa của các hệ số hồi quy và R^2 ?
5. Có thể cho rằng chi phí đào tạo ngành CNTT tăng thì lao động ngành CNTT giảm?
6. Trong điều kiện các yếu tố khác không đổi, chi phí đào tạo giảm 100USD thì lao động ngành CNTT tăng lên trong khoảng nào?
7. Trong điều kiện các yếu tố khác không đổi, mức tăng trưởng ngành CNTT thay đổi thì số lượng lao động ngành CNTT thay đổi như thế nào?
8. Trong điều kiện các yếu tố khác không đổi, IW tăng 100USD thì L tăng tối đa bao nhiêu?
9. Trong điều kiện các yếu tố khác không đổi, nếu C tăng 200USD thì L giảm tối đa bao nhiêu lao động?
10. Có thể cho rằng IW giảm 500USD thì L giảm 100 lao động được không?
11. Có thể cho rằng khi C giảm 200USD thì L tăng nhiều hơn 250 lao động được không?
12. Số lao động ngành CNTT năm trước tăng thì số lao động CNTT năm sau có thực sự tăng lên không?

8. Phân tích hồi quy bội

B 8.4. Mô hình hồi quy giá nhà (PRICE tính bằng nghìn đô) phụ thuộc vào các biến SQFT (diện tích nhà) BEDRMS (số phòng ngủ) BATHS (số phòng tắm).

i	PRICE	SQFT	BEDRMS	BATHS
1	199.9	1065	3	1.75
2	228	1254	3	2
3	235	1300	3	2
4	285	1577	4	2.5
5	239	1600	3	2
6	293	1750	4	2
7	285	1800	4	2.75
8	365	1870	4	2
9	295	1935	4	2.5
10	290	1948	4	2
11	385	2254	4	3
12	505	2600	3	2.5
13	425	2800	4	3
14	415	3000	4	3

1. Hồi quy mẫu

$$\text{PRICE}_i = \beta_0 + \beta_1 \text{SQFT}_i + \beta_2 \text{BEDRMS}_i + \beta_3 \text{BATHS}_i + \epsilon_i$$

và giải thích ý nghĩa của các hệ số hồi quy?

2. Tính ước lượng điểm giá PRICE trung bình khi SQFT tăng thêm 300 và BEDRMS tăng thêm 1
3. Tính hệ số xác định và hệ số xác định điều chỉnh, giải thích ý nghĩa?
4. Tìm khoảng tin cậy đối với β_1, β_2 và β_3 với mức ý nghĩa 5%?
5. Phải chăng cả ba biến đều không ảnh hưởng đến giá PRICE?
6. Dự báo giá trị trung bình và giá trị cá biệt về giá nhà trung bình khi SQRT = 2500, BDRMS = 3 và BATHS = 2?

Chọn mô hình

B 8.5. Khảo sát tại 12 công ty trong một khu vực bán hàng về doanh thu (Y), chi phí quảng cáo (X_1) và lương nhân viên tiếp thị (X_2) có số liệu sau:

8. Phân tích hồi quy bội

Y	126	148	105	162	101	175	160	127	138	143	158	137
X_1	17	23	18	22	14	24	23	15	16	21	22	13
X_2	11	14	9	16	9	17	15	11	12	14	15	13

(đơn vị tính triệu đồng)

- Giả sử mối quan hệ giữa Y với X_1 và X_2 có thể biểu diễn bằng hàm hồi quy tuyến tính. Hãy ước lượng hàm này?
- Giải thích ý nghĩa của các hệ số hồi quy về mặt lý thuyết và thống kê?
- Tính hệ số xác định bội và hệ số xác định bội điều chỉnh và cho biết ý nghĩa?
- Phải chăng cả hai biến X_1 và X_2 đều không ảnh hưởng đến Y ?
- Để dự báo doanh thu ta nên dùng hàm nào trong các hàm sau đây:
 - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
 - $y_i = \beta_0 + \beta_2 x_2 + \epsilon_i$
 - $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$
- Dự báo doanh thu trung bình của một công ty có chi phí quảng cáo là 23 triệu đồng/tháng và lương của nhân viên tiếp thị là 15 triệu đồng/tháng với mức ý nghĩa 5%?

Dạng Log-Log

B 8.6. Cho kết quả hồi quy với Y là *sản lượng*, K là *vốn*, L là *lao động*

Dependent Variable: log(Y)				
Method: Least squares				
Included observations: 20				
Variable	Coefficient	Std. Error	t-Statistic	Prob
C	0.764682	0.713780	1.071314	0.2990
$\log(K)$	0.510023	0.126959	4.017220	0.0009
$\log(L)$	0.599932	0.248400	2.415183	0.0273
R-Squared: 0.910215		Mean dependent var: 6.298380		
Adjusted Squared: 0.899652		S.D. dependent var: 0.180753		
S.E. of regression: 0.057258		F-statistic: 86.17070		
Sum squared resid: 0.055735		Prob (F- statistic): 0.00000		

- Viết hàm hồi quy tổng thể, hàm hồi quy mẫu với các biến Y , K , L và giải thích ý nghĩa kết quả ước lượng các hệ số hồi quy?
- Phải chăng cả hai biến độc lập đều giải thích cho sự biến động của biến phụ thuộc?
- Khi vốn tăng thêm 1%, lao động không đổi thì sản lượng tăng tối đa bao nhiêu?

8. Phân tích hồi quy bội

4. Khi lao động tăng thêm 1%, vốn không đổi thì sản lượng tăng tối thiểu bao nhiêu?
5. Khi vốn và lao động cùng tăng 1% thì sản lượng thay đổi như thế nào?
6. Tăng vốn 1% đồng thời giảm lao động 1% thì sản lượng có thay đổi không?
7. Có thể cho rằng quá trình sản xuất có hiệu quả tăng theo quy mô hay không?
8. Khi bỏ biến $\log(L)$ khỏi mô hình thì hệ số xác định còn 0.8794 và tổng bình phương phần dư bằng 0.07486. Vậy có nên bỏ biến đó không?

Biến giả

B 8.7. Để nghiên cứu nhu cầu của một loại hàng người ta tiến hành khảo sát *giá cả* và *lượng hàng* bán được ở 20 *khu vực* bán hàng và thu được các số liệu cho trong bảng dưới đây

y_i	x_i	d_i	y_i	x_i	d_i
20	2	1	14	5	0
19	3	0	14	6	1
18	3	1	13	6	0
18	4	0	12	7	1
17	4	1	12	7	0
17	3	1	15	5	1
16	4	0	16	4	0
16	4	1	12	7	1
15	5	1	10	8	0
15	5	1	11	8	1

Trong đó: Y là *lượng hàng bán được* (tấn/tháng), X là *giá bán* (ngàn đồng/kg), $D = \{0: \text{thành phố}, 1: \text{nông thôn}\}$

1. Tìm các hàm hồi quy :

$$y_i = \alpha_0 + a_1 x_i + \epsilon_i \quad (8.43)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \eta_i \quad (8.44)$$

2. Cho biết ý nghĩa của các hệ số hồi quy β_1 và β_2 .
3. Dùng hệ số xác định bội điều chỉnh kết hợp với kiểm định giả thuyết hệ số hồi quy của biến D bằng 0 để kết luận xem có nên đưa biến D vào mô hình hay không?
4. Dùng hàm (8.26) để dự báo lượng hàng bán được trung bình của một khu vực khi giá bán là 7 ngàn đồng/kg với độ tin cậy 95%?

B 8.8. Bảng dưới đây là số liệu giả thiết về thu nhập hàng năm của giáo viên đại học (Y ngàn đô), số năm kinh nghiệm giảng dạy (X_1), giới tính (X_2 1=nam, 0=nữ)

8. Phân tích hồi quy bội

y_i	23.0	19.5	24.0	21.0	25.0	22.0	26.5	23.1	25.0	28.0	29.5	26.0	27.5	31.5	29.0
$x_{1,i}$	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
$x_{2,i}$	1	0	1	0	1	0	1	0	0	1	1	0	0	1	0

- Giới tính có ảnh hưởng đến thu nhập của giáo viên đại học hay không?
- Dự báo mức thu nhập của một giáo viên nam có số năm kinh nghiệm giảng dạy là 8 năm với độ tin cậy 95%?
- Dự báo mức thu nhập của một giáo viên nữ có số năm kinh nghiệm giảng dạy là 9 năm với độ tin cậy 98%?

Phương sai thay đổi

B 8.9. Cho các số liệu về chi tiêu cho *tiêu dùng* (Y) và *thu nhập* (X) hàng tháng của 20 hộ gia đình ở một vùng nông thôn (đơn vị: 10.000 đồng):

y_i	39.9	62.4	63.6	24.2	81.4	12.2	77.2	51	20.6	77.6
x_i	44.6	64.6	67.2	24.2	84.6	12.4	89.4	52.2	20.6	80.4

y_i	16	66.2	67	26.2	29.6	43.2	58.6	50	35.8	39.6
x_i	16.2	69	76	28.2	32.8	48.2	60.2	56.6	36.4	40.2

- Có xảy ra hiện tượng phương sai của sai số thay đổi với mô hình hồi quy đang xét hay không? (Dùng đồ thị hoặc kiểm định Park và Glejser).
- Dùng phương pháp bình phương có trọng số để ước lượng hàm hồi quy với $\sigma_i = x_i$

Đa cộng tuyến

B 8.10. Xét một tập hợp số liệu lý thuyết với 3 biến X_1, X_2, Y được cho bảng dưới đây

y_i	-10	-8	-6	-4	-2	0	2	4	6	8	10
$x_{i,1}$	1	2	3	4	5	6	7	8	9	10	11
$x_{i,2}$	1	3	5	7	9	11	13	15	17	19	21

Giả sử chúng ta muốn áp dụng mô hình hồi quy tuyến tính Y phụ thuộc vào X_1 và X_2

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

- Chúng ta có thể ước lượng 3 tham số $\beta_0, \beta_1, \beta_2$ hay không? Tại sao?
- Nếu không, hàm tuyến tính nào bạn có thể ước lượng?

8. Phân tích hồi quy bội

Bài toán liên quan y tế

B 8.11. Bảng dưới đây cho khu vực (KV), số giường (X) và số nhập viện (Y) năm ngoái cho mỗi một trong 24 bệnh viện nhỏ chăm sóc cấp tính

KV	X	Y
A	19	120
A	120	3374
A	49	2244
A	100	3606
A	33	950
A	22	703

KV	X	Y
B	96	2958
B	48	1487
B	148	4700
B	101	3308
B	66	2696
B	138	4845
B	25	1159
B	193	5692
B	44	1576

KV	X	Y
C	76	2648
C	75	2757
C	84	2881
C	13	402
C	40	1600
C	69	1646
C	125	4825
C	13	370
C	32	987

1. Tìm hàm hồi quy ước lượng.
2. Giải thích ý nghĩa của các tham số.
3. Tìm một ước lượng của trung bình số nhập viện cho bệnh viện 100 giường ở các vùng B và C.
4. Trung bình nhập viện có khác nhau giữa ba khu vực cho các bệnh viện với một số giường cho sẵn? Bình luận.

B 8.12. Đối với cấy ghép phổi, tốt nhất là phổi người cho có kích thước tương tự như phổi người nhận. Tổng dung tích phổi (TLC-Total Lung Capacity) thì khó đo lường, do đó, rất hữu ích để có thể dự đoán TLC từ các thông tin khác. Bảng dưới đây cho thấy TLC trước khi cấy ghép của 32 người nhận cấy ghép tim-phổi thu được từ phép đo plethysmography toàn thân, cùng với độ tuổi giới tính, và chiều cao của họ.

8. Phân tích hồi quy bội

#	Tuổi	Giới	Cao	TLC	#	Tuổi	Giới	Cao	TLC
1	35	F	149	3.40	17	30	F	172	6.30
2	11	F	138	3.41	18	21	F	163	6.55
3	12	M	148	3.80	19	21	F	164	6.60
4	16	F	156	3.90	20	20	M	189	6.62
5	32	F	152	4.00	21	34	M	182	6.89
6	16	F	157	4.10	22	43	M	184	6.90
7	14	F	165	4.46	23	35	M	174	7.00
8	16	M	152	4.55	24	39	M	177	7.20
9	35	F	177	4.83	25	43	M	183	7.30
10	33	F	158	5.10	26	37	M	175	7.65
11	40	F	166	5.44	27	32	M	173	7.80
12	28	F	165	5.50	28	24	M	173	7.90
13	23	F	160	5.73	29	20	F	162	8.05
14	52	M	178	5.77	30	25	M	180	8.10
15	46	F	169	5.80	31	22	M	173	8.70
16	29	M	173	6.00	32	25	M	171	9.45

1. Từ một mô hình hồi quy bao gồm giới tính, tuổi và chiều cao, ta có thể dự đoán tốt đến mức nào TLC của một cá nhân?
2. So sánh kết quả thu được ở trên với kết quả xuất phát từ hồi quy tuyến tính chỉ trên một mình chiều cao mà thôi.
3. Tính khoảng tin cậy dự đoán 95% từ hồi quy tuyến tính trên chiều cao cho một người nào đó có chiều cao trung bình.
4. Làm thế nào chúng ta có thể điều tra xem liệu các mối quan hệ giữa TLC và chiều cao là giống nhau cho nam và nữ.
5. Giả sử rằng bạn đang viết một bài báo dựa trên các phân tích này, bạn sẽ viết gì trong phần “phương pháp thống kê” của bạn.

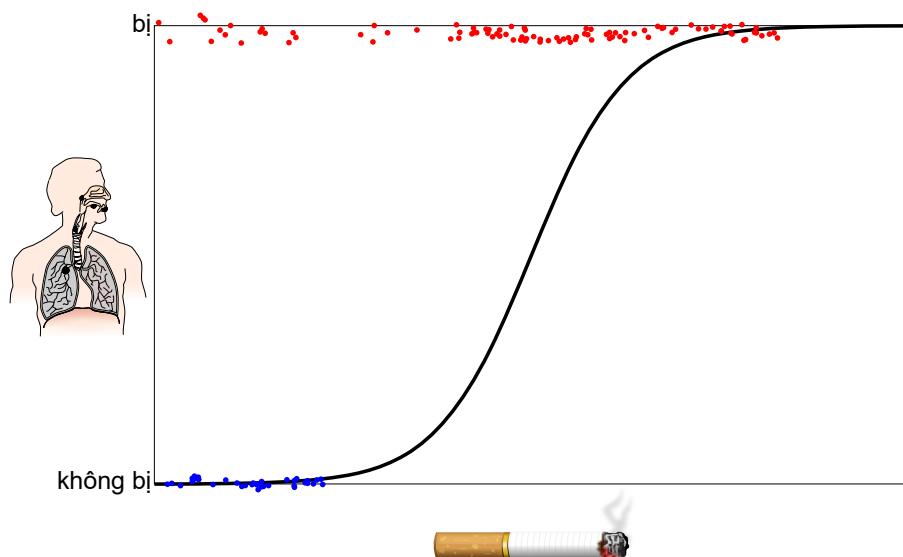
Phân tích hồi quy logistic

Định nghĩa. Phân tích hồi qui logistic (hồi quy logit) là một kỹ thuật xem xét mối liên hệ giữa biến độc lập (liên tục hoặc phân loại) với biến phụ thuộc là biến nhị phân.

Bởi vì biến phụ thuộc không phải là một biến liên tục, mà là biến nhị phân do đó các phương pháp phân tích như mô hình hồi qui tuyến tính không thể áp dụng được. Vào thập niên 70s, nhà thống kê học David R. Cox phát triển một mô hình logistic regression (logit) trong đó giả định rằng

“log odds của biến phụ thuộc là phụ thuộc tuyến tính vào các biến độc lập”

Nhiều nghiên cứu y khoa có mục tiêu là phân tích mối tương quan giữa một (hay nhiều) yếu tố nguy cơ (*risk scores*) và nguy cơ mắc bệnh (*probability of outcome*). Chẳng hạn nghiên cứu về mối tương quan giữa số điếu thuốc lá hút trung bình hàng ngày (X) và ung thư phổi (Y).



Gọi $p = Pr(Y = \text{bị} | \text{số điếu thuốc})$ là xác suất bị bệnh tương ứng với *số điếu thuốc* thì $1 - p$ là xác suất không bị bệnh; ta có công thức sau

$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 \text{số điếu thuốc} \quad (9.1)$$

hay

$$Pr(Y = \text{bị} | \text{số điếu thuốc}) = \frac{1}{1 + \exp(-w_0 - w_1 \text{số điếu thuốc})} \quad (9.2)$$

9. Phân tích hồi quy logistic

Lưu ý về miền giá trị

- Xác suất p có miền giá trị là $[0, 1]$
- odds có miền giá trị là $[0, \infty)$
- logodds có miền giá trị là $(-\infty, \infty)$

9.1

Mô hình tổng quát và Ước lượng tham số

Bài toán. Cho một tập dữ liệu huấn luyện ngẫu nhiên \mathcal{D} . Tìm các tham số sao cho cực đại giá trị likelihood

Giá trị likelihood của mô hình logistic regression tương ứng với $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ (lưu ý: $\mathbf{x}_i \in \mathbb{R}^k$ và y_i được mã hóa thành $\{0, 1\}$)

$$Pr(\mathcal{D} \mid \mathbf{w}) = \prod_{i=1}^n Pr(y_i \mid x_i; \mathbf{w}) \quad (9.3)$$

với $\mathbf{w} = \{\mathbf{w}_0, \mathbf{w}_1, \dots, \mathbf{w}_k\}$

Giá trị âm của log-likelihood

$$\begin{aligned} -LL(\mathcal{D} \mid \mathbf{w}) &= -\log Pr(\mathcal{D} \mid \mathbf{w}) \\ &= -\sum_{i=1}^n \log Pr(y_i \mid x_i; \mathbf{w}) \\ &= \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \end{aligned} \quad (9.4)$$

với

$$p_i = \frac{1}{1 + \exp(\mathbf{w}^\top \mathbf{x}_i)}$$

Không có công thức cho ước lượng vector tham số \mathbf{w}

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} -\log Pr(\mathcal{D} \mid \mathbf{w}) \quad (9.5)$$

Chúng ta sẽ sử dụng phương pháp tối ưu dựa trên Gradient để ước lượng.

- Đặt ma trận \mathbf{X}

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & x_{2,1} & \dots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{bmatrix} \quad (9.6)$$

9. Phân tích hồi quy logistic

- Đặt ma trận \mathbf{V}

$$\mathbf{V} = \begin{bmatrix} \hat{p}_1(1 - \hat{p}_1) & 0 & \dots & 0 \\ 0 & \hat{p}_2(1 - \hat{p}_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{p}_n(1 - \hat{p}_n) \end{bmatrix} \quad (9.7)$$

- Ma trận hiệp phương sai của các hệ số ước lượng

$$\begin{pmatrix} \text{Var}(\hat{w}_0) & \text{cov}(\hat{w}_0, \hat{w}_1) & \dots & \text{cov}(\hat{w}_0, \hat{w}_k) \\ \text{cov}(\hat{w}_1, \hat{w}_0) & \text{Var}(\hat{w}_1) & \dots & \text{cov}(\hat{w}_1, \hat{w}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{w}_k, \hat{w}_0) & \text{cov}(\hat{w}_k, \hat{w}_1) & \dots & \text{Var}(\hat{w}_k) \end{pmatrix} = (\mathbf{X}^\top \mathbf{V} \mathbf{X})^{-1} \quad (9.8)$$

- Phương sai của các hệ số ước lượng

$$\text{Var}(\hat{\mathbf{w}}) = \text{diag}(\mathbf{X}^\top \mathbf{V} \mathbf{X})^{-1} \quad (9.9)$$

9.2

Ví dụ hồi quy

Ví dụ. Tìm mối liên hệ thời gian ôn thi ảnh hưởng đến xác suất sinh viên vượt qua kỳ thi như thế nào? qua dữ liệu 20 sinh viên

giờ ôn	kết quả	giờ ôn	kết quả
.5	rớt	2.75	đậu
.75	rớt	3	rớt
1	rớt	3.25	đậu
1.25	rớt	3.5	rớt
1.5	rớt	4	đậu
1.75	rớt	4.25	đậu
1.75	đậu	4.5	đậu
2	rớt	4.75	đậu
2.25	đậu	5	đậu
2.5	rớt	5.5	đậu

Hồi quy logistic có 2 dạng biểu diễn

- Dạng xác suất: xác suất đậu kỳ thi theo giờ ôn tập

$$p = p(\text{giờ}) = \frac{1}{1 + \exp(-\mathbf{w}_0 - \mathbf{w}_1 \text{giờ})} \quad (9.10)$$

9. Phân tích hồi quy logistic

- Dạng tỉ lệ cơ hội: *tỉ lệ cơ hội đậu kỳ thi theo giờ ôn tập*

$$\log odds(\text{giờ}) = \log \left(\frac{p}{1-p} \right) = w_0 + w_1 \text{giờ} \quad (9.11)$$

Kết quả chạy hồi quy

Logit Regression Results						
Dep. Variable:	ket_qua	No. Observations:	20			
Model:	Logit	Df Residuals:	18			
Method:	MLE	Df Model:	1			
Date:	Sun, 13 Jun 2021	Pseudo R-squ.:	0.4208			
Time:	08:38:16	Log-Likelihood:	-8.0299			
converged:	True	LL-Null:	-13.863			
Covariance Type:	nonrobust	LLR p-value:	0.0006365			
	coef	std err	z	P> z	[0.025	0.975]
Intercept	-4.0777	1.761	-2.316	0.021	-7.529	-0.626
gio_on	1.5046	0.629	2.393	0.017	0.272	2.737

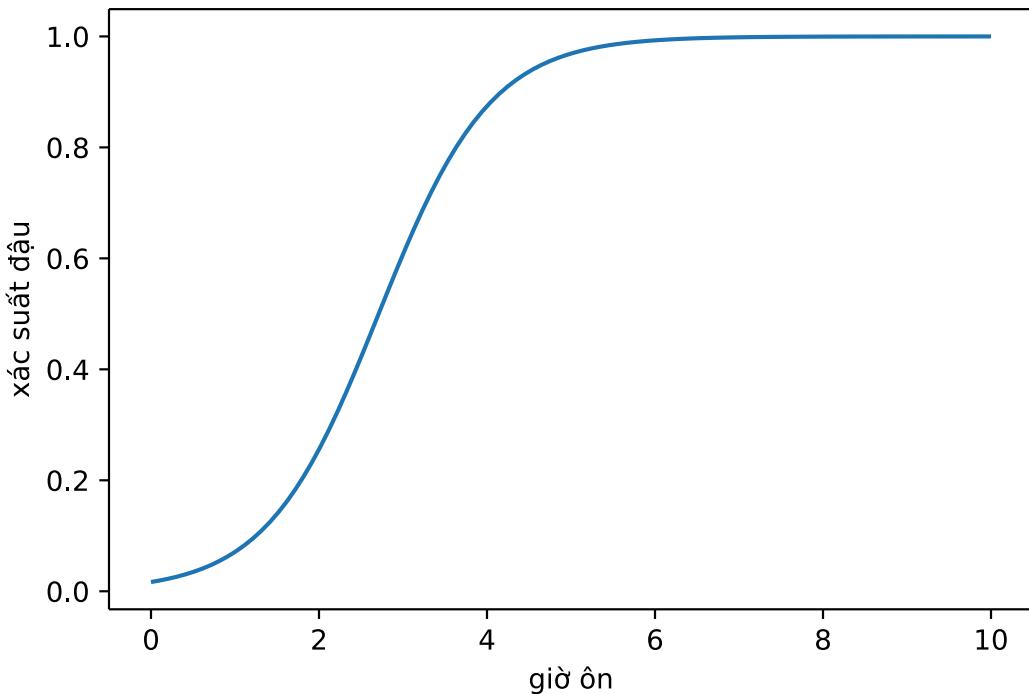
Kết luận:

- Biến **gio_on** (giờ ôn) có ý nghĩa đối với mô hình về mặt thống kê (*p-value* = 0.017)
- Mô hình giải thích được 42% sự thay đổi của biến **ket_qua** (kết quả) (*pseudo-R²* = 0.4208)
- Phương trình hồi quy xác suất đậu kỳ thi theo giờ ôn tập

$$p = p(\text{giờ}) = \frac{1}{1 + \exp(4.0777 - 1.5046 \times \text{giờ})} \quad (9.12)$$

- Trực quan hàm xác suất

9. Phân tích hồi quy logistic



9.3

Đánh giá mô hình

Trong đánh giá hồi quy logistic người ta thường so sánh *mô hình* với *mô hình null*, mô hình không có sử dụng biến độc lập nào

$$p = \frac{1}{1 + \exp(-\mathbf{w}_0)} \quad (9.13)$$

Từ giá trị likelihood cực đại của *mô hình* và *mô hình null*, chúng ta có thể ước lượng pseudo- R^2

$$\text{pseudo-}R^2 = 1 - \frac{LL(\mathcal{D} | \mathbf{w}_0, \mathbf{w}_1)}{LL(\mathcal{D} | \mathbf{w}_0)} \quad (9.14)$$

Lưu ý, trong hồi quy logistic không có khái niệm về R^2 như trong hồi quy tuyến tính và có rất nhiều định nghĩa khác nhau về pseudo- R^2 .

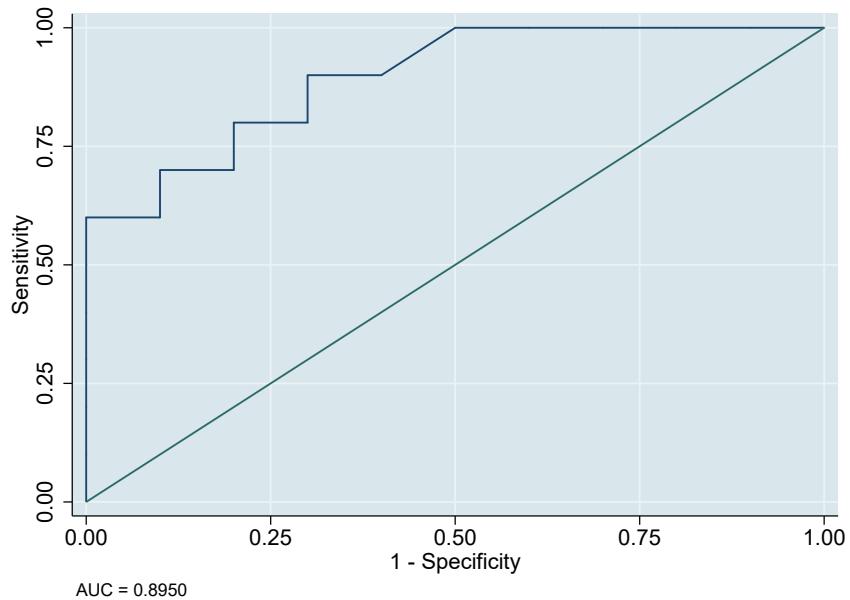
9.3.1. Đường cong ROC

Định nghĩa. Mỗi điểm (x, y) trên đường cong ROC (*receiver operating characteristic*) tương ứng với xác suất dương tính thật (*độ nhạy*) trên trực tung và xác suất dương tính giả ($1 - \text{độ đặc hiệu}$) trên trực hoành.

Định nghĩa. Phần diện tích dưới đường cong ROC được gọi là AUC (*area under the roc curve*) dùng để đánh giá mô hình

9. Phân tích hồi quy logistic

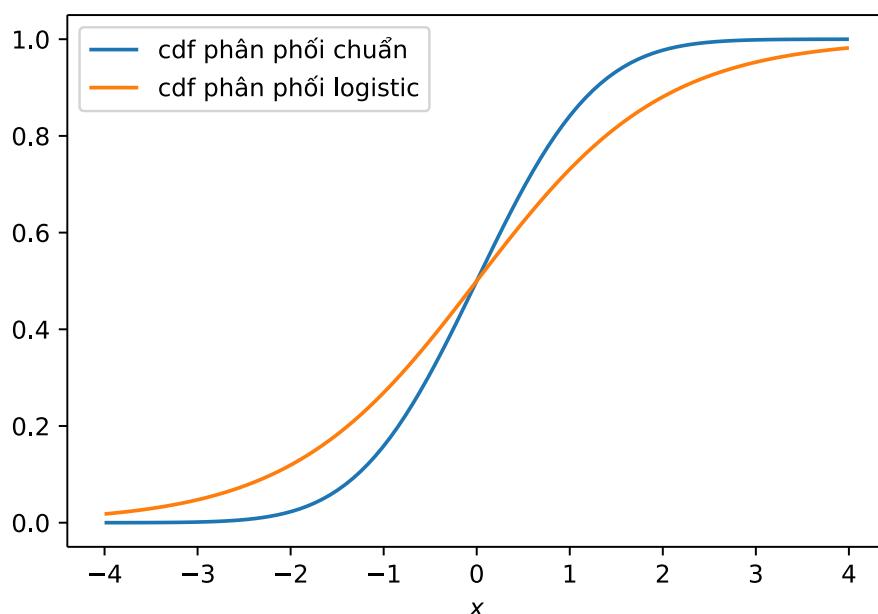
AUC	Đánh giá
0.9-1.0	rất tốt
0.8-0.9	tốt
0.6-0.8	tạm được
0.5-0.6	không giá trị



9.4

Mô hình hồi quy Probit

Với biến phụ thuộc là nhị phân các nhà người ta còn sử dụng mô hình hồi quy probit; giả định là phân phối của $p(y | \mathbf{x}; \mathbf{w})$ tuân theo phân phối chuẩn. Các mô hình logit và probit nói chung cho các kết quả tương tự.



9. Phân tích hồi quy logistic

Bài tập

B 9.1 (Nghiên cứu về vai trò của phụ nữ trong xã hội). Trong một điều tra xã hội thực hiện vào năm 1971-1972, các nhà nghiên cứu hỏi đối tượng, nam và nữ, đồng ý hay không đồng ý với câu hỏi sau đây

“Phụ nữ nên lo việc nhà và để việc điều hành nhà nước cho đàn ông” (Harberman SJ.

The analysis of residuals in cross-classified tables. Biometrics 1973;29:205-220).

Các nhà nghiên cứu ghi nhận trình độ học vấn và giới của mỗi đối tượng. Kết quả nghiên cứu có thể tóm lược bằng bảng số liệu sau đây

edu	sex	agree	disagree
0	Male	4	2
1	Male	2	0
2	Male	4	0
3	Male	6	3
4	Male	5	5
5	Male	13	7
6	Male	25	9
7	Male	27	15
8	Male	75	49
9	Male	29	29
10	Male	32	45
11	Male	36	59
12	Male	115	245
13	Male	31	70
14	Male	28	79
15	Male	9	23
16	Male	15	110
17	Male	3	29
18	Male	1	28
19	Male	2	13
20	Male	3	20

edu	sex	agree	disagree
0	Female	4	2
1	Female	1	0
2	Female	0	0
3	Female	6	1
4	Female	10	0
5	Female	14	7
6	Female	17	5
7	Female	26	16
8	Female	91	36
9	Female	30	35
10	Female	55	67
11	Female	50	62
12	Female	190	403
13	Female	17	92
14	Female	18	81
15	Female	7	34
16	Female	13	115
17	Female	3	28
18	Female	0	21
19	Female	1	2
20	Female	2	4

Biến **edu** là trình độ học vấn (đo bằng số năm theo học) của người trả lời, **agree** và **disagree** là số đối tượng đồng ý hay không đồng ý với câu hỏi. Hãy phân tích tỉ lệ đồng ý theo trình độ học vấn và giới tính?

Phân tích phương sai (ANOVA)

Phân tích phương sai (*analysis of variance*) hay còn gọi là kiểm định ANOVA là một kiểm định sự giống nhau của 3 (hoặc nhiều hơn) trung bình tổng thể bằng cách phân tích phương sai mẫu.

10.1

Phân tích phương sai một yếu tố (one-way ANOVA)

Ví dụ. Bảng dữ liệu dưới đây so sánh độ galactose trong 3 nhóm bệnh nhân: nhóm 1 gồm 9 bệnh nhân với bệnh Crohn; nhóm 2 gồm 11 bệnh nhân với bệnh viêm ruột kết (colitis); và nhóm 3 gồm 20 đối tượng không có bệnh (gọi là nhóm đối chứng). Câu hỏi đặt ra là độ galactose trung bình giữa 3 nhóm bệnh nhân có khác nhau hay không?

	Nhóm 1	Nhóm 2	Nhóm 3
1343	1264	1809 2850	
1393	1314	1926 2964	
1420	1399	2283 2973	
1641	1605	2384 3171	
1897	2385	2447 3257	
2160	2511	2479 3271	
2169	2514	2495 3288	
2279	2767	2525 3358	
2890	2827	2541 3643	
	2895	2769 3657	
	3011		
Số bệnh nhân	9	11	20
Trung bình	1910	2226	2804
Độ lệch chuẩn	516	727	527

Gọi giá trị galactose trung bình của ba nhóm là μ_1 , μ_2 , và μ_3 , và chúng ta có bài toán kiểm định là

- Giả thuyết $H_0 : \mu_1 = \mu_2 = \mu_3$

10. Phân tích phương sai (ANOVA)

- Đối thuyết H_a : có một sự khác biệt giữa $3 \mu_j, (j = 1, 2, 3)$

Do phương pháp kiểm định t-test chỉ có thể so sánh trung bình của 2 nhóm cho nên chúng ta cần một phương pháp thích hợp có thể so sánh cho nhiều nhóm cùng một lúc và đó chính là phương pháp phân tích phương sai, một dạng đặc biệt của phân tích hồi tuyến tính với biến phụ thuộc là biến định lượng và biến độc lập là các biến định tính.

10.1.1. Mô hình phân tích phương sai

Chúng ta gọi độ galactose của bệnh nhân i thuộc nhóm j ($j = 1, 2, 3$) là x_{ij} . Mô hình phân tích phương sai được viết bằng phương trình

$$x_{ij} = \mu + \alpha_j + \epsilon_{ij} \quad (10.1)$$

hay

$$\begin{cases} x_{i1} &= \mu + \alpha_1 + \epsilon_{i1} \\ x_{i2} &= \mu + \alpha_2 + \epsilon_{i2} \\ x_{i3} &= \mu + \alpha_3 + \epsilon_{ij} \end{cases} \quad (10.2)$$

Nghĩa là giá trị galactose của bất cứ bệnh nhân nào bằng giá trị trung bình của toàn quần thể μ cộng/trừ cho ảnh hưởng của nhóm j được đo bằng hệ số ảnh hưởng α_j , và sai số ϵ_{ij} . Giả định khác là các ϵ_{ij} có cùng phân phối chuẩn $\mathcal{N}(0, \sigma^2)$.

Cũng như phân tích hồi qui tuyến tính chúng ta phải tìm

$$\hat{\mu}, \hat{\alpha}_j = \arg \min_{\mu, \alpha_j} \sum (x_{ij} - \mu + \alpha_j)^2 \quad (10.3)$$

Tuy nhiên, vì đây là phân tích kiểm định cho nên chúng ta sẽ phân tích giá trị x_{ij} như sau

$$x_{ij} = \bar{x} + (\bar{x}_j - \bar{x}) + (x_{ij} - \bar{x}_j), \quad (10.4)$$

trong đó,

- \bar{x} là số trung bình của toàn mẫu
- \bar{x}_j là số trung bình của nhóm j .

Nói cách khác, phần $(\bar{x}_j - \bar{x})$ phản ánh độ khác biệt giữa trung bình từng nhóm và trung bình toàn mẫu, phần $(x_{ij} - \bar{x}_j)$ phản ánh khác biệt của một đối tượng và số trung bình của từng nhóm.

Xem lại bảng số liệu trên

10. Phân tích phương sai (ANOVA)

	Nhóm 1	Nhóm 2	Nhóm 3	Toàn bộ
Số đối tượng	$n_1 = 9$	$n_2 = 11$	$n_3 = 20$	$n = 40$
Trung bình	$\bar{x}_1 = 1910$	$\bar{x}_2 = 2226$	$\bar{x}_3 = 2804$	$\bar{x} = 2444$
Phương sai	$s_1^2 = 265944$	$s_2^2 = 473387$	$s_3^2 = 277500$	

- Tổng bình phương cho toàn bộ mẫu là

$$\begin{aligned} TSS &= \sum_i \sum_j (x_{ij} - \bar{x})^2 \\ &= 12133923 \end{aligned} \tag{10.5}$$

- Tổng bình phương vì khác nhau giữa các nhóm

$$\begin{aligned} BSS &= \sum_i \sum_j (\bar{x}_j - \bar{x})^2 \\ &= \sum_j n_j (\bar{x}_j - \bar{x})^2 \\ &= 5681168 \end{aligned} \tag{10.6}$$

- Tổng bình phương vì dao động trong mỗi nhóm

$$\begin{aligned} WSS &= \sum_i \sum_j (x_{ij} - \bar{x}_j)^2 \\ &= \sum_j (n_j - 1) s_j^2 \\ &= 12133922 \end{aligned} \tag{10.7}$$

Lưu ý, ta có

$$TSS = BSS + WSS \tag{10.8}$$

- Trung bình bình phương cho từng nhóm (có n đối tượng và k nhóm)

$$\begin{aligned} WMS &= \frac{WSS}{n - k} \\ &= 327944 \end{aligned} \tag{10.9}$$

- Trung bình bình phương giữa các nhóm

$$\begin{aligned} BMS &= \frac{BSS}{k - 1} \\ &= 2841810 \end{aligned} \tag{10.10}$$

10. Phân tích phương sai (ANOVA)

Nếu có sự khác biệt giữa các nhóm, thì chúng ta kì vọng rằng BMS sẽ lớn hơn WMS . Thành ra, để kiểm tra giả thiết, chúng ta có thể dựa vào kiểm định F -test

$$F = \frac{BMS}{WMS} = 8.6655, \quad (10.11)$$

với bậc tự do $(k - 1, n - k)$. Kết quả kiểm định

Nguồn biến thiên (source of variation)	Bậc tự do (degrees of freedom)	Tổng bình phương (sum of squares)	Trung bình bình phương (mean square)	Kiểm định F
Khác biệt giữa các nhóm (between-group)	2	5681168	2841810	8.6655 (0.00082)
Khác biệt trong từng nhóm (within-group)	37	12133923	327944	
Tổng số	39	12133923		

Giá trị $p = 0.00082$ có nghĩa là tín hiệu cho thấy có sự khác biệt về độ galactose giữa ba nhóm.

10.2 Phân tích phương sai hai yếu tố (two-way ANOVA)

Phương pháp phân tích phương sai hai yếu tố đơn giản là sự mở rộng của phương pháp phân tích phương sai một yếu tố. Có hai loại

1. Phân tích phương sai 2 yếu tố không lặp
2. Phân tích phương sai 2 yếu tố lặp

Chúng ta chỉ xem xét loại thứ hai

Ví dụ. Để đánh giá hiệu quả của một kỹ thuật sơn mới, các nhà nghiên cứu áp dụng sơn trên 3 loại vật liệu $\{1, 2, 3\}$ trong 2 điều kiện $\{1, 2\}$. Mỗi điều kiện và loại vật liệu, nghiên cứu được lặp lại 3 lần. Độ bền được đo là chỉ số bền bỉ (tạm gọi là score). Câu hỏi đặt ra là độ bền có ảnh hưởng bởi điều kiện và vật liệu thí nghiệm hay không?

		Vật liệu		
		1	2	3
Điều kiện	1	4.1, 3.9, 4.3	3.1, 2.8, 3.3	3.5, 3.2, 3.6
	2	2.7, 3.1, 2.6	1.9, 2.2, 2.3	2.7, 2.3, 2.5

10.2.1. Mô hình phân tích phương sai

Gọi x_{ij} là score của điều kiện i ($i = 1, 2$) cho vật liệu j ($j = 1, 2, 3$). Mô hình phân tích phương sai hai chiều phát biểu rằng

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (10.12)$$

10. Phân tích phương sai (ANOVA)

trong đó μ là số trung bình cho toàn quần thể, các hệ số α_i (ảnh hưởng của điều kiện i) và β_j (ảnh hưởng của vật liệu j) cần phải ước tính từ số liệu thực tế. Giả định khác là các ϵ_{ij} có cùng phân phối chuẩn $\mathcal{N}(0, \sigma^2)$ và không có sự tương tác giữa các yếu tố.

- Trung bình cho từng nhóm *điều kiện* và *vật liệu*, trung bình cho nhóm *điều kiện*, trung bình cho nhóm *vật liệu*

		Vật liệu			trung bình vật liệu
		1	2	3	
Điều kiện	1	$x_{11} = 4.10$	$x_{12} = 3.07$	$x_{13} = 3.43$	$\bar{x}_{1*} = 3.533$
	2	$x_{21} = 2.80$	$x_{22} = 2.13$	$x_{23} = 2.50$	$\bar{x}_{2*} = 2.478$
trung bình điều kiện		$\bar{x}_{*1} = 3.450$	$\bar{x}_{*2} = 2.600$	$\bar{x}_{*3} = 2.967$	$\bar{x} = 3.00$

- Phương sai cho từng nhóm *điều kiện* và *vật liệu*

		Vật liệu			
		1	2	3	
Điều kiện	1	$s_{11}^2 = 0.040$	$s_{12}^2 = 0.063$	$s_{13}^2 = 0.043$	
	2	$s_{21}^2 = 0.070$	$s_{22}^2 = 0.043$	$s_{23}^2 = 0.040$	

- Số lượng đổi tương của từng nhóm *điều kiện* và *vật liệu*

		Vật liệu			
		1	2	3	
Điều kiện	1	$n_{11} = 3$	$n_{12} = 3$	$n_{13} = 3$	$n_{1*} = 9$
	2	$n_{21} = 3$	$n_{22} = 3$	$n_{23} = 3$	$n_{2*} = 9$
		$n_{*1} = 6$	$n_{*2} = 6$	$n_{*3} = 6$	$n = 18$

Trong phân tích phương sai hai chiều, chúng ta cần chia tổng bình phương ra thành 3 nguồn

- Nguồn thứ nhất là tổng bình phương do biến đổi giữa các *điều kiện*

$$\begin{aligned} SS_c &= \sum_i n_{i*} (\bar{x}_{i*} - \bar{x})^2 \\ &= 5.01 \end{aligned} \tag{10.13}$$

- Nguồn thứ hai là tổng bình phương do biến đổi giữa các *vật liệu*

$$\begin{aligned} SS_m &= \sum_j n_{*j} (\bar{x}_{*j} - \bar{x})^2 \\ &= 2.18 \end{aligned} \tag{10.14}$$

- Nguồn thứ ba là tổng bình phương phần dư (residual sum of squares)

10. Phân tích phương sai (ANOVA)

$$\begin{aligned}
 SS_e &= \sum_i \sum_j (x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2 \\
 &= \sum_i \sum_j (n_{ij} - 1) s_{ij}^2 \\
 &= 0.73
 \end{aligned} \tag{10.15}$$

Gọi p là số giá trị điều kiện và q là số giá trị vật liệu. Ta có SS_c có $p - 1$ bậc tự do, SS_m có $q - 1$ bậc tự do, và SS_e có $n - pq + 2$ bậc tự do. Do đó, các trung bình bình phương

- Cho điều kiện

$$\begin{aligned}
 MS_c &= \frac{SS_c}{p - 1} \\
 &= 5.01
 \end{aligned} \tag{10.16}$$

- Cho vật liệu

$$\begin{aligned}
 MS_m &= \frac{SS_m}{q - 1} \\
 &= 1.09
 \end{aligned} \tag{10.17}$$

- Cho phần dư

$$\begin{aligned}
 MS_e &= \frac{SS_e}{n - pq + 2} \\
 &= 0.052
 \end{aligned} \tag{10.18}$$

Do đó, so sánh độ khác biệt giữa các điều kiện dựa vào kiểm định

$$F = \frac{MS_c}{MS_e} \tag{10.19}$$

với bậc tự do $(p - 1, n - pq + 2)$ và giữa các vật liệu có thể dựa vào kiểm định

$$F = \frac{MS_m}{MS_e} \tag{10.20}$$

với bậc tự do $(q - 1, n - pq + 2)$. Kết quả chạy kiểm định ta có

10. Phân tích phương sai (ANOVA)

Nguồn biến thiên (source of variation)	Bậc tự do (degrees of freedom)	Tổng bình phương (sum of squares)	Trung bình bình phương (mean square)	Kiểm định F
Điều kiện	1	5.01	5.01	95.575 (1.235e-07)
Vật liệu	2	2.18	1.09	20.788 (6.437e-05)
Phần dư	14	0.73	0.052	
Tổng số	17	7.92		

Ba nguồn biến thiên của score được phân tích trong bảng trên. Qua trung bình bình phương (mean square), chúng ta thấy ảnh hưởng của điều kiện có vẻ quan trọng hơn là ảnh hưởng của vật liệu thí nghiệm. Tuy nhiên, cả hai ảnh hưởng đều có ý nghĩa thống kê, vì trị số p rất thấp cho hai yếu tố.

- Để cho hoàn chỉnh ta có thể xét thêm sự tương tác giữa hai yếu tố điều kiện và vật liệu

Bài tập

Phân tích phương sai một yếu tố

B 10.1. So sánh 3 loại thuốc bổ A, B, C trên 3 nhóm, người ta được kết quả tăng trọng (kg) như sau

A	1.0	1.2	1.4	1.1	0.8	0.6
B	2.0	1.8	1.9	1.2	1.4	1.0
C	0.4	0.6	0.7	0.2	0.3	0.1

Hãy so sánh kết quả tăng trọng của 3 loại thuốc bổ trên với mức ý nghĩa $\alpha = 0.01$

B 10.2. Một nhà máy thủy điện sử dụng các turbines được giải nhiệt bằng nước. Nếu nước được dung để giải nhiệt bị ô nhiễm thì hệ thống máy móc sẽ bị xói mòn. Do đó, người ta sử dụng các máy lọc để làm giảm mức ô nhiễm của nước. Giám đốc nhà máy muốn trắc nghiệm tính hiệu quả của 4 máy lọc đang sử dụng. Ở mỗi máy lọc người ta lấy ngẫu nhiên độc lập nhau 3 mẫu nước đã được lọc và đo mức độ ô nhiễm. Các kết quả có được như sau

Máy lọc 1	Máy lọc 2	Máy lọc 3	Máy lọc 4
10	11	13	23
9	16	8	18
5	9	9	25

B 10.3. Một nghiên cứu được thực hiện để so sánh tuổi thọ (giờ) của 4 nhãn hiệu Pin: A, B, C, D. Kết quả ghi nhận được như sau

10. Phân tích phương sai (ANOVA)

Hiệu A	Hiệu B	Hiệu C	Hiệu D
15	14	19	16
16	15	20	15
18	16	16	16
20	15	13	18
19	14	17	
20			

Yêu cầu: Giả định tuổi thọ pin có phân phối chuẩn, phương sai bằng nhau. Với phương pháp ANOVA, ở mức ý nghĩa $\alpha = 0.05$, có thể kết luận rằng tuổi thọ trung bình của 4 nhãn hiệu pin là không khác nhau được không?

B 10.4. Ba mẫu thiết kế bao bì của một loại sản phẩm được xem xét bằng cách thu thập doanh số (triệu đồng/tuần) của mỗi loại bao bì trong một mẫu ngẫu nhiên các cửa hàng. Kết quả được ghi nhận trong bảng sau

Mẫu bao bì I	Mẫu bao bì II	Mẫu bao bì III
18	24	19
16	25	24
29	21	24
26	31	28
29	22	15
14		29
12		32
23		

Với kiểm định ANOVA ở mức ý nghĩa $\alpha = 0.01$, có thể kết luận rằng các mẫu bao bì không ảnh hưởng đến doanh số được không? (Giả định doanh số theo các mẫu bao bì có phân phối chuẩn, phương sai bằng nhau).

B 10.5. Một nhà sản xuất muốn kiểm tra xem 3 máy có công suất khác nhau không. Ông ta chỉ định ngẫu nhiên 15 công nhân được đào tạo cùng một phương pháp làm việc trên 3 máy (5 người/1 máy). Với mức rủi ro 5%, liệu 3 máy có công suất khác nhau?

Máy 1	Máy 2	Máy 3
25.40	26.31	24.10
23.74	25.10	23.40
21.80	23.50	22.75
21.60	20.00	22.20
19.75	20.60	20.40

B 10.6. Để so sánh hiệu năng của 3 loại thuốc diệt muỗi A, B, C người ta thực hiện một thực nghiệm như sau: Có 21 thùng, mỗi thùng nhốt vài trăm con muỗi. Chia ngẫu nhiên các thùng

10. Phân tích phương sai (ANOVA)

này thành 3 nhóm, mỗi nhóm 7 thùng. Muỗi ở trong mỗi nhóm thùng được xịt một loại thuốc khác nhau A, B hoặc C, tỉ lệ % muỗi chết được ghi nhận như sau

Thuốc diệt muỗi A	Thuốc diệt muỗi B	Thuốc diệt muỗi C
68	58	71
80	60	62
69	70	58
76	51	74
68	57	65
77	71	59
60	61	57

Với kiểm định ANOVA ở mức ý nghĩa $\alpha = 0.05$, có thể nói khả năng diệt muỗi (thể hiện thông qua tỉ lệ muỗi chết trung bình) của 3 loại thuốc là như nhau được không? (giả định muỗi chết có phân phối chuẩn, phương sai bằng nhau).

B 10.7. Trưởng phòng kỹ thuật của một nhà máy sản xuất vỏ xe thực hiện một nghiên cứu để đánh giá sự khác biệt về chất lượng sản phẩm giữa 3 ca sản xuất: sáng, chiều, đêm. Chọn ngẫu nhiên một số sản phẩm để kiểm tra, kết quả ghi nhận như sau

Thời gian sản xuất	Số sản phẩm	Độ bền trung bình (ngàn km)	Tổng bình phương các sai lệch
Sáng	10	25.95	6.255
Chiều	12	25.50	6.595
Tối	15	23.75	7.555

Yêu cầu: Với mức ý nghĩa $\alpha = 0.05$, có thể kết luận rằng có sự khác biệt về độ bền giữa các sản phẩm sản xuất ra ở ca sáng, ca chiều và ca đêm hay không? Nếu có, sự khác biệt đó như thế nào?

B 10.8. Bốn trạm sửa chữa và bảo hành xe Honda trong một thành phố lớn tuyên bố rằng khách hàng sẽ được phục vụ nhanh chóng ngay khi xe được đưa tới trạm. Giám đốc phụ trách dịch vụ hậu mãi của hãng tiến hành kiểm tra chất lượng dịch vụ của các trạm bảo hành, bằng cách chọn ngẫu nhiên khách hàng đến trạm trong giờ cao điểm (9 đến 11 giờ sáng) và ghi nhận thời gian chờ đợi của họ. Một phần kết quả tính toán cho trong bảng sau

Trạm bảo hành	Số khách hàng	Thời gian chờ TB (phút)	Phương sai
A	3	5.133333	0.323333
B	4	8	1.433333
C	5	5.04	0.748
D	4	6.475	0.595833

Lập bảng ANOVA. Số liệu trên có chứng tỏ rằng thời gian chờ đợi của khách hàng ở các trạm bảo hành của hãng là không khác nhau? Kết luận với mức ý nghĩa $\alpha = 0.05$.

10. Phân tích phương sai (ANOVA)

B 10.9. Một hãng sản xuất ô tô thực hiện một nghiên cứu để đo lường sự khác biệt mức nhiên liệu tiêu thụ trung bình giữa 3 loại xe: cỡ nhỏ (4 chỗ), trung bình (8 chỗ), và xe cở lớn (12 chỗ). Chọn ngẫu nhiên 27 xe, kết quả tính toán cho trong bảng sau

Loại xe	Số xe	Mức nhiên liệu tiêu thụ TB (lit/100km)	Phương sai
Nhỏ	12	8.133333	2.343333
Trung Bình	9	9.583253	2.453333
Lớn	6	10.04578	3.74853

Phân tích phương sai hai yếu tố

B 10.10. Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa và năng suất. Năng suất lúa được ghi nhận từ các thực nghiệm sau

		Giống lúa		
		A	B	C
Loại phân bón	1	65 68 62	69 71 67	75 75 78
	2	74 79 76	72 69 69	70 69 65
	3	64 72 65	68 73 75	78 82 80
	4	83 82 84	78 78 75	76 77 75

Hãy cho biết sự ảnh hưởng của loại phân bón, giống lúa trên năng suất với mức ý nghĩa $\alpha = 0.01$

B 10.11. Để khảo sát ảnh hưởng của 4 loại thuốc trừ sâu (1, 2, 3 và 4) và ba loại giống (B1, B2 và B3) đến sản lượng của cam, các nhà nghiên cứu tiến hành một thí nghiệm loại giai thừa. Trong thí nghiệm này, mỗi giống cam có 4 cây cam được chọn một cách ngẫu nhiên, và 4 loại thuốc trừ sâu áp dụng (cũng ngẫu nhiên) cho mỗi cây cam. Kết quả nghiên cứu (sản lượng cam) cho từng giống và thuốc trừ sâu như sau

		Thuốc trừ sâu			
		1	2	3	4
Giống	B1	29	50	43	53
	B2	41	58	42	73
	B3	66	85	63	85

Hãy cho biết thuốc trừ sâu, giống cam có ảnh hưởng đến sản lượng cam không? $\alpha = 0.05$

B 10.12. 4 chuyên gia tài chính được yêu cầu dự đoán về tốc độ tăng trưởng (%) trong năm tới của 5 công ty trong ngành nhựa. Dự đoán được ghi nhận như sau

10. Phân tích phương sai (ANOVA)

		Chuyên gia			
		A	B	C	D
Công ty	1	8	12	8.5	13
	2	14	10	9	11
	3	11	9	12	10
	4	9	13	10	13
	5	12	10	10	10

Hãy lập bảng ANOVA. Có thể nói rằng dự đoán tốc độ tăng trưởng trung bình là như nhau cho cả 5 công ty nhựa được không?

B 10.13. Một công ty vận chuyển thực hiện một nghiên cứu để xem xét ảnh hưởng của lộ trình đến thời gian vận chuyển (phút) giữa 2 địa điểm. Số liệu thống kê về thời gian vận chuyển của 9 chuyến trong một tuần được thực hiện trên các lộ trình và thời gian khác nhau trong ngày cho trong bảng sau

		Lộ trình		
		A	B	C
Thời gian	10 - 12 giờ sáng	50	52	54
	1 - 3 giờ chiều	45	65	62
	7 - 10 giờ tối	55	47	50

Yêu cầu: ở mức ý nghĩa 5%, hãy kết luận xem:

- Có sự khác biệt về thời gian vận chuyển trung bình giữa 3 lộ trình hay không? Nếu có, công ty nên chọn lộ trình nào?
- Có sự khác biệt về thời gian vận chuyển trung bình giữa các thời gian khác nhau trong ngày hay không? Nếu có, công ty nên thực hiện vận chuyển vào thời gian nào?

Phân tích dữ liệu chuỗi thời gian

Chuỗi thời gian (*time series*) trong kinh tế lượng và toán tài chính là chuỗi số liệu được thu thập trong một thời kì hoặc một khoảng thời gian lặp lại nhau trên cùng một đối tượng, một không gian, một địa điểm.

Phân tích chuỗi thời gian bao gồm các phương pháp để phân tích dữ liệu chuỗi thời gian, để từ đó trích xuất ra được các thuộc tính thống kê có ý nghĩa và các đặc điểm của dữ liệu.

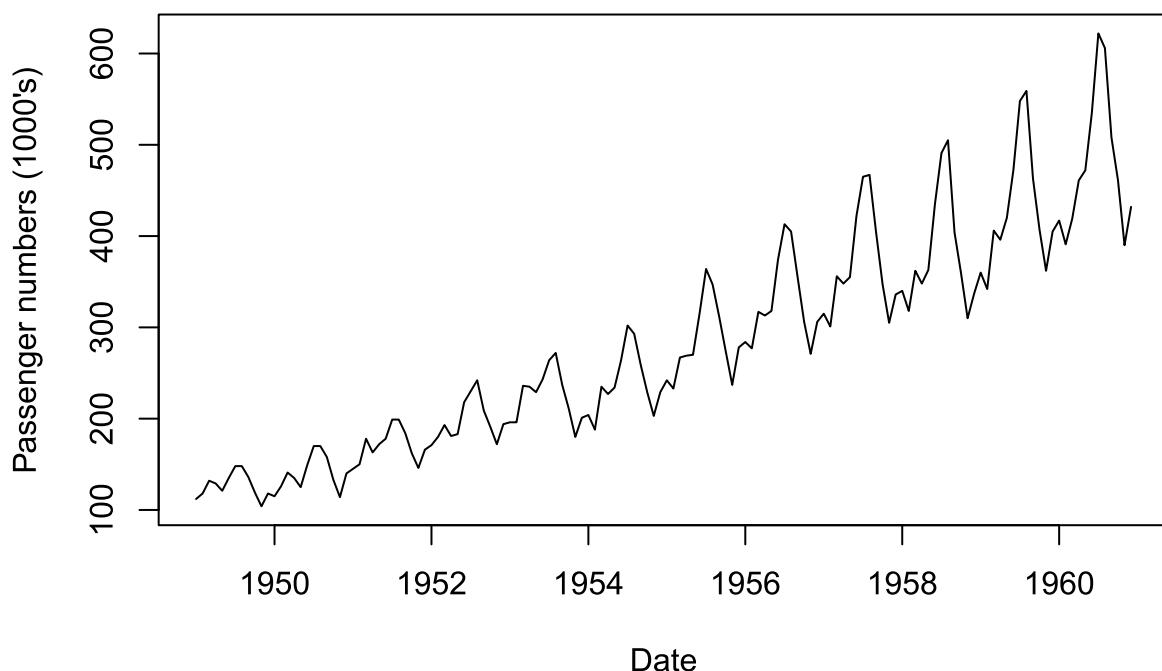
Dự đoán chuỗi thời gian là việc sử dụng mô hình để dự đoán các sự kiện thời gian dựa vào các sự kiện đã biết trong quá khứ để từ đó dự đoán các điểm dữ liệu trước khi nó xảy ra.

Ví dụ. Dữ liệu về tổng số hành khách sử dụng máy bay tại Mỹ từng tháng từ năm 1949 đến năm 1960

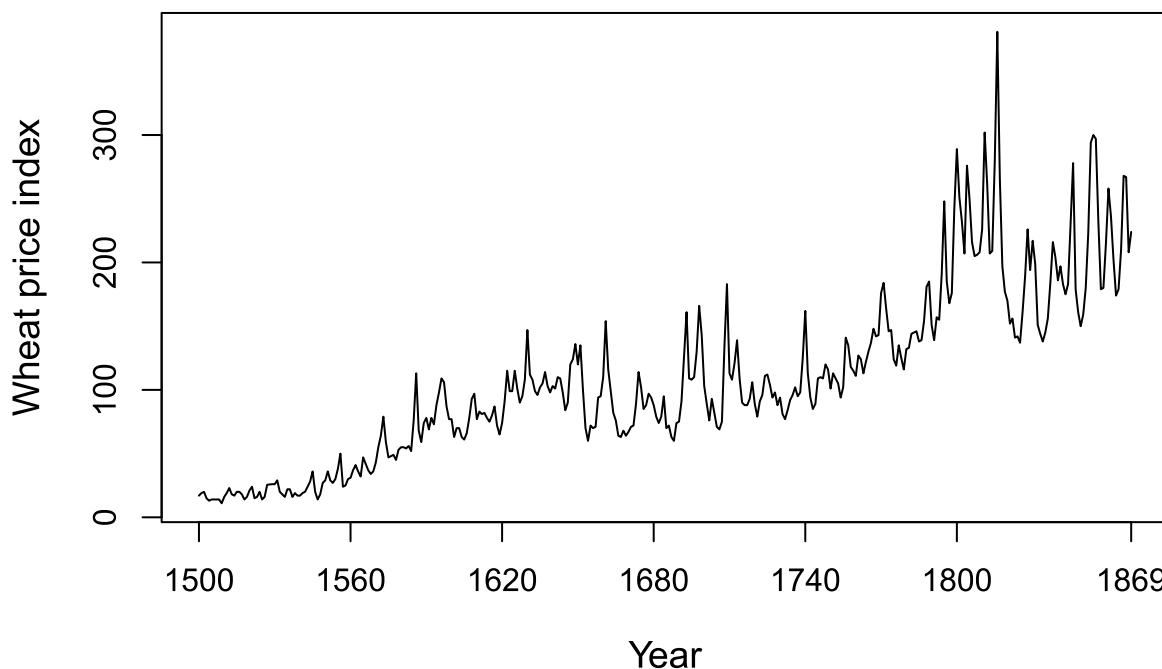
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

11. Phân tích dữ liệu chuỗi thời gian

Air Passenger numbers from 1949 to 1961



Ví dụ. Giá lúa mì từ năm 1500 đến 1869



Ví dụ. Chuỗi thời gian có thể được thu thập theo đơn vị thời gian là năm, tháng, ngày hay chi tiết hơn như giờ, phút,...

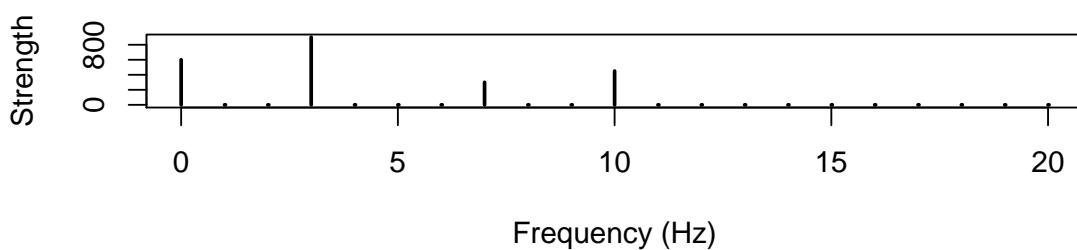
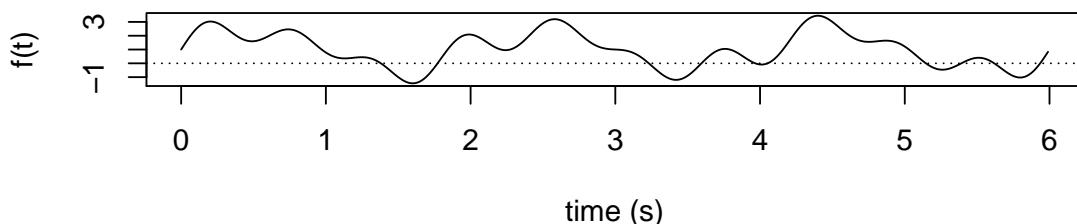
- Giá trị GDP của Việt Nam theo năm trong giai đoạn 1980 - 2013
- Giá đóng cửa của cổ phiếu VNM theo ngày giao dịch trong giai đoạn 2008 - 2013
- Tỷ giá trung bình của VND/USD theo tháng

11. Phân tích dữ liệu chuỗi thời gian

- Chi tiêu trung bình của nền kinh tế theo quý

Năm	Dân số	GDP	tỷ lệ tăng trưởng
2005	82,393,500	57.63	7.55
2006	83,313,000	66.37	6.98
2007	84,221,100	77.41	7.13
2008	85,122,300	99.13	5.66
2009	86,025,000	106.01	5.4
2010	86,932,500	115.93	6.42
2011	87,840,000	135.54	6.24
2012	88,772,900	155.82	5.25
2013	89,708,900	171.39	5.42

Trong các ngành khoa học kỹ thuật, xử lý tín hiệu thời gian thường tập trung vào phân tích tìm các **đặc trưng** (*feature*); ví dụ như phổ tần số của tín hiệu



11.1

Một số đặc trưng của chuỗi thời gian

Số liệu chuỗi thời gian có một số đặc trưng sau

- Tính tự tương quan:** Số liệu chuỗi thời gian thường có tính tự tương quan
 - Đầu tư năm nay có liên hệ với đầu tư năm trước

11. Phân tích dữ liệu chuỗi thời gian

- Tỷ lệ lạm phát của quý 1 năm nay có liên hệ với lạm phát của quý trước và của quý 1 năm trước
- **Yếu tố mùa vụ:** Các số liệu kinh tế - xã hội thường chịu tác động của yếu tố mùa vụ. Giá trị của chuỗi thời gian tại một thời điểm hoặc một thời kì năm nay có xu hướng biến động giống như cùng thời điểm hay cùng kì năm trước
 - Giá cả các năm thường cao vào dịp Tết
 - Chi tiêu của người dân thường cao vào quý 1 và quý 3
- **Yếu tố xu thế:** Đa phần các chuỗi thời gian thường có xu thế tăng hoặc giảm trong thời gian dài. Xu thế này có thể quan sát qua đồ thị của chuỗi
 - GDP của các Việt Nam tăng lên theo năm do phát triển công nghệ, cải thiện nguồn nhân lực, gia tăng nhân tố đầu vào, ...
 - Phát thải khí nhà kính của thế giới tăng theo năm do nhu cầu của khu vực sản xuất.
 - Diện tích rừng trên thế giới có xu hướng giảm do ngày càng cần đất đai để phục vụ các mục đích khác.

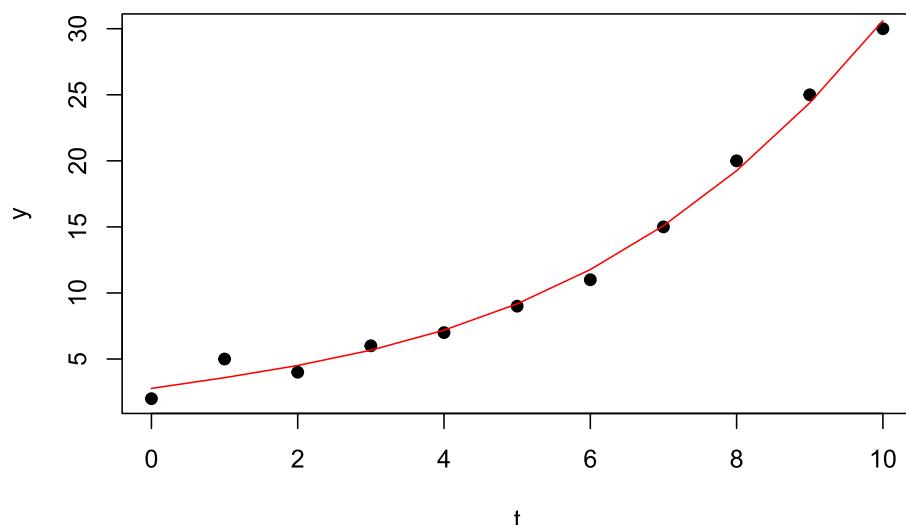
11.2

Các kỹ thuật tính toán xử lý dữ liệu

Với số liệu chuỗi thời gian ta thường sử dụng chỉ số t để chỉ thứ tự của các quan sát, chẳng hạn x_t, y_t, GDP_t, \dots trong đó $t = 1, 2, \dots, n$ hoặc $t = 0, 1, \dots, \infty$ hoặc $t = -\infty, \dots, -1, 0, 1, \dots, \infty$

11.2.1. Làm khớp đường cong với chuỗi dữ liệu

Một trong những kỹ thuật xử lý chuỗi không chứa yếu tố mùa vụ là làm khớp dữ liệu bằng đường cong đa thức, log hoặc logistic. Đường cong kết quả sẽ là một độ đo cho yếu tố xu thế của dữ liệu



11.2.2. Hàm biến đổi chuỗi dữ liệu

Sử dụng các hàm; ví dụ lấy log hoặc căn bậc hai, chuyển đổi chuỗi x_t thành chuỗi mới y_t .

$$x_t \longmapsto y_t = f(x_t) \quad (11.1)$$

Có 3 lý do để biến đổi chuỗi

1. Để ổn định phương sai
2. Để biến đổi yếu tố vụ mùa thành cộng tính
3. Để biến dữ liệu thành dạng phân phối chuẩn

11.2.3. Lọc (làm trơn) chuỗi dữ liệu

Kỹ thuật thứ hai có thể được dùng để đối phó với yếu tố xu thế là sử dụng là các bộ lọc tuyến tính, biến đổi chuỗi x_t thành chuỗi mới y_t

$$\{x_t\} \longmapsto y_t = f(\{x_t\}) \quad (11.2)$$

- **Lọc trung bình đơn giản:** tại thời điểm t là trung bình của m quan sát liên tiếp

$$y_t = \frac{x_{t-m+1} + x_{t-m+2} + \dots + x_t}{m} \quad (11.3)$$

- **Lọc trung bình trung tâm:** tại một thời điểm t là trung bình của m quan sát trước t , m quan sát sau t và quan sát tại t

$$y_t = \frac{x_{t-m} + x_{t-m+2} + \dots + x_t + \dots + x_{t+m-1} + x_{t+m}}{2m+1} \quad (11.4)$$

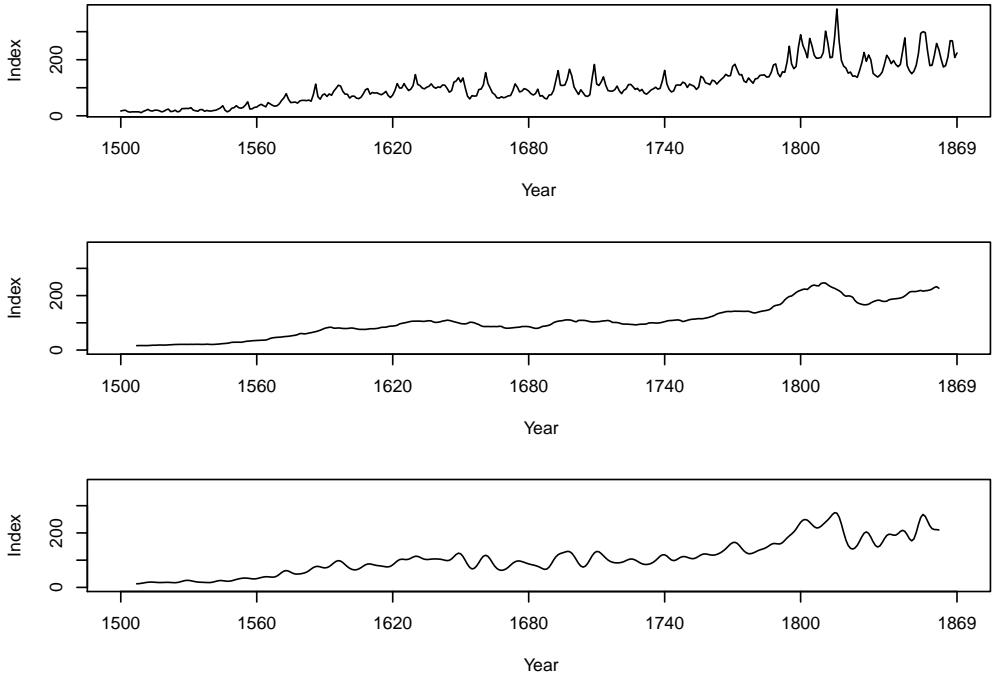
- **Lọc trung bình có trọng số:** thường được gọi là *moving average* được định nghĩa tổng quát như sau

$$y_t = Sm(\{x_t\}) = \sum_{i=-m}^n a_i x_{t+i} \quad (11.5)$$

với $\{a_i\}$ là tập các trọng số và $\sum_i a_i = 1$

Ví dụ. Sử dụng các bộ lọc trọng số để làm trơn chuỗi dữ liệu giá lùa mì

11. Phân tích dữ liệu chuỗi thời gian



- **Lọc mũ đơn giản:** được định nghĩa như sau

$$\begin{aligned} y_t &= \alpha x_t + \alpha(1 - \alpha)x_{t-1} + \alpha(1 - \alpha)^2x_{t-2} + \dots \\ &= \alpha \sum_{i=0}^{\infty} (1 - \alpha)^i x_{t-i} \end{aligned} \quad (11.6)$$

với $0 < \alpha < 1$

Biến đổi công thức trên thành công thức truy hồi

$$y_t = \alpha x_t + (1 - \alpha)y_{t-1} \quad (11.7)$$

11.2.4. Sai phân chuỗi dữ liệu

Sai phân (differencing) là một dạng hàm lọc đặc biệt thường được dùng để loại bỏ các yếu tố xu hướng ra khỏi chuỗi dữ liệu

Định nghĩa.

- Sai phân bậc 1 của một chuỗi x_t

$$\nabla x_t = x_t - x_{t-1} \quad (11.8)$$

- Sai phân bậc 2 của một chuỗi x_t

$$\nabla^2 x_t = \nabla x_t - \nabla x_{t-1} \quad (11.9)$$

$$= x_t - 2x_{t-1} + x_{t-2} \quad (11.10)$$

11. Phân tích dữ liệu chuỗi thời gian

- Sai phân bậc k của một chuỗi x_t

$$\nabla^k x_t = \nabla^{k-1} x_t - \nabla^{k-1} x_{t-1} \quad (11.11)$$

11.2.5. Hàm tự tương quan

Định nghĩa. Cho một chuỗi dữ liệu thời gian y_t có $\mu = \mathbb{E}[y_t]$, chúng ta định nghĩa hàm autocovariance ứng với độ trễ (lag) k

$$\gamma(k) = \mathbb{E}[(y_{t+k} - \mu)(y_t - \mu)] \quad (11.12)$$

và hàm tự tương quan ACF (autocorrelation function) ứng với độ trễ k

$$\text{ACF}(k) = \rho(k) = \frac{\gamma(k)}{\gamma(0)} \quad (11.13)$$

Hàm tự tương quan có tính chất sau

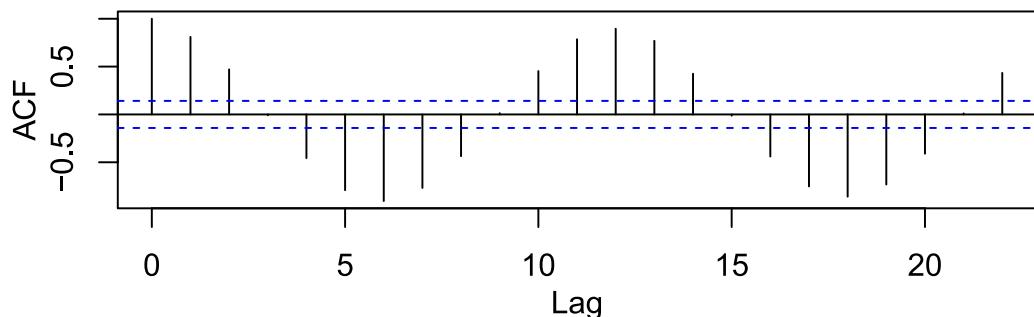
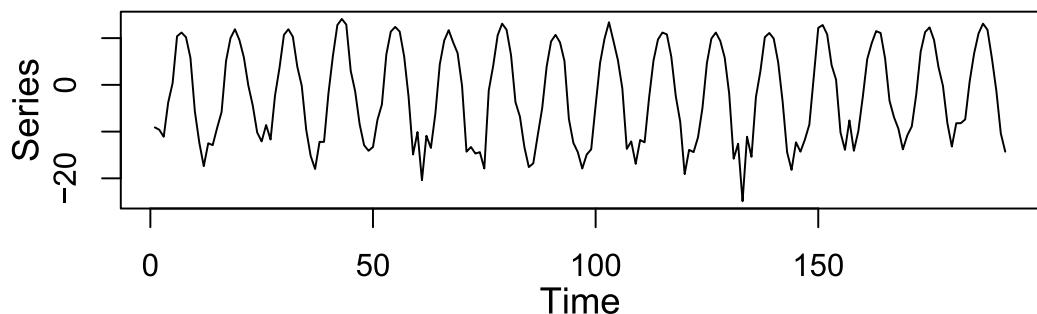
- Là hàm chẵn

$$\rho(k) = \rho(-k) \quad (11.14)$$

- Là hàm bị chặn

$$|\rho(k)| \leq 1$$

- Để trực quan hóa hàm tự tương quan chúng ta sử dụng **biểu đồ tự tương quan** (correlogram)



Ước lượng hàm tự tương quan

Cho một tập mẫu chuỗi dữ liệu thời gian $\{y_1, y_2, \dots, y_T\}$ thì ta có thể tính các ước lượng bằng

$$\hat{\gamma}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (y_{t+k} - \bar{y})(y_t - \bar{y}) \quad (11.15)$$

và

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} \quad (11.16)$$

11.3 Các thành phần của chuỗi thời gian

Định nghĩa. Chuỗi thời gian thông thường gồm có 4 thành phần sau:

- Thành phần xu thế (*trend*) - T
- Thành phần mùa vụ (*seasonality*) - S
- Thành phần chu kỳ (*cyclical*) - C
- Thành phần ngẫu nhiên (*random*) - R

Định nghĩa. Có hai loại mô hình để phân tích thành phần chuỗi thời gian

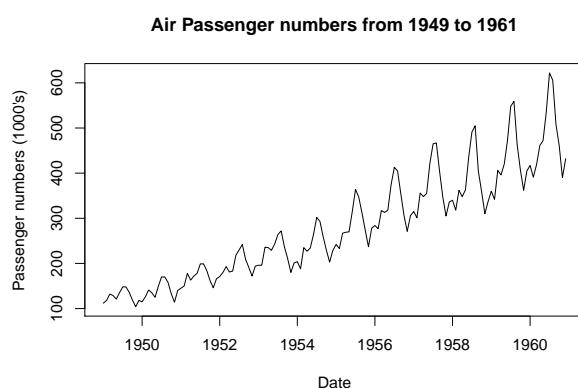
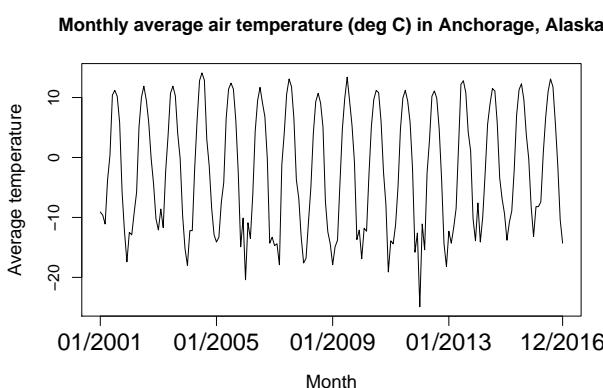
- Mô hình cộng:

$$Y_t = T_t + S_t + C_t + R_t \quad (11.17)$$

- Mô hình nhân:

$$Y_t = T_t \times S_t \times C_t \times R_t \quad (11.18)$$

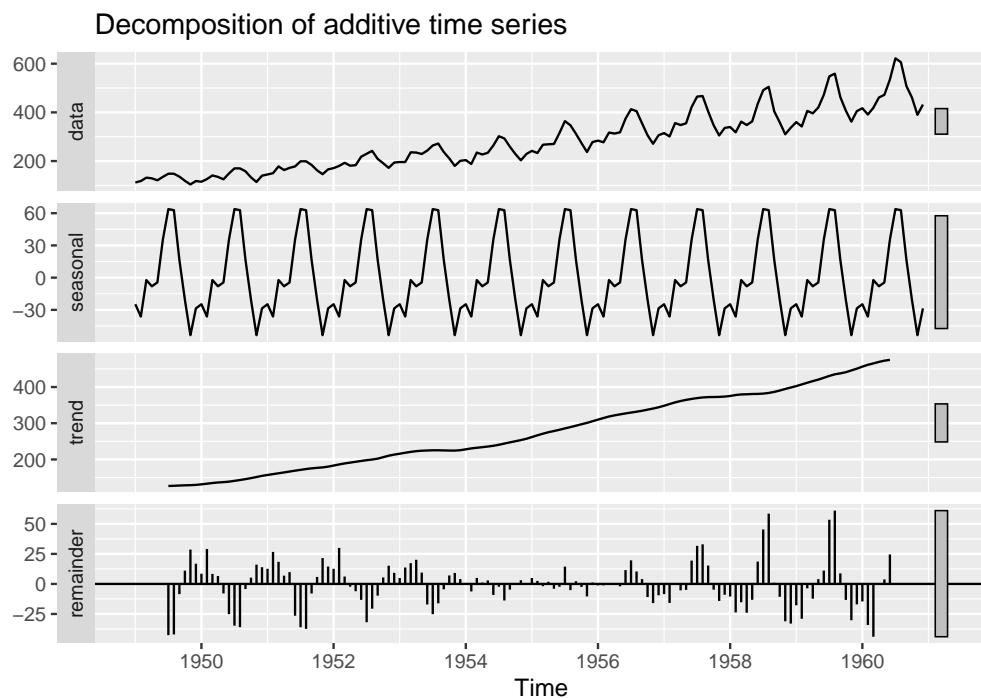
$$\log(Y_t) = \log(T_t) + \log(S_t) + \log(C_t) + \log(R_t) \quad (11.19)$$



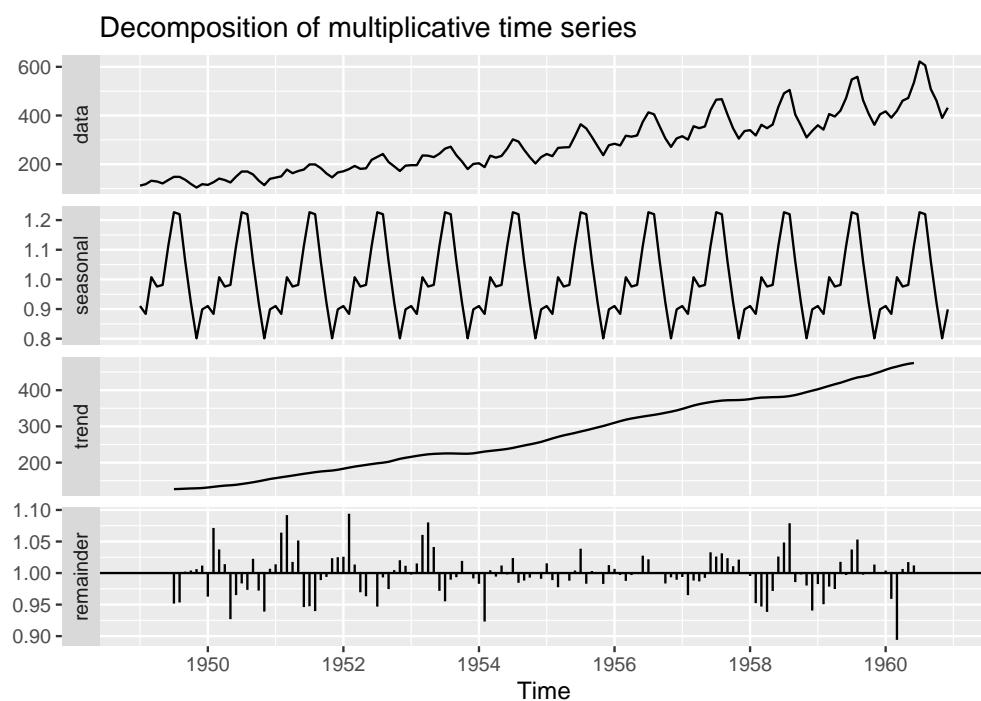
11. Phân tích dữ liệu chuỗi thời gian

11.3.1. Phân rã chuỗi dữ liệu

- Sử dụng mô hình cộng phân rã chuỗi dữ liệu hành khách hàng không



- Sử dụng mô hình nhân phân rã chuỗi dữ liệu hành khách hàng không

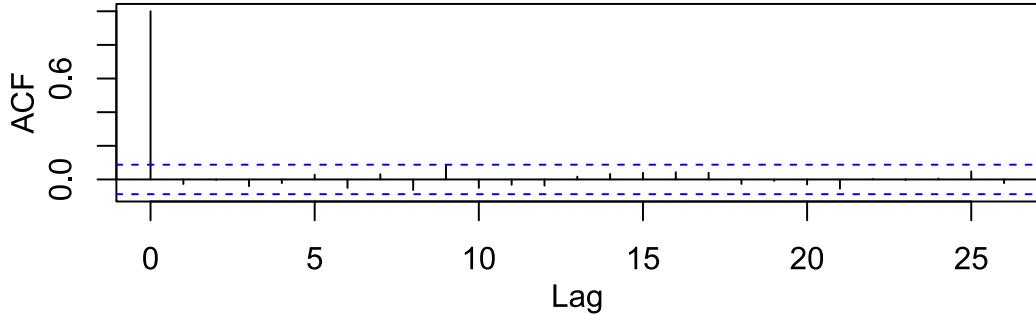
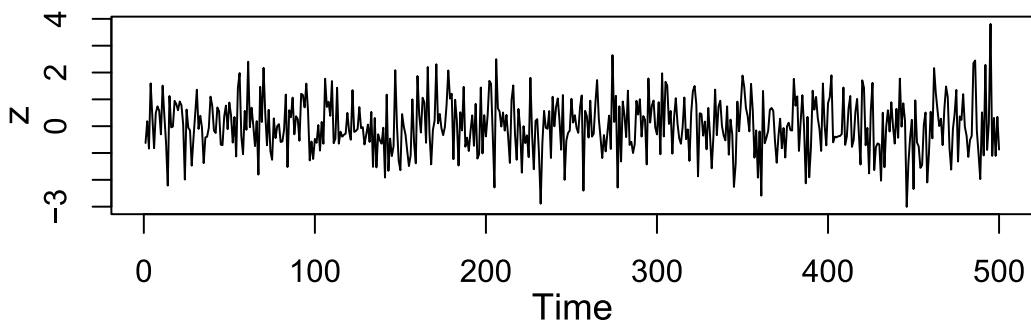


11.4 Quá trình ngẫu nhiên

Định nghĩa. Quá trình ngẫu nhiên rời rạc theo thời gian (*discrete-time random process*) là một chuỗi các biến ngẫu nhiên $\{Z_t\}_{t=0}^{\infty}$

Định nghĩa. Quá trình ngẫu nhiên $\{Z_t\}_{t=0}^{\infty}$ được gọi là *quá trình ngẫu nhiên thuần túy* nếu các biến ngẫu nhiên có phân bố đồng nhất và độc lập với nhau (*independent and identically distributed - i.i.d.*)

- Quá trình ngẫu nhiên và biểu đồ tự tương quan



Định nghĩa. Quá trình ngẫu nhiên $\{Z_t\}_{t=0}^{\infty}$ được gọi là *nhiễu trắng* (*white noise*) nếu mỗi biến ngẫu nhiên có kỳ vọng bằng 0, phương sai giống nhau và không tự tương quan

- $\mathbb{E}(Z_t) = 0, \forall t$
- $\text{Var}(Z_t) = \sigma^2, \forall t$
- $\text{cov}(Z_i, Z_j) = 0, \forall i \neq j$

Đôi khi nhiễu trắng còn thêm điều kiện i.i.d.

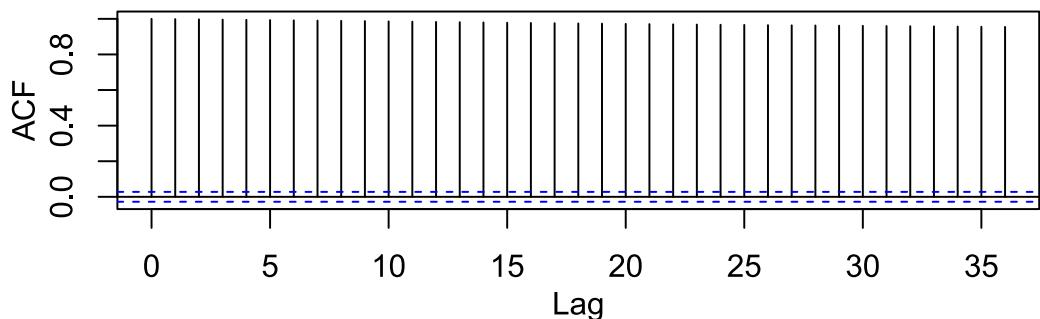
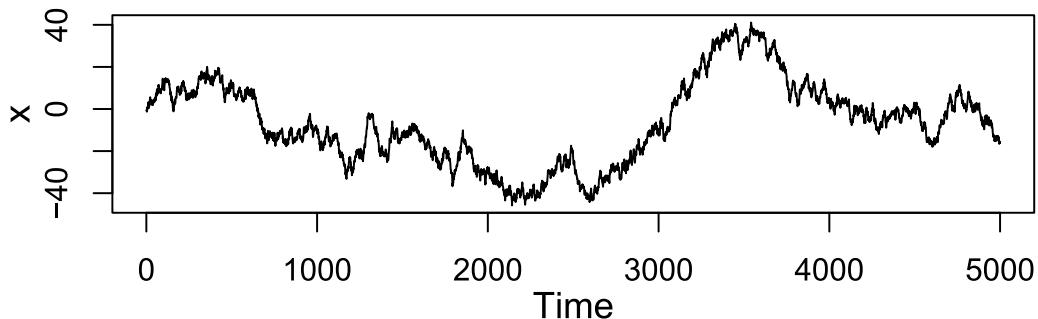
11.4.1. Bước ngẫu nhiên

Định nghĩa. Giả sử quá trình ngẫu nhiên thuận túy $\{Z_t\}_{t=0}^{\infty}$ với trung bình là μ và phương sai là σ^2 , quá trình ngẫu nhiên $\{X_t\}_{t=0}^{\infty}$ được định nghĩa như sau

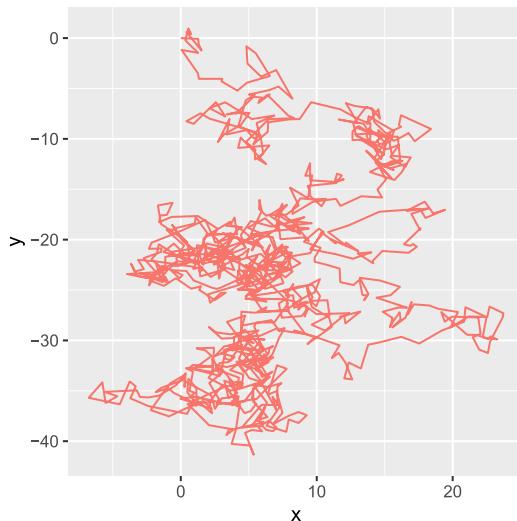
$$X_t = X_{t-1} + Z_t \quad (11.20)$$

được gọi là **bước ngẫu nhiên** (*random walk*)

- Bước ngẫu nhiên với $\{Z_t\}_{t=0}^{\infty}$ với trung bình $\mu = 0$ và phương sai $\sigma^2 = 1$ và biểu đồ tự tương quan

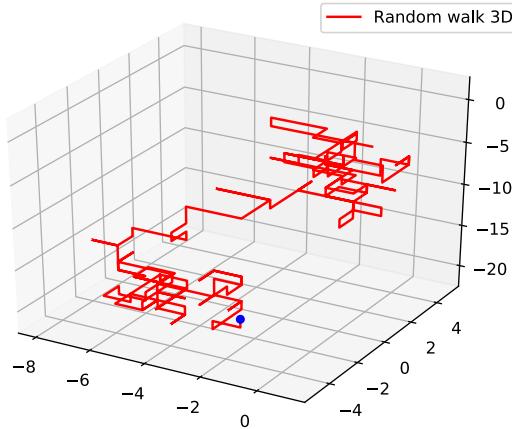


- Bước ngẫu nhiên trong không gian 2D



11. Phân tích dữ liệu chuỗi thời gian

- Bước ngẫu nhiên trong không gian 3D



11.4.2. Quá trình trung bình trượt

Định nghĩa. Giả sử quá trình ngẫu nhiên thuần túy $\{Z_t\}_{t=0}^{\infty}$ với trung bình $\mu = 0$ và phương sai là σ^2 , thì **quá trình trung bình trượt** (*moving average process*) bậc q ký hiệu là $MA(q)$

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \dots + \beta_q Z_{t-q}, \quad (11.21)$$

hệ số β_0 thường được chọn là 1.

- Một số tính chất của quá trình trung bình trượt $MA(q)$

$$\mathbb{E}(X_t) = 0 \quad (11.22)$$

$$\text{Var}(X_t) = \sigma^2 \sum_{i=0}^q \beta_i \quad (11.23)$$

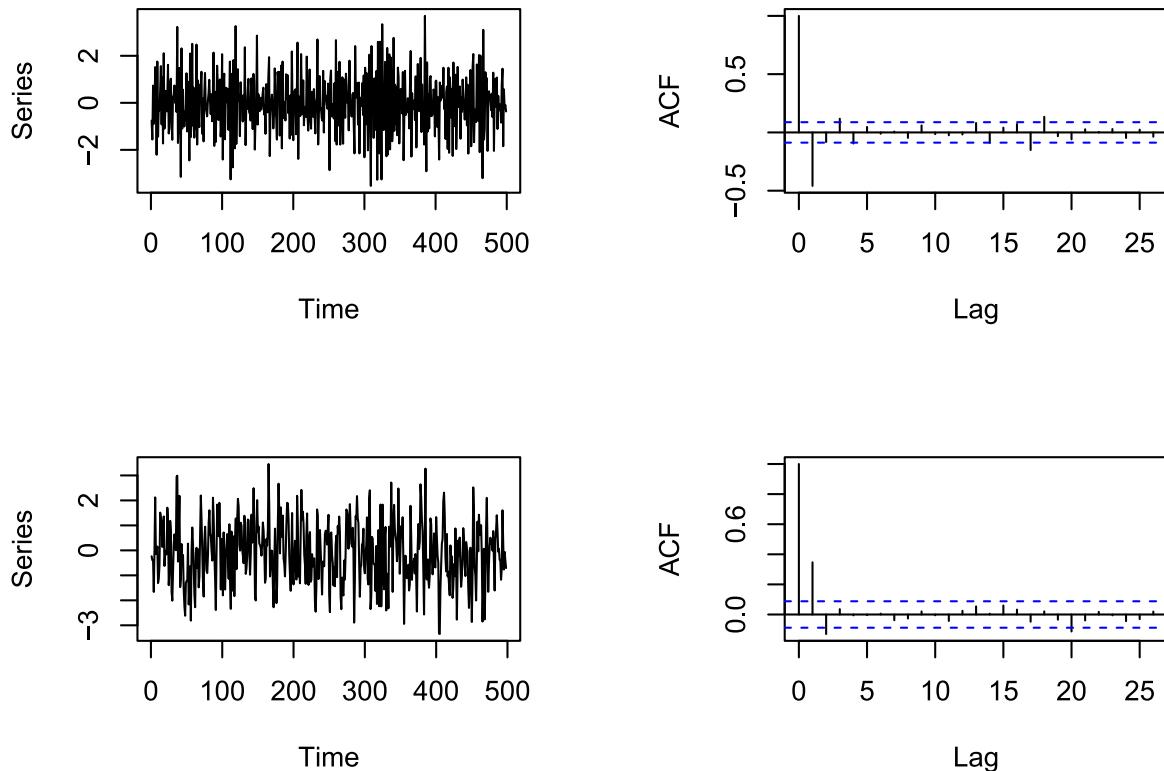
$$\rho(k) = \begin{cases} 1 & k = 0 \\ \frac{\sum_{i=0}^{q-k} \beta_i \beta_{i+k}}{\sum_{i=0}^q \beta_i^2} & k = 1, \dots, q \\ 0 & k > q \end{cases} \quad (11.24)$$

- Trực quan quá trình trung bình trượt $MA(1)$ và $MA(2)$

$$X_t = Z_t - 0.8Z_{t-1}, \quad Z_t \sim \mathcal{N}(0, 1)$$

$$X_t = Z_t + 0.7Z_{t-1} - 0.2Z_{t-2}, \quad Z_t \sim \mathcal{N}(0, 1)$$

11. Phân tích dữ liệu chuỗi thời gian



11.4.3. Quá trình tự hồi quy

Định nghĩa. Giả sử quá trình ngẫu nhiên thuần túy $\{Z_t\}_{t=0}^{\infty}$ với trung bình $\mu = 0$ và phương sai là σ^2 , thì **quá trình tự hồi quy** (*autoregressive process*) bậc p ký hiệu là AR(p)

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + Z_t \quad (11.25)$$

Quá trình bậc một

Xét quá trình tự hồi quy bậc 1 AR(1)

$$X_t = \alpha X_{t-1} + Z_t \quad (11.26)$$

- Một số tính chất

$$\mathbb{E}(X_t) = 0 \quad (11.27)$$

$$\text{Var}(X_t) = \sigma^2(1 + \alpha + \alpha^2 + \dots) \quad (11.28)$$

- Nếu $|\alpha| < 1$ thì

$$\text{Var}(X_t) = \frac{\sigma^2}{1 - \alpha} \quad (11.29)$$

$$\rho(k) = \alpha^k \quad (11.30)$$

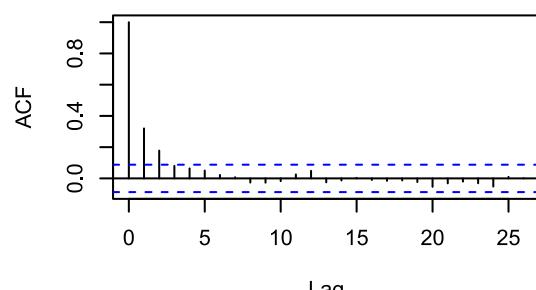
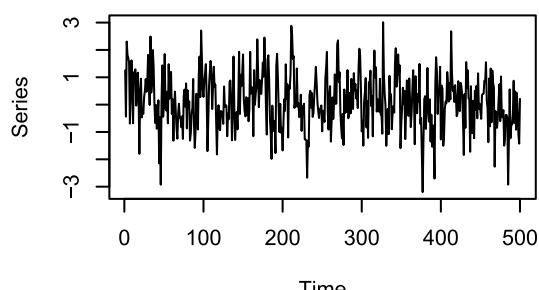
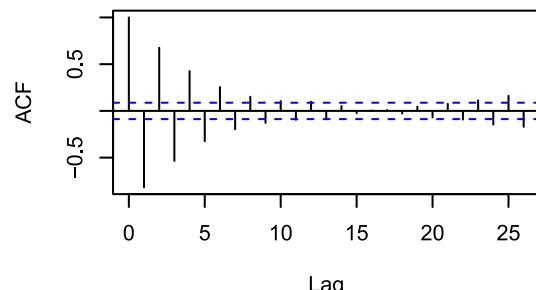
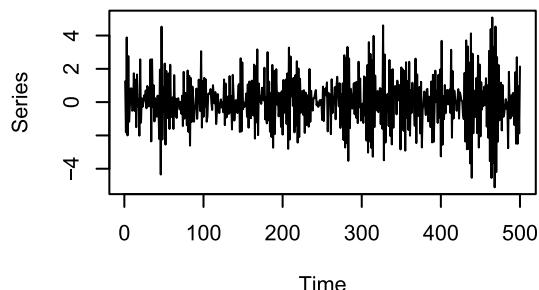
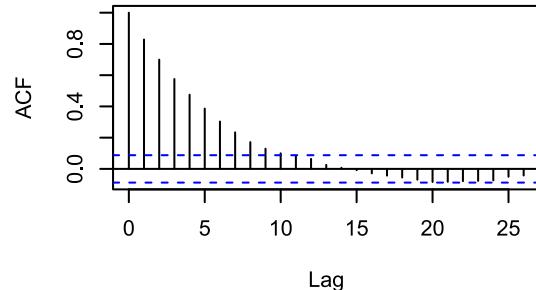
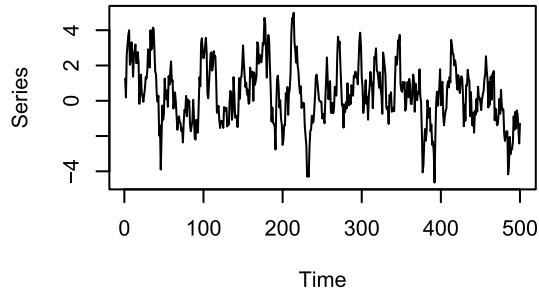
- Trực quan 3 quá trình tự hồi quy bậc 1 AR(1) và biểu đồ tự tương quan

11. Phân tích dữ liệu chuỗi thời gian

$$X_t = 0.8X_{t-1} + Z_t, \quad Z_t \sim \mathcal{N}(0, 1)$$

$$X_t = -0.8X_{t-1} + Z_t, \quad Z_t \sim \mathcal{N}(0, 1)$$

$$X_t = 0.3X_{t-1} + Z_t, \quad Z_t \sim \mathcal{N}(0, 1)$$



Quá trình bậc tổng quát

- Phương trình Yule-Walker

$$\rho(k) = \alpha_1\rho(k-1) + \dots + \alpha_p\rho(k-p), \quad \forall k > 0 \quad (11.31)$$

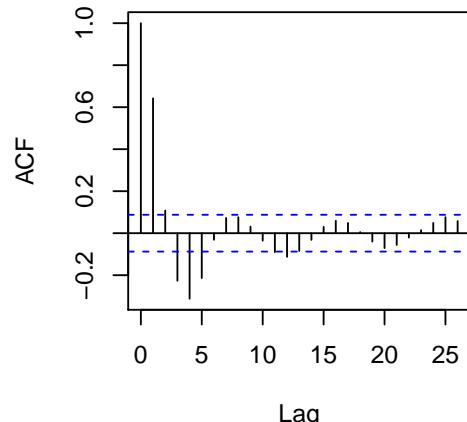
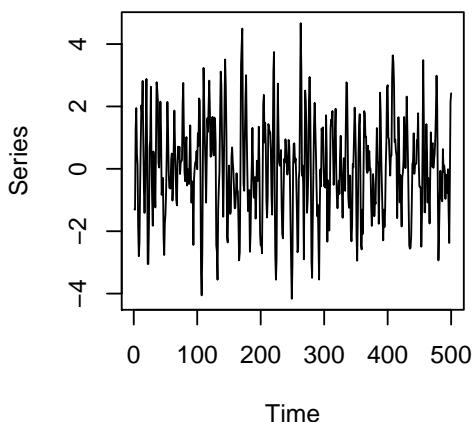
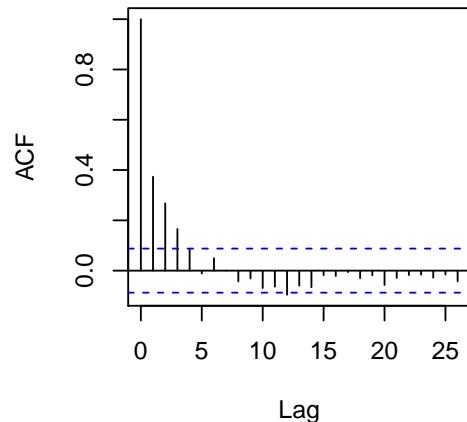
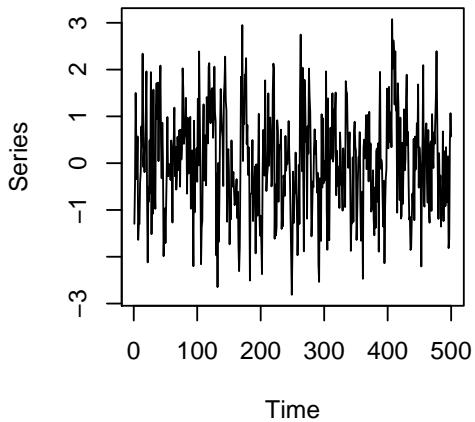
11. Phân tích dữ liệu chuỗi thời gian

$$\begin{pmatrix} \rho(0) & \rho(1) & \cdots & \rho(p-1) \\ \rho(1) & \rho(0) & \cdots & \rho(p-2) \\ \vdots & \vdots & \vdots & \vdots \\ \rho(p-1) & \rho(p-2) & \cdots & \rho(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(p) \end{pmatrix} \quad (11.32)$$

- Trực quan 2 quá trình tự hồi quy bậc 2 và biểu đồ tự tương quan

$$X_t = \frac{1}{3}X_{t-1} + \frac{2}{9}X_{t-2} + Z_t, \quad Z_t \sim \mathcal{N}(0, 1)$$

$$X_t = X_{t-1} + \frac{1}{2}X_{t-2} + Z_t, \quad Z_t \sim \mathcal{N}(0, 1)$$



Định nghĩa. Hàm tự tương quan riêng phần (partial autocorrelation function - PACF) là hệ số tương quan giữa X_t và X_{t-k} và loại bỏ tất cả các tương quan với $X_{t-1}, \dots, X_{t-k-1}$, hay nói cách khác nó đã loại bỏ tác động của các giá trị X trung gian. Về tính toán nó là hệ số tương

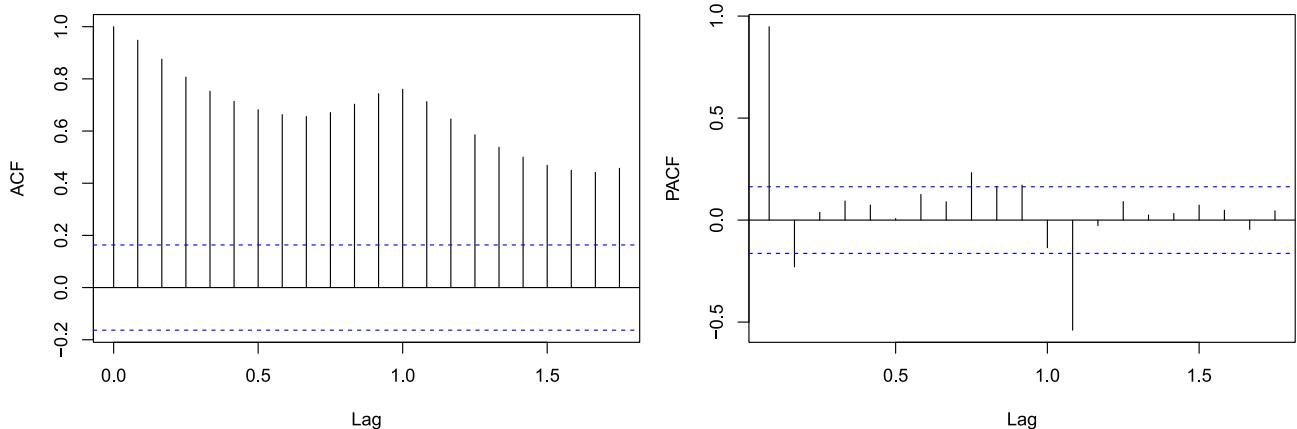
11. Phân tích dữ liệu chuỗi thời gian

quan cuối cùng trong mô hình AR(k) và ký hiệu là ρ_{kk}

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_k X_{t-k} + Z_t \quad (11.33)$$

$$\text{PACF}(k) = \rho_{kk} = \alpha_k \quad (11.34)$$

- Hàm tự tương quan và tương quan riêng phần cho chuỗi dữ liệu khách hàng không



11.5 Chuỗi dữ liệu dừng

Trong phân tích dữ liệu chuỗi thời gian, một mô hình tốt được đưa ra khi phân tích trên các chuỗi dữ liệu dừng.

Định nghĩa. Một chuỗi thời gian là dừng (*stationary time series*) khi các đặc trưng thống kê (trung bình, phương sai tại các độ trễ khác nhau) giữ nguyên không đổi cho dù chuỗi được xác định vào thời điểm nào đi nữa.

Chuỗi dừng có xu hướng trở về giá trị trung bình và những dao động quanh giá trị trung bình sẽ là như nhau. Nói cách khác, một chuỗi thời gian không dừng sẽ có giá trị trung bình thay đổi theo thời gian, hoặc giá trị phương sai thay đổi theo thời gian hoặc cả hai.

Nói chung, một chuỗi thời gian dừng sẽ không có mô hình dự đoán được trong dài hạn (long-term).

11.5.1. Kiểm định tính dừng

Biểu đồ

Kiểm định nghiệm đơn vị

Kiểm định Dickey - Fuller cho chuỗi dữ liệu thời gian

- Giả thuyết $H_0 : \alpha = 1$, chuỗi không dừng
- Đổi thuyết $H_a : |\alpha| < 1$, chuỗi dừng

11.6 Cú pháp Backshift

Phép toán backshift B (một số tài liệu sử dụng là L) rất hữu dụng trong việc viết công thức cho các chuỗi dữ liệu thời gian

Khái niệm 11.1. Gọi y_t là chuỗi dữ liệu thời gian, thực hiện phép toán backshift B cho y_t ta có

$$B(y_t) = By_t = y_{t-1} \quad (11.35)$$

Từ định nghĩa ta có

$$B(By_t) = B^2y_t = y_{t-2}, \quad (11.36)$$

và

$$B^p y_t = y_{t-p}. \quad (11.37)$$

Phép toán backshift rất tiện lợi khi biểu diễn quá trình sai phân.

- Sai phân bậc 1

$$\nabla y_t = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t \quad (11.38)$$

- Sai phân bậc 2

$$\nabla^2 y_t = y_t - 2y_{t-1} + y_{t-2} = (1 - 2B + B^2)y_t = (1 - B)^2 y_t \quad (11.39)$$

- Sai phân bậc d

$$\nabla^d y_t = (1 - B)^d y_t \quad (11.40)$$

Ví dụ, để tính sai phân mùa vụ m theo sau một sai phân bậc 1 ta có thể viết như sau

$$\begin{aligned} (1 - B)(1 - B^m)y_t &= (1 - B - B^m + B^{m+1})y_t \\ &= y_t - y_{t-1} - y_{t-m} + y_{t-m-1} \end{aligned} \quad (11.41)$$

11.7 Mô hình ARMA, ARIMA, SARIMA

11.7.1. Mô hình ARMA

Định nghĩa. Mô hình ARMA(p, q) là mô hình cho một chuỗi dữ liệu dùng $\{X_t\}$ bằng sự kết hợp hai quá trình ngẫu nhiên AR(p) và MA(q)

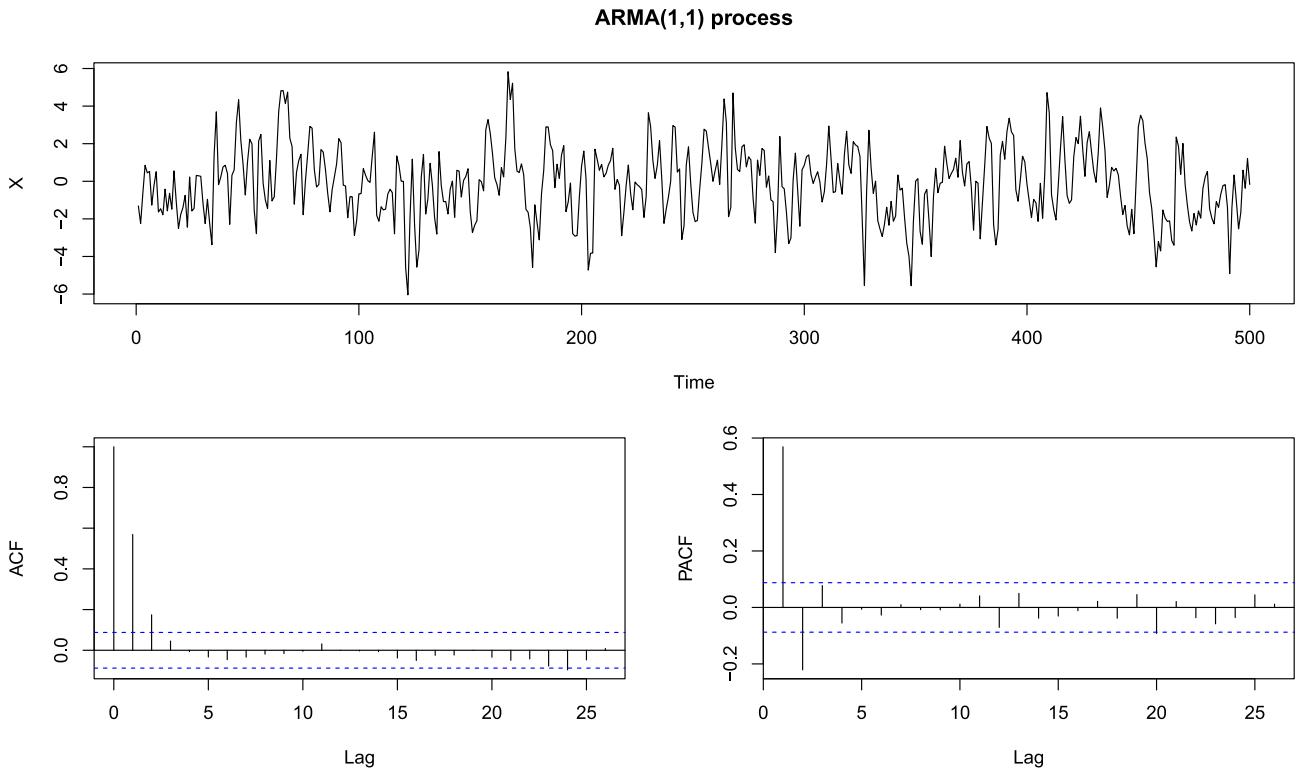
$$\begin{aligned} X_t &= \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} \\ &\quad + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \dots + \beta_q Z_{t-q} + Z_t \end{aligned} \quad (11.42)$$

11. Phân tích dữ liệu chuỗi thời gian

Việc ước lượng các tham số $\alpha_1, \alpha_2, \dots, \alpha_p, \beta_1, \beta_2, \dots, \beta_q$ được thực hiện bằng phương pháp *maximum likelihood* (tham khảo lại các quá trình ngẫu nhiên AR và MA).

- Trực quan quá trình ARMA(1, 1) và biểu đồ tự tương quan và tự tương quan riêng phần

$$X_t = 0.35X_{t-1} + Z_t + 0.4Z_{t-1}, \quad Z_t \sim \mathcal{N}(0, 1)$$



11.7.2. Mô hình ARIMA

Định nghĩa. Mô hình ARIMA(p, d, q) (*autoregressive integrated moving average*) là mô hình cho một chuỗi dữ liệu $\{X_t\}$ bằng cách áp dụng mô hình ARMA(p, q) trên sai phân bậc d trên chuỗi $\{X_t\}$ ($\nabla^d(X_t)$). Phương trình tổng quát của mô hình

$$\begin{aligned} X'_t &= c + \alpha_1 X'_{t-1} + \alpha_2 X'_{t-2} + \dots + \alpha_p X'_{t-p} \\ &\quad + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \dots + \beta_q Z_{t-q} + Z_t, \end{aligned} \tag{11.43}$$

trong đó X' là sai phân bậc d của X .

Phương trình mô hình cũng có thể được viết theo cú pháp backshift như sau

$$\underbrace{(1 - \alpha_1 B - \dots - \alpha_p B^p)}_{\text{AR}(p)} \underbrace{(1 - B)^d}_{\text{sai phân } d} X_t = c + \underbrace{(1 + \beta_1 B + \dots + \beta_q B^q)}_{\text{MA}(q)} Z_t \tag{11.44}$$

Một số mô hình ARIMA cơ bản

11. Phân tích dữ liệu chuỗi thời gian

Chuỗi dữ liệu	Mô hình
Nhiều trắng (white noise)	ARIMA(0, 0, 0)
Bước ngẫu nhiên (random walk)	ARIMA(0, 1, 0) không có hằng số
Bước ngẫu nhiên trượt (random walk with drift)	ARIMA(0, 1, 0) có hằng số
Tự hồi quy (autoregression)	ARIMA($p, 0, 0$)
Trung bình trượt (moving average)	ARIMA(0, 0, q)

Ý nghĩa các tham số

Hằng số c có ảnh hưởng quan trọng đến các dự báo dài hạn của mô hình

- Nếu $c = 0$ và $d = 0$ thì các dự báo dài hạn sẽ về 0.
- Nếu $c = 0$ và $d = 1$ thì các dự báo dài hạn sẽ đi đến một hằng số khác 0.
- Nếu $c = 0$ và $d = 2$ thì các dự báo dài hạn sẽ đi theo một đường thẳng.
- Nếu $c \neq 0$ và $d = 0$ thì các dự báo dài hạn sẽ đi đến giá trị trung bình của dữ liệu.
- Nếu $c \neq 0$ và $d = 1$ thì các dự báo dài hạn sẽ đi theo một đường thẳng.
- Nếu $c \neq 0$ và $d = 2$ thì các dự báo dài hạn sẽ theo xu hướng bậc hai.

Giá trị của d ảnh hưởng đến khoảng giá trị dự báo (*prediction interval*).

- Khi d tăng thì khoảng giá trị dự báo cũng tăng nhanh theo.
- Khi $d = 0$ thì các khoảng giá trị dự báo cơ bản là giống nhau.

Giá trị của p ảnh hưởng đến chu kỳ của chuỗi dữ liệu. Để có được các dự báo theo chu kỳ, cần phải có $p \geq 2$

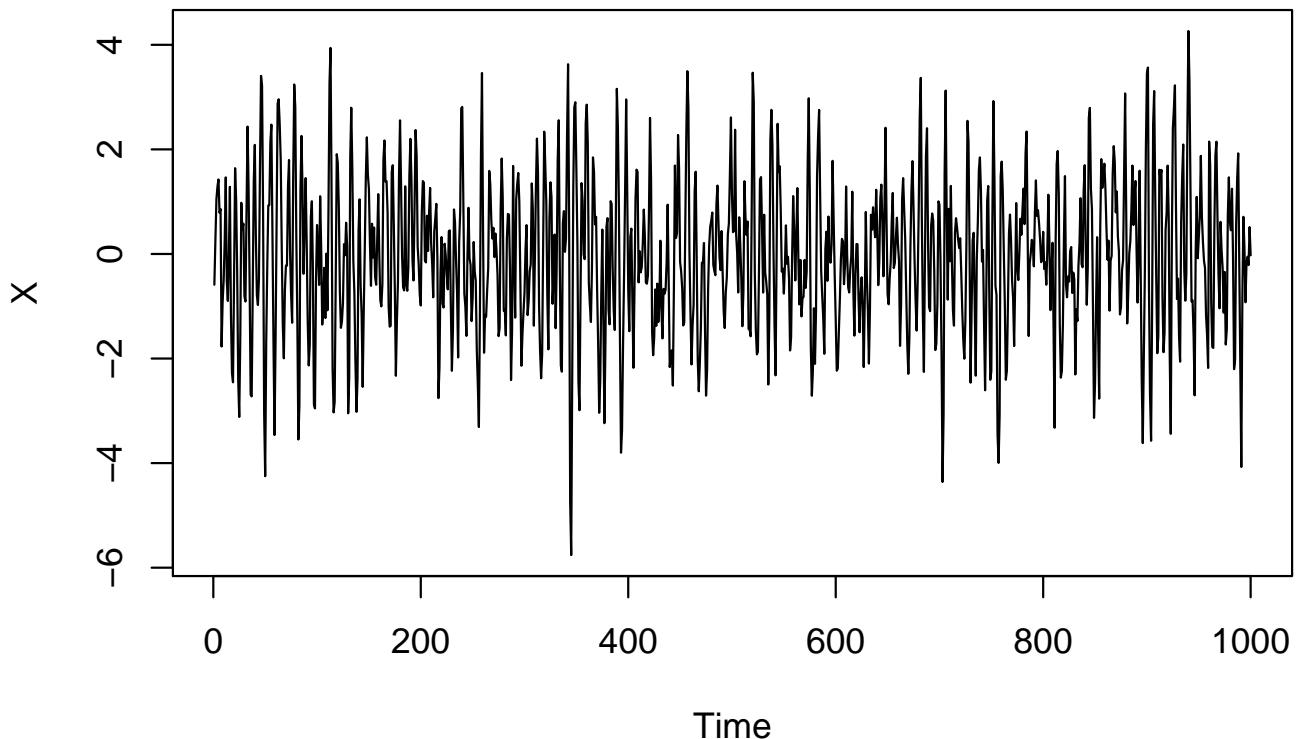
Biểu đồ ACF và PACF

Mô hình	ACF	PACF
$AR(p) = ARIMA(p, d, 0)$	Suy giảm theo số mũ hay hình sin tắt dần	Đỉnh cao đáng kể tại độ trễ p , và không còn đỉnh sau p
$MA(q) = ARIMA(0, d, q)$	Đỉnh cao đáng kể tại độ trễ q , và không còn đỉnh sau q	Suy giảm theo số mũ
$ARIMA(p, d, q)$	Suy giảm theo số mũ	Suy giảm theo số mũ

- Các biểu đồ ACF và PACF có thể hữu ích trong việc xác định giá trị p hoặc q cho $ARIMA(p, d, 0)$ hoặc $ARIMA(0, d, q)$

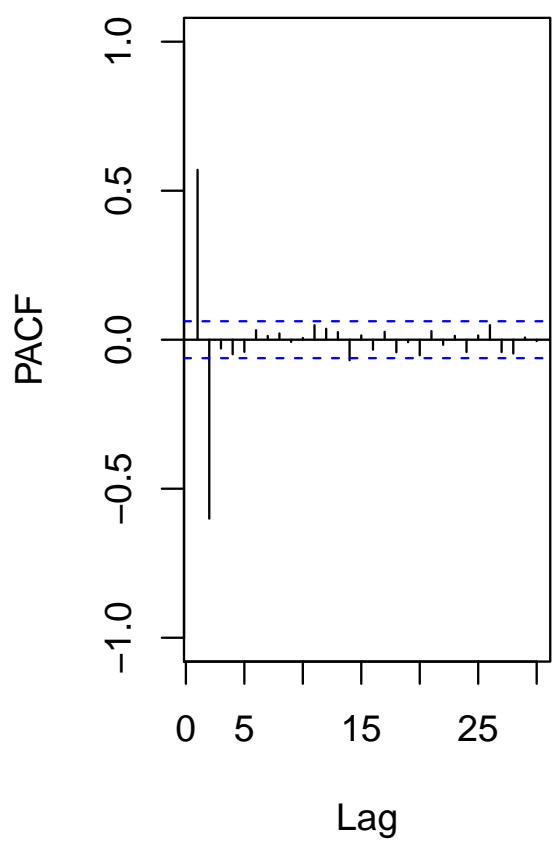
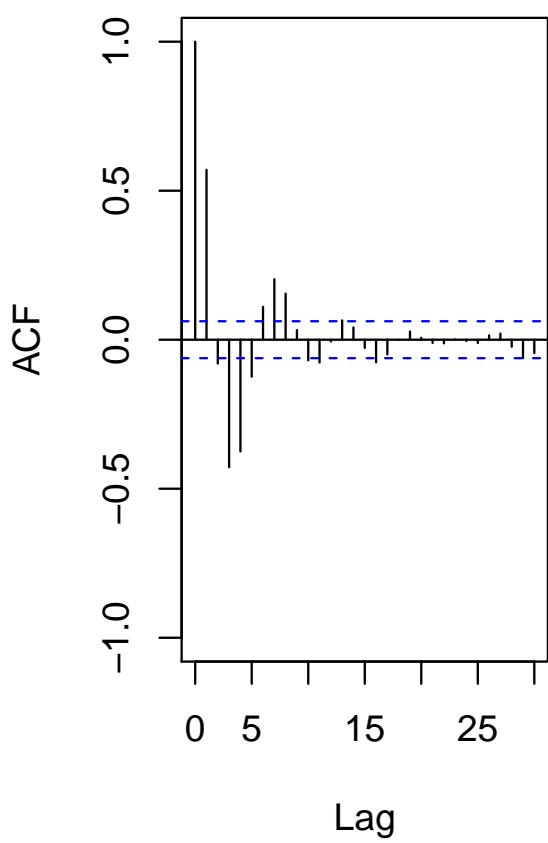
11. Phân tích dữ liệu chuỗi thời gian

- Tuy nhiên, nó không giúp nhiều cho việc xác định p, q phù hợp cho mô hình tổng quát ARIMA(p, d, q)

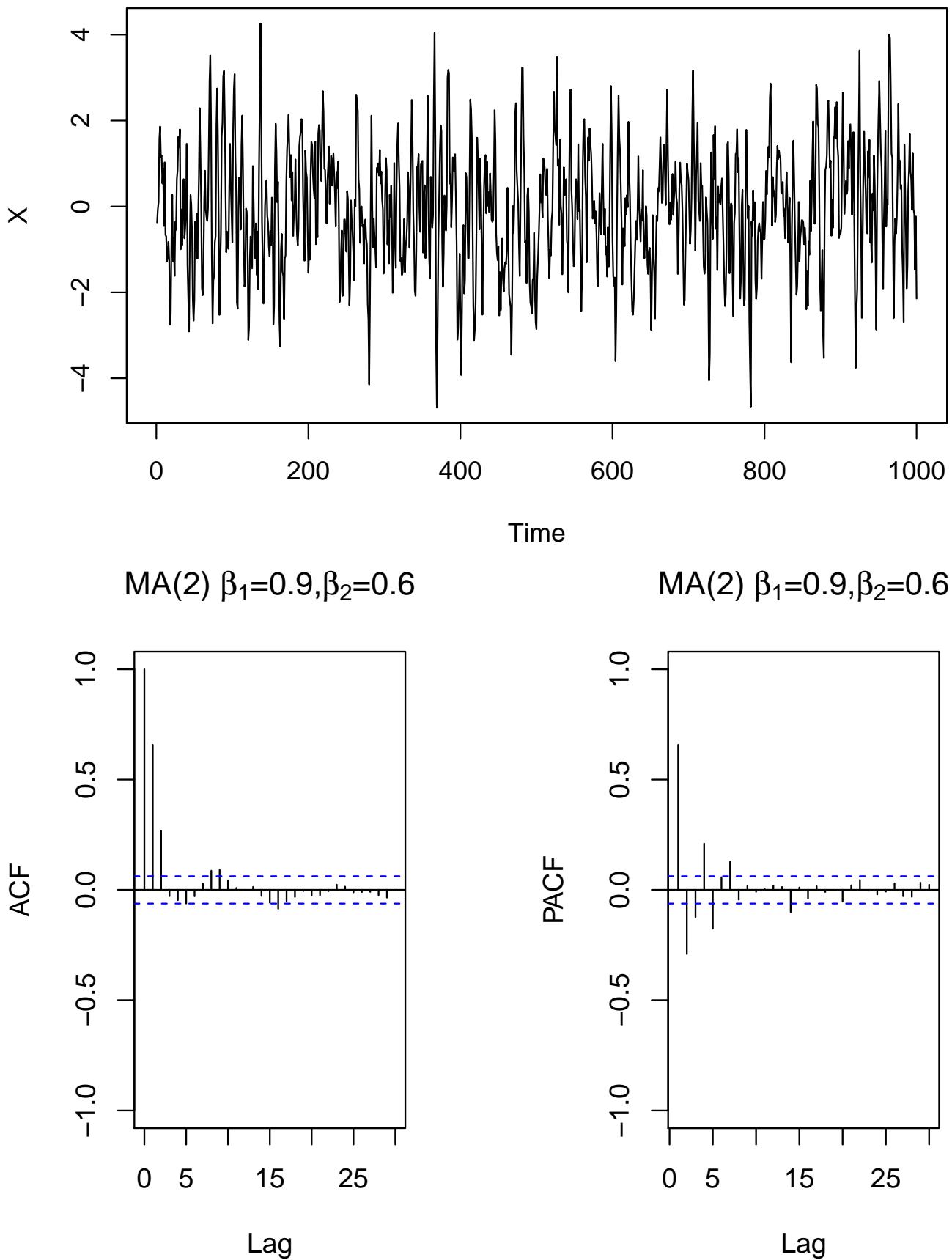


AR(2) $\alpha_1=0.9, \alpha_2=-0.6$

AR(2) $\alpha_1=0.9, \alpha_2=-0.6$



11. Phân tích dữ liệu chuỗi thời gian



11.7.3. Mô hình SARIMA

Định nghĩa. Mô hình SARIMA(p, d, q)(P, D, Q)[m] là mô hình ARIMA mở rộng trên chuỗi dữ liệu có chu kỳ theo mùa.

$$\left[\begin{array}{c} \text{Phần mô hình không có tính mùa} & \text{Phần mô hình có tính mùa} & \text{Số đoạn mỗi mùa} \\ \text{SARIMA} & (p, d, q) & (P, D, Q) & [m] \\ & & & (11.45) \end{array} \right],$$

trong đó $m = 4$ nếu chuỗi dữ liệu theo quý và $m = 12$ nếu chuỗi dữ liệu theo tháng.

11.8 Phương pháp Box-Jenkins

Định nghĩa. Phương pháp Box-Jenkins là quá trình xây dựng mô hình bằng cách chọn ra một mô hình từ lớp các mô hình (S)ARIMA. Kỹ thuật của Box-Jenkins là tiến trình xây dựng mô hình chứ không chỉ đơn thuần là tiến trình tìm kiếm mô hình phù hợp.

Phương pháp Box-Jenkins

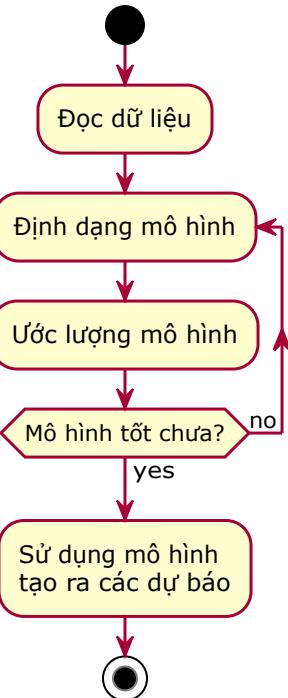
Bao gồm 3 bước chính:

- Định dạng mô hình (S)ARIMA.
- Ước lượng mô hình (S)ARIMA.
- Kiểm định mô hình.

Lặp lại các bước trên đến khi nào tìm được mô hình tốt.

Lưu đồ thuật toán cho phương pháp Box-Jenkins

- **Đầu vào:** là chuỗi dữ liệu $\mathcal{D} = \{x_1, x_2, \dots, x_T\}$
- **Đầu ra:** là mô hình (S)ARIMA(p, d, q)(P, D, Q)[m] tốt nhất (theo tiêu chí xác định)



11.8.1. Tiêu chí lựa chọn mô hình

Khi chúng ta có nhiều mô hình khác nhau thì để chọn mô hình tốt nhất ta có thể sử dụng các tiêu chí

- Likelihood L của dữ liệu khi sử dụng mô hình (càng lớn càng tốt)
- Chỉ số AIC (Akaike Information Criteria) (càng nhỏ càng tốt)

$$AIC = -2 \log(L) + 2(p + q + k + 1), \quad (11.46)$$

trong đó, L là likelihood của dữ liệu, $k = 1$ nếu $c \neq 0$ và $k = 0$ nếu $c = 0$

- Chỉ số AIC hiệu chỉnh

$$AICc = AIC + \frac{2(p + q + k + 1)(p + q + k + 2)}{T - p - q - k - 2}, \quad (11.47)$$

trong đó T là chiều dài của chuỗi dữ liệu

- Chỉ số BIC (Bayesian Information Criterion) (càng nhỏ càng tốt)

$$BIC = AIC + [\log(T) - 2](p + q + k + 1). \quad (11.48)$$

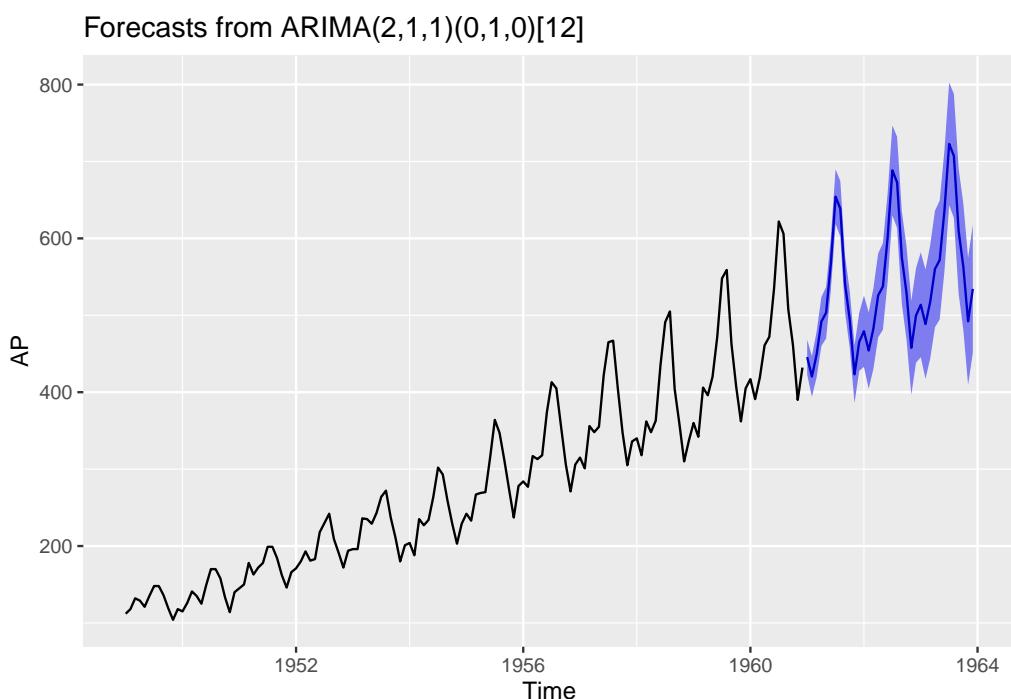
11. Phân tích dữ liệu chuỗi thời gian

11.8.2. Ví dụ hồi quy mô hình và áp dụng

Sử dụng phương pháp Box-Jenkins cho kết quả hồi quy mô hình ARIMA cho dữ liệu khách hàng không

SARIMAX Results						
<hr/>						
Dep. Variable:	#Passengers	No. Observations:	144			
Model:	ARIMA(2, 1, 1)x(0, 1, 0)[12]	Log Likelihood	-504.923			
Date:	Sun, 20 Jun 2021	AIC	1017.847			
Time:	14:23:57	BIC	1029.348			
Sample:	0 - 144	HQIC	1022.520			
Covariance Type:	opg					
<hr/>						
coef	std err	z	P> z	[0.025	0.975]	
ar.L1	0.5960	0.085	6.987	0.000	0.429	0.763
ar.L2	0.2143	0.091	2.343	0.019	0.035	0.394
ma.L1	-0.9819	0.038	-25.602	0.000	-1.057	-0.907
sigma2	129.3150	14.557	8.883	0.000	100.784	157.846
<hr/>						
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	7.68			
Prob(Q):	0.98	Prob(JB):	0.02			
Heteroskedasticity (H):	2.33	Skew:	-0.01			
Prob(H) (two-sided):	0.01	Kurtosis:	4.19			
<hr/>						

- Mô hình khá tốt các hệ số ước lượng đều có ý nghĩa thống kê (p -value rất nhỏ)
- Dự đoán số lượng khách hàng không trong vòng 36 tháng tiếp theo với khoảng tin cậy là 95% bằng trực quan



11.9 Hồi quy với lỗi ARIMA

Bài tập

B 11.1. Thu thập dữ liệu và sử dụng mô hình ARIMA trong dự báo chỉ số VN-Index và tỉ số sinh lợi của thị trường

12.1 Các loại dữ liệu

1. Dữ liệu chuỗi thời gian: Là các số liệu được thu thập trong một thời kỳ, một khoảng thời gian nhất định. Ví dụ như các số liệu về GDP, giá điện, dân số hay giá trị sản xuất...

Năm	CPI	Lai_suất	GTSX_CN	PA
2007M1	111.0	6.5	49,212.0	12,000.0
2007M2	113.4	6.5	35,392.0	12,000.0
2007M3	113.1	6.5	45,154.0	12,000.0
2007M4	113.7	6.5	47,344.6	12,000.0
2007M5	114.5	6.5	47,953.4	12,000.0

2. Dữ liệu chéo: Là các số liệu về một hoặc nhiều biến được thu thập tại một thời điểm ở nhiều địa phương, đơn vị khác nhau...

Tỉnh	GTSX_CN	GTSX_TM
Hà Nội	2345	1244
HCM	2436	1242
Đà Nẵng	3454	1222
Hải Phòng	2333	1234

3. Dữ liệu bảng: Là sự kết hợp giữ *dữ liệu chuỗi thời gian* và *dữ liệu chéo* (hỗn hợp theo không gian và thời gian). Ví dụ như các số liệu về giá USD hàng ngày ở Hà Nội, HCM...

Năm	Tỉnh	GTSX_TM	GTSX_CN
2008	Hà Nội	1244	4577
2009	Hà Nội	1242	4575
2010	Hà Nội	1222	4555
2011	Hà Nội	1111	4444
2008	HCM	2244	5577
2009	HCM	2242	5575
2010	HCM	1422	4755
2011	HCM	1151	4484

12.2 Các ưu điểm của dữ liệu bảng

Baltagi liệt kê các ưu điểm sau đây của dữ liệu bảng

- Vì dữ liệu bảng liên quan đến các cá nhân, doanh nghiệp, tiểu bang, đất nước ... theo thời gian, nên nhất định phải có tính đặc biệt (không đồng nhất) trong các đơn vị này. Kỹ thuật ước lượng dữ liệu bảng có thể chính thức xem xét đến tính đặc biệt đó bằng cách xem xét các biến số có tính đặc thù theo từng cá nhân, được trình bày ngay sau đây. Ta sử dụng thuật ngữ cá nhân theo ý nghĩa chung bao gồm các đơn vị vi mô như các cá nhân, các doanh nghiệp, tiểu bang, và đất nước.
- Thông qua kết hợp các chuỗi theo thời gian của các quan sát theo không gian, dữ liệu bảng cung cấp ‘những dữ liệu có nhiều thông tin hơn, đa dạng hơn, ít cộng tuyến hơn giữa các biến số, nhiều bậc tự do hơn và hiệu quả hơn.’
- Thông qua nghiên cứu các quan sát theo không gian lặp lại, dữ liệu bảng phù hợp hơn để nghiên cứu tính động của thay đổi. Tình trạng thất nghiệp, luân chuyển công việc, và tính lưu chuyển lao động sẽ được nghiên cứu tốt hơn với dữ liệu bảng.
- Dữ liệu bảng có thể phát hiện và đo lường tốt hơn những ảnh hưởng mà không thể quan sát trong dữ liệu chuỗi thời gian thuần túy hay dữ liệu chéo theo không gian thuần túy. Ví dụ, ảnh hưởng của luật tiền lương tối thiểu đối với việc làm và thu nhập có thể được nghiên cứu tốt hơn nếu chúng ta xem xét các đợt gia tăng tiền lương tối thiểu liên tiếp nhau trong mức lương tối thiểu của liên bang và (hoặc) tiểu bang.
- Dữ liệu bảng giúp ta nghiên cứu những mô hình hành vi phức tạp hơn. Ví dụ, các hiện tượng như lợi thế kinh tế theo qui mô và thay đổi kỹ thuật có thể được xem xét thông qua dữ liệu bảng tốt hơn so với dữ liệu theo chuỗi thời gian thuần túy hay theo không gian thuần túy.
- Bằng cách thu thập những số liệu có sẵn cho vài nghìn đơn vị, dữ liệu bảng có thể tối thiểu hóa sự thiên lệch có thể xảy ra nếu ta tổng hợp các cá nhân hay các doanh nghiệp thành số liệu tổng.

12.3 Mô hình VAR

12.4 Mô hình hồi quy

Ví dụ. Xem xét biến *tổng đầu tư thực* (Y) phụ thuộc như thế nào vào *giá trị thực của doanh nghiệp* (X_1) và *trữ lượng vốn thực* (X_3). Dữ liệu công ty General Electric (GE), General Motor

12. Phân tích dữ liệu bảng

(GM), US Steel (US), và Westinghouse (WEST) trong giai đoạn 1935-1954 được trình bày trong bảng sau

Year	GE			US			GM			WEST		
	Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2
1935	33.1	1170.6	97.8	209.9	1362.4	53.8	317.6	3078.5	2.8	12.93	191.5	1.8
1936	45.0	2015.8	104.4	355.3	1807.1	50.5	391.8	4661.7	52.6	25.90	516.0	0.8
1937	77.2	2803.3	118.0	469.9	2673.3	118.1	410.6	5387.1	156.9	35.05	729.0	7.4
1938	44.6	2039.7	156.2	262.3	1801.9	260.2	257.7	2792.2	209.2	22.89	560.4	18.1
1939	48.1	2256.2	172.6	230.4	1957.3	312.7	330.8	4313.2	203.4	18.84	519.9	23.5
1940	74.4	2132.2	186.6	361.6	2202.9	254.2	461.2	4643.9	207.2	28.57	628.5	26.5
1941	113.0	1834.1	220.9	472.8	2380.5	261.4	512.0	4551.2	255.2	48.51	537.1	36.2
1942	91.9	1588.0	287.8	445.6	2168.6	298.7	448.0	3244.1	303.7	43.34	561.2	60.8
1943	61.3	1749.4	319.9	361.6	1985.1	301.8	499.6	4053.7	264.1	37.02	617.2	84.4
1944	56.8	1687.2	321.3	288.2	1813.9	279.1	547.5	4379.3	201.6	37.81	626.7	91.2
1945	93.6	2007.7	319.6	258.7	1850.2	213.8	561.2	4840.9	265.0	39.27	737.2	92.4
1946	159.9	2208.3	346.0	420.3	2067.7	232.6	688.1	4900.0	402.0	53.46	760.5	86.0
1947	147.2	1656.7	456.4	420.5	1796.3	264.8	568.9	3526.5	761.5	55.56	581.4	111.1
1948	146.3	1604.4	543.4	494.5	1625.8	306.9	529.2	3245.7	922.4	49.56	662.3	130.6
1949	98.3	1431.8	618.3	405.1	1667.0	351.1	555.1	3700.2	1020.1	32.04	583.8	141.8
1950	93.5	1610.5	647.4	418.8	1677.4	357.8	642.9	3755.6	1099.0	32.24	635.2	136.7
1951	135.2	1819.4	671.3	588.2	2289.5	341.1	755.9	4833.0	1207.7	54.38	732.8	129.7
1952	157.3	2079.7	726.1	645.2	2159.4	444.2	891.2	4926.9	1430.5	71.78	864.1	145.5
1953	179.5	2371.6	800.3	641.0	2031.3	623.6	1304.4	6241.7	1777.3	90.08	1193.5	174.8
1954	189.6	2759.9	888.9	459.3	2115.5	669.7	1486.7	5593.6	226.3	68.60	1188.9	213.5

Ta quy định rằng N là số cá nhân tối đa, T là số lượng đoạn thời gian tối đa. Nếu mỗi đơn vị theo không gian có cùng một số lượng quan sát như nhau theo chuỗi thời gian, thì dữ liệu bảng này được gọi là **bảng cân đối**. Nếu số quan sát khác nhau giữa các đơn vị, thì đó là **bảng không cân đối**. Theo qui ước, chọn i là ký hiệu đơn vị theo không gian và t là ký hiệu theo thời gian. Bảng dữ liệu trên có

- Giá trị $i = 1, \dots, 4$ biểu diễn cho cá nhân thứ i ; nghĩa là $i = 1$ là công ty GM
- Giá trị $t = 1, \dots, 20$ biểu diễn cho năm thứ t ; nghĩa là $t = 1$ là năm 1935

12.5 Mô hình ảnh hưởng cố định (Fixed effects model)

Việc xây dựng mô hình phụ thuộc vào những giả định về hệ số chặn và các hệ số góc

12. Phân tích dữ liệu bảng

1. Các hệ số góc và hệ số chẵn là giống nhau.
2. Các hệ số góc là giống nhau nhưng hệ số chẵn phụ thuộc vào các cá nhân.
3. Các hệ số góc là giống nhau nhưng hệ số chẵn phụ thuộc vào các cá nhân và thời gian.
4. Tất cả các hệ số (hệ số góc cũng như các hệ số chẵn) đều phụ thuộc vào các cá nhân.
5. Tất cả các hệ số (hệ số góc cũng như các hệ số chẵn) đều phụ thuộc vào các cá nhân và thời gian.

12.6

Mô hình ảnh hưởng ngẫu nhiên (Random effects model)

Bài tập

13.1 Một số khái niệm

- Đối tượng (*object*) là tất cả những sự vật, sự kiện mà hoạt động của con người có liên quan tới.
- Hệ thống (*system*) là tập hợp các đối tượng (con người, máy móc), sự kiện mà giữa chúng có những mối quan hệ nhất định.
- Trạng thái của hệ thống (*state of system*) là tập hợp các tham số, biến số dùng để mô tả hệ thống tại một thời điểm và trong điều kiện nhất định.
- Mô hình (*model*) là một sơ đồ phản ánh đối tượng, con người dùng sơ đồ đó để nghiên cứu, thực nghiệm nhằm tìm ra quy luật hoạt động của đối tượng hay nói cách khác mô hình là đối tượng thay thế của đối tượng gốc để nghiên cứu về đối tượng gốc.
- Mô hình hóa (*modeling*) là thay thế đối tượng gốc bằng một mô hình nhằm các thu nhận thông tin quan trọng về đối tượng bằng cách tiến hành các thực nghiệm trên mô hình. Lý thuyết xây dựng mô hình và nghiên cứu mô hình để hiểu biết về đối tượng gốc gọi lý thuyết mô hình hóa. Nếu các quá trình xảy ra trong mô hình đồng nhất (theo các chỉ tiêu định trước) với các quá trình xảy ra trong đối tượng gốc thì người ta nói rằng mô hình đồng nhất với đối tượng. Lúc này người ta có thể tiến hành các thực nghiệm trên mô hình để thu nhận thông tin về đối tượng.
- Mô phỏng (*simulation, imitation*) là phương pháp mô hình hóa dựa trên việc xây dựng mô hình số (*numerical model*) và dùng phương pháp số (*numerical method*) để tìm các lời giải. Chính vì vậy máy tính số là công cụ hữu hiệu và duy nhất để thực hiện việc mô phỏng hệ thống.

Lý thuyết cũng như thực nghiệm đã chứng minh rằng, chỉ có thể xây dựng được mô hình gần đúng với đối tượng mà thôi, vì trong quá trình mô hình hóa bao giờ cũng phải chấp nhận một số giả thiết nhằm giảm bớt độ phức tạp của mô hình, để mô hình có thể ứng dụng thuận tiện

trong thực tế. Mặc dù vậy, mô hình hóa luôn luôn là một phương pháp hữu hiệu để con người nghiên cứu đối tượng, nhận biết các quá trình, các quy luật tự nhiên. Đặc biệt, ngày nay với sự trợ giúp đắc lực của khoa học kỹ thuật, nhất là khoa học máy tính và công nghệ thông tin, người ta đã phát triển các phương pháp mô hình hóa cho phép xây dựng các mô hình ngày càng gần với đối tượng nghiên cứu, đồng thời việc thu nhận, lựa chọn, xử lý các thông tin về mô hình rất thuận tiện, nhanh chóng và chính xác. Chính vì vậy, mô hình hóa là một phương pháp nghiên cứu khoa học mà tất cả những người làm khoa học đều phải nghiên cứu và ứng dụng vào thực tiễn hoạt động của mình.

13.2 Mô hình hóa hệ thống

13.2.1. Tại sao cần mô hình hóa hệ thống

1. Khi nghiên cứu trên hệ thống thực gặp nhiều khó khăn do nhiều nguyên nhân
 - Giá thành nghiên cứu trên hệ thống thực quá đắt.
 - Nghiên cứu trên hệ thống thực đòi hỏi thời gian quá dài.
 - Nghiên cứu trên hệ thực ảnh hưởng đến sản xuất hoặc gây nguy hiểm cho người và thiết bị.
 - Trong một số trường hợp không cho phép làm thực nghiệm trên hệ thống thực.
2. Phương pháp mô hình hóa cho phép đánh giá hệ thống khi thay đổi tham số hoặc cấu trúc của hệ thống cũng như đánh giá phản ứng của hệ thống khi thay đổi tín hiệu điều khiển. Những số liệu này dùng để thiết kế hệ thống hoặc lựa chọn thông số tối ưu để vận hành hệ thống.
3. Phương pháp mô hình hóa cho phép nghiên cứu hệ thống ngay cả khi chưa có hệ thống thực

Mô hình phải đạt được hai tính chất cơ bản sau:

- Tính đồng nhất: mô hình phải đồng nhất với đối tượng mà nó phản ánh theo những tiêu chuẩn định trước.
- Tính thực dụng: có khả năng sử dụng mô hình để nghiên cứu hệ thống.

13.2.2. Các loại mô hình hóa hệ thống

Có hai loại mô hình, *mô hình vật lý* và *mô hình toán học*.

- Mô hình vật lý là mô hình được cấu tạo bởi các phần tử vật lý. Các thuộc tính của đối tượng phản ánh các định luật vật lý xảy ra trong mô hình. Mô hình vật lý bao gồm *mô hình thu nhỏ* và *mô hình tương tự*.

- Mô hình toán học thuộc loại mô hình trừu tượng. Các thuộc tính được phản ánh bằng các biểu thức, phương trình toán học. Mô hình toán học được chia thành *mô hình giải tích* và *mô hình số*. Mô hình số được xây dựng bằng các chương trình chạy trên máy tính. Ngày nay, nhờ sự phát triển của kỹ thuật máy tính và công nghệ thông tin, người ta đã xây dựng được các mô hình số có thể mô phỏng được quá trình hoạt động của đối tượng thực. Những mô hình loại này được gọi là *mô hình mô phỏng*.

13.3 Phương pháp mô phỏng

Định nghĩa. Mô phỏng là quá trình xây dựng mô hình toán học của hệ thống thực và sau đó tiến hành tính toán thực nghiệm trên mô hình để mô tả, giải thích và dự đoán hành vi của hệ thống thực

Phương pháp mô phỏng được đề xuất vào những năm 80 của thế kỷ 20. Từ đó đến nay phương pháp mô phỏng đã được nghiên cứu, hoàn thiện, và ứng dụng thành công vào nhiều lĩnh vực khác nhau như lĩnh vực khoa học kỹ thuật, khoa học xã hội, kinh tế, y tế... Sau đây trình bày một số lĩnh vực mà phương pháp mô phỏng đã được ứng dụng và phát huy được ưu thế của mình.

- Phân tích và thiết kế hệ thống sản xuất, lập kế hoạch sản xuất.
- Đánh giá phần cứng, phần mềm của hệ thống máy tính.
- Quản lý và xác định chính sách sự trữ mua sắm vật tư của hệ thống kho vật tư, nguyên liệu.
- Phân tích và đánh giá hệ thống phòng thủ quân sự, xác định chiến lược phòng thủ, tấn công.
- Phân tích và thiết kế hệ thống thông tin liên lạc, đánh giá khả năng làm việc của mạng thông tin.
- Phân tích và thiết kế các hệ thống giao thông như đường sắt, đường bộ, hàng không, cảng biển.
- Đánh giá, phân tích và thiết kế các cơ sở dịch vụ như bệnh viện, bưu điện, nhà hàng, siêu thị.
- Phân tích hệ thống kinh tế, tài chính.
- Dự báo thời tiết.
- Mô phỏng hoạt động của các máy gia tốc, vụ nổ hạt nhân.

13. Phương pháp mô phỏng Monte Carlo

- Lập trình trò chơi (*game programming*), học tăng cường (*reinforcement learning*).



Phương pháp mô phỏng được ứng dụng vào các giai đoạn khác nhau của việc nghiên cứu, thiết kế và vận hành các hệ thống như sau:

- Phương pháp mô phỏng được ứng dụng vào giai đoạn nghiên cứu, khảo sát hệ thống trước khi tiến hành thiết kế nhằm xác định độ nhạy của hệ thống đối với sự thay đổi cấu trúc và tham số của hệ thống.
- Phương pháp mô phỏng được ứng dụng vào giai đoạn thiết kế hệ thống để phân tích và tổng hợp các phương án thiết kế hệ thống, lựa chọn cấu trúc hệ thống thỏa mãn các chỉ tiêu cho trước.
- Phương pháp mô phỏng được ứng dụng vào giai đoạn vận hành hệ thống để đánh giá khả năng hoạt động, giải bài toán vận hành tối ưu, chẩn đoán các trạng thái đặc biệt của hệ thống.

13.4 Phương pháp mô phỏng Monte Carlo

Phương pháp Monte Carlo được xây dựng dựa trên nền tảng

- Các số ngẫu nhiên** (*random numbers*): đây là nền tảng quan trọng, góp phần hình thành nên “thương hiệu” của phương pháp. Các số ngẫu nhiên không chỉ được sử dụng trong việc mô phỏng lại các hiện tượng ngẫu nhiên xảy ra trong thực tế mà còn được sử dụng để lấy mẫu ngẫu nhiên của một phân bố nào đó, chẳng hạn như trong tính toán các tích phân số.
- Luật số lớn** (*law of large numbers*): luật này đảm bảo rằng khi ta chọn ngẫu nhiên các giá trị (mẫu thử) trong một dãy các giá trị (quần thể), kích thước dãy mẫu thử càng lớn thì các đặc trưng thống kê (trung bình, phương sai,...) của mẫu thử càng “gần” với các

13. Phương pháp mô phỏng Monte Carlo

đặc trưng thống kê của quần thể. Luật số lớn rất quan trọng đối với phương pháp Monte Carlo vì nó đảm bảo cho sự ổn định của các giá trị trung bình của các biến ngẫu nhiên khi số phép thử đủ lớn.

- **Định lý giới hạn trung tâm** (*central limit theorem*): định lý này phát biểu rằng dưới một số điều kiện cụ thể, trung bình số học của một lượng đủ lớn các phép lặp của các biến ngẫu nhiên độc lập (independent random variables) sẽ được xấp xỉ theo phân bố chuẩn (normal distribution). Do phương pháp Monte Carlo là một chuỗi các phép thử được lặp lại nên định lý giới hạn trung tâm sẽ giúp chúng ta dễ dàng xấp xỉ được trung bình và phương sai của các kết quả thu được từ phương pháp.

13.4.1. Các loại số ngẫu nhiên

Có 3 loại số ngẫu nhiên chính

- Số ngẫu nhiên thực (*real random number*): các đại lượng của các hiện tượng ngẫu nhiên trong thế giới tự nhiên.
- Số gần ngẫu nhiên (*quasi-random number*): các đại lượng có phân bố tốt (có sự không nhất quán thấp).
- Số giả ngẫu nhiên (*pseudo-random number*): các đại lượng mà phân bố của nó vượt qua được các kiểm tra về tính ngẫu nhiên.

Có hai điều chúng ta cần lưu ý khi mô phỏng các số ngẫu nhiên

- Máy tính không thể tạo ra các dãy số ngẫu nhiên thật sự mà chỉ là các số giả ngẫu nhiên.
- Bản thân các số không phải là ngẫu nhiên mà chỉ có dãy số mới có thể được xem là ngẫu nhiên

Một dãy số ngẫu nhiên tốt phải hội tụ đầy đủ các yếu tố sau đây

- Chu kì lặp lại phải dài tức là việc gieo số ngẫu nhiên phải tạo ra được nhiều số trước khi lặp lại dãy số cũ của nó để cho không có phần nào của dãy bị trùng trong tính toán.
- Các số được tạo ra phải hướng tới phân bố đều, tức là một dãy số bất kì gồm vài trăm số phải hướng tới phân bố đồng nhất trong toàn vùng khảo sát.
- Các số không tương quan với nhau, tức là các số trong dãy phải độc lập về mặt thống kê với các số trước nó.
- Thuật toán phải nhanh, tức là thời gian máy tính tạo ra số ngẫu nhiên phải nhỏ.

Các số giả ngẫu nhiên trong phương pháp Monte Carlo chỉ cần tỏ ra “đủ mức ngẫu nhiên”, nghĩa là tuân theo phân bố đều hay theo phân bố định trước, khi số lượng của chúng lớn.

13.4.2. Phương pháp tạo số giả ngẫu nhiên và lấy mẫu

Phương pháp đồng dư tuyến tính

Để tạo ra dãy số giả ngẫu nhiên độc lập từ một phân phối đều $X_1, X_2, \dots, X_n \sim \mathcal{U}(0, 1)$, ta sử dụng thuật toán sau

Tạo n số giả ngẫu nhiên

1. Khởi tạo hạt (*seed*) x_0
2. Lặp $i = 1 \dots n$, tính x_i bằng công thức truy hồi

$$x_i = (ax_{i-1} + c) \bmod m \quad (13.1)$$

3. Chuẩn hóa giá trị bằng $x_i/m, i = 1 \dots n$

- Các giá trị đề xuất $m = 2^{32}$, $a = 1\,664\,525$ và $c = 1\,013\,904\,223$

Thực sự đây không phải là một thuật toán tạo số ngẫu nhiên tốt nhất nhưng ưu điểm của thuật toán này là đơn giản, dễ sử dụng, tính toán nhanh và dãy số ngẫu nhiên do nó tạo ra là khá tốt.

Ta có thể thấy rằng trong dãy số được tạo ra bởi phương pháp này mỗi số chỉ có thể xuất hiện duy nhất một lần trước khi dãy bị lặp lại.

Để tạo ra số ngẫu nhiên từ phân phối chuẩn $\mathcal{N}(0, 1)$ ta sử dụng thuật toán

Tạo số ngẫu nhiên từ phân phối chuẩn

1. Tạo số ngẫu nhiên $u_1, u_2 \sim \mathcal{U}(0, 1)$
2. Tính $\theta = 2\pi u_1$ và $r = \sqrt{-2 \ln(u_2)}$
3. Tính $x = r \cos(\theta)$ và $y = r \sin(\theta)$ là các đại lượng ngẫu nhiên độc lập của phân phối $\mathcal{N}(0, 1)$

Phương pháp biến đổi ngược (inverse transform method)

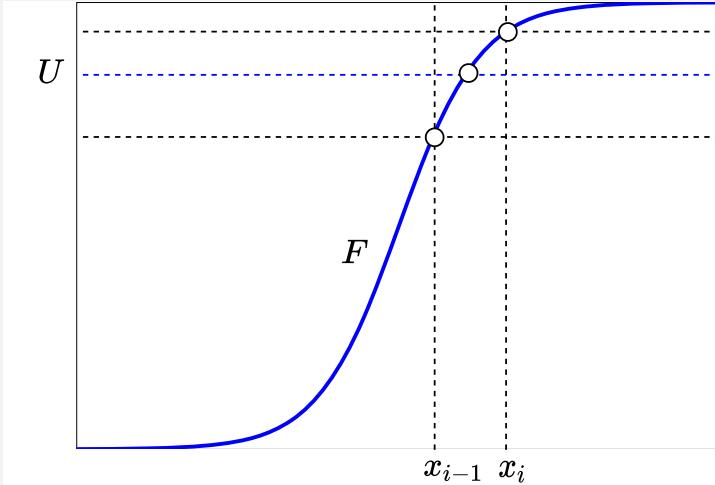
Giả sử ta cần tạo giá trị của một biến ngẫu nhiên rời rạc X có hàm phân bố xác suất

$$p_i = P(X = x_i), \quad i = 0, 1, \dots \quad (13.2)$$

Gọi F là hàm phân phối tích lũy

Tạo số ngẫu nhiên từ phân phối rời rạc

1. Tạo số ngẫu nhiên $u \sim \mathcal{U}(0, 1)$
2. Chọn $X = x_i$ nếu $F(x_{i-1}) \leq u < F(x_i)$



Ví dụ. Tạo một biến Poisson ngẫu nhiên với tham số λ

$$p_i = P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, \dots \quad (13.3)$$

1. Tạo số ngẫu nhiên $u \sim \mathcal{U}(0, 1)$

2. $i = 0, \alpha = e^{-\lambda}, F = \alpha$

3. Nếu $u < F$ thì

Chọn $X = i$ và dừng

ngược lại

$$i = i + 1, \alpha = \frac{\lambda \alpha}{i}, F = F + \alpha \text{ và lặp lại (c)}$$

Phương pháp bắc bỏ (acceptance-rejection method)

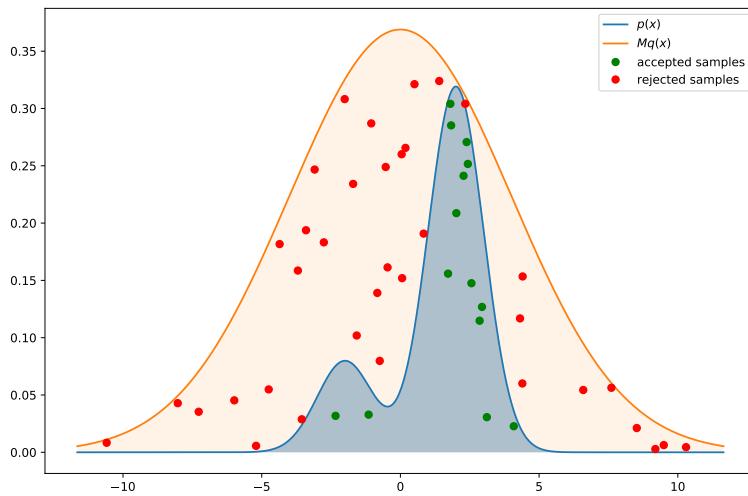
Giả sử ta cần tạo giá trị của một biến ngẫu X ~ p(x)

- $p(x)$ là phân phối mục tiêu
- $q(x)$ là phân phối dễ lấy mẫu
- M là hằng số sao cho $\forall x \in \mathcal{X}, p(x) \leq Mq(x)$

13. Phương pháp mô phỏng Monte Carlo

Tạo số ngẫu nhiên từ phân phối bất kỳ

1. Tạo số ngẫu nhiên $x \sim q(x)$
2. Tạo số ngẫu nhiên $u \sim \mathcal{U}(0, Mq(x))$
3. Nếu $u < p(x)$ thì chọn x ; ngược lại bác bỏ x



Phương pháp importance sampling

Đây là một phương pháp ước lượng giá trị kỳ vọng (hoặc các đại lượng thống kê quan tâm) của hàm $f(x)$ trong đó x có phân phối p . Tuy nhiên, Thay vì lấy mẫu từ p , chúng ta sẽ tính kết quả từ việc lấy mẫu từ phân phối q . Phương pháp này thường được xem là một kĩ thuật giảm phương sai trong lấy mẫu Monte Carlo.

Ta đã biết rằng kỳ vọng của $f(x)$ được tính theo công thức

$$\mathbb{E}_{x \sim p}[f(x)] = \int f(x)p(x)dx \quad (13.4)$$

Thay vì lấy mẫu biến x từ phân bố p , ta sẽ đi lấy mẫu theo một phân bố q đơn giản hơn, khi đó kỳ vọng của $f(x)$ được tính lại theo công thức

$$\begin{aligned} \mathbb{E}_{x \sim p}[f(x)] &= \int f(x)p(x)dx \\ &= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\ &= \mathbb{E}_{x \sim q}\left[f(x)\frac{p(x)}{q(x)}\right] \end{aligned} \quad (13.5)$$

13.4.3. Phương pháp Monte Carlo

Các thành phần chính của phương pháp mô phỏng Monte Carlo gồm có

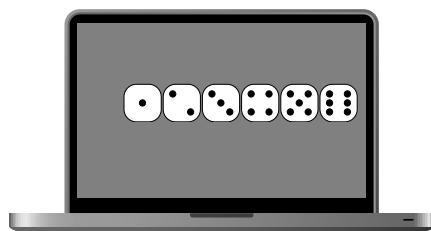
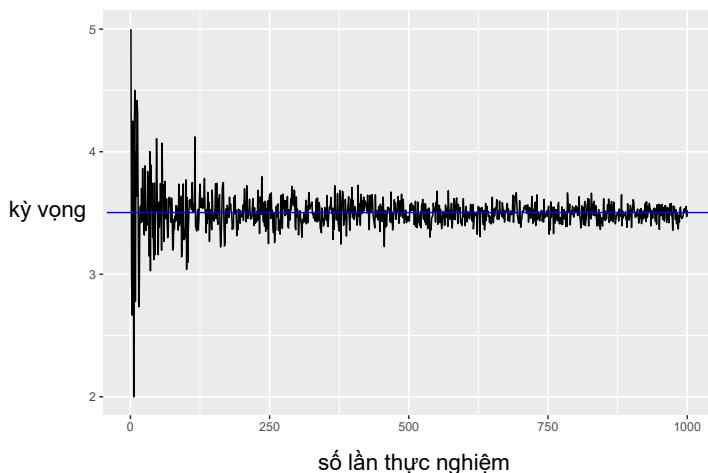
13. Phương pháp mô phỏng Monte Carlo

- Hàm mật độ xác suất (*probability density function - PDF*): một hệ vật lý (hay toán học) phải được mô tả bằng một bộ các PDF.
- Nguồn phát số ngẫu nhiên (*random number generator - RNG*): Tạo ra các số ngẫu nhiên theo các phân phối PDF.
- Ghi nhận (*scoring*): dữ liệu đầu ra phải được tích luỹ trong các khoảng giá trị của đại lượng cần quan tâm.
- Ước lượng sai số (*error estimation*): ước lượng sai số thống kê (phương sai) theo số phép thử và theo đại lượng quan tâm.

Để tăng hiệu quả cho phương pháp Monte Carlo, chúng ta có thể sử dụng

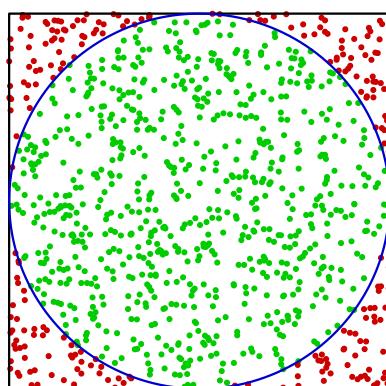
- Các kỹ thuật giảm phương sai (*variance reduction technique*): các phương pháp nhằm giảm phương sai của đáp số được ước lượng để giảm thời gian tính toán của mô phỏng Monte Carlo.
- Song song hoá (*parallelization*) và vector hoá (*vectorization*): các thuật toán cho phép phương pháp Monte Carlo được thực thi một cách hiệu quả trên một cấu trúc máy tính hiện nay.

Ví dụ. Ước lượng kỳ vọng của mặt xúc xắc



chương trình mô phỏng

Ví dụ. Tính giá trị π .



13. Phương pháp mô phỏng Monte Carlo

- Sử dụng phân phối đều \mathcal{U} để tạo ra các điểm ngẫu nhiên nằm trong hình vuông
- Tính tỉ số

$$\pi \approx 4 \times \frac{\text{số điểm nằm trong hình tròn}}{\text{số điểm nằm trong hình vuông}}$$

Bài toán. Công ty A đang xem xét việc thuê một thiết bị hiện đại dùng trong sản xuất. Hợp đồng thuê một năm là 400000 đô la và không thể hủy bỏ sớm. Công ty đang đánh giá liệu *mức sản xuất* hàng năm và mức tiết kiệm trong *bảo trì*, *lao động* và *nguyên vật liệu* có đủ bù đắp cho việc thuê máy này không? Từ các chuyên gia về kinh tế, công ty có được các thông tin sau

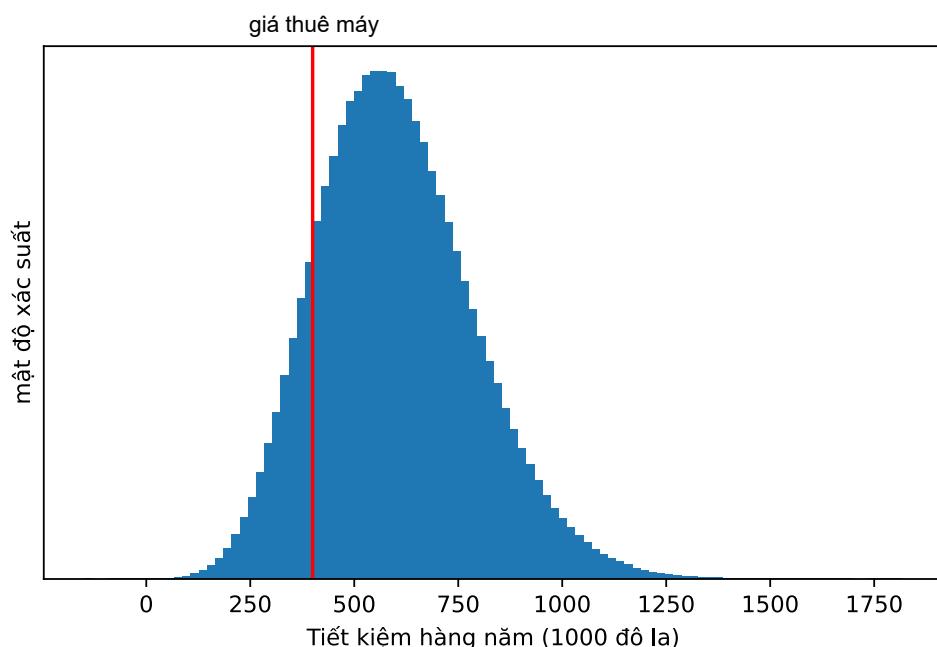
- Tiết kiệm bảo trì (*MS*): 10-20 USD mỗi đơn vị sản phẩm
- Tiết kiệm lao động (*LS*): 2-8 USD mỗi đơn vị sản phẩm
- Tiết kiệm nguyên vật liệu (*RMS*): 3-9 USD mỗi đơn vị sản phẩm
- Mức sản xuất (*PL*): 15000-35000 đơn vị sản phẩm mỗi năm

Lưu ý: tất cả các biến đều có phân phối chuẩn và có khoảng tin cậy là 90%

Như vậy, ta có tiết kiệm trong một năm là

$$\text{Tiết kiệm hàng năm} = (MS + LS + RMS) \times PL \quad (13.6)$$

Sử dụng phương pháp Monte Carlo chạy mô phỏng



Bài tập

B 13.1. Giải các bài tập chương 2, 3, 4 bằng phương pháp Monte Carlo

13. Phương pháp mô phỏng Monte Carlo

B 13.2. Giả sử bạn đang quản lý dự án liên quan đến việc tạo ra module e-learning. Việc tạo ra module e-learning bao gồm ba công việc (task) theo *tuần tự*: viết nội dung, tạo đồ họa, và tích hợp các yếu tố đa phương tiện. Dựa trên kinh nghiệm chuyên môn ta có ước lượng thời gian hoàn thành cho từng công việc như sau

Công việc	trường hợp tốt nhất	khả thi nhất	trường hợp xấu nhất
viết nội dung	4	6	8
làm hình ảnh	5	7	9
tích hợp	2	4	6

Hãy xác định khả năng hoàn thành dự án trong 17 ngày?

B 13.3. Giả sử bạn đang kinh doanh một dịch vụ dựa trên đăng ký trực tuyến.

- Phí đăng ký là 10 đô la mỗi tháng cho mỗi người dùng và có thể bị hủy bỏ bất kỳ tháng nào.
- Hiện tại đang có 10000 người dùng.
- Doanh thu định kỳ hàng tháng là

$$\text{số người dùng} \times \text{phí đăng ký}.$$

- Chi phí duy trì dịch vụ hàng năm là cố định ở mức 1 triệu đô la mỗi năm bất kể người dùng. Bảng dưới đây là tỉ lệ tăng trưởng người dùng từng tháng (có thể tăng hoặc giảm) trong 5 năm qua

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	-2.0%	3.5%	0.5%	2.2%	1.9%	2.8%	2.2%	0.2%	5.0%	3.0%	3.6%	-4.6%
2017	-11.4%	4.7%	-3.1%	-3.9%	-4.3%	-1.4%	-4.5%	4.4%	7.7%	10.3%	2.3%	-1.4%
2018	1.3%	-5.8%	5.2%	-8.6%	-4.1%	-1.5%	2.4%	-2.7%	-3.3%	-0.6%	6.0%	-3.7%
2019	0.0%	1.1%	2.8%	0.3%	9.6%	-6.7%	1.0%	-1.3%	-4.3%	-3.0%	-1.2%	-8.3%
2020	-7.9%	-0.2%	0.9%	1.0%	-4.6%	2.0%	-0.9%	7.3%	-4.5%	-7.9%	-4.1%	5.6%

Bạn kỳ vọng 12 tháng trong năm tới sẽ *giống* với bất kỳ tháng nào trong 60 tháng trước đó. Hãy xác định khả năng lặp lại khi tiếp tục vận hành dịch vụ này trong năm tới hay không?

Tài liệu tham khảo

- [1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [2] R. C. Hill, W. E. Griffiths, and G. C. Lim. *Principles of econometrics*. John Wiley & Sons, 2018.
- [3] Nguyễn Văn Thìn. Bài tập xác suất và thống kê toán. Đại học Khoa học Tự nhiên, 2010.
- [4] Nguyễn Văn Tuấn. *Mô hình hồi qui và khám phá khoa học*. Nhà xuất bản TPHCM, 2020.
- [5] Nguyễn Văn Tuấn. *Phân tích dữ liệu với R*. Nhà xuất bản TPHCM, 2020.
- [6] J. H. Stock and M. W. Watson. *Introduction to econometrics 3rd ed.* Pearson Education, 2015.
- [7] Vũ Duy Thành. Bài giảng kinh tế lượng. Đại học Kinh tế Quốc dân, 2015.
- [8] Vũ Hữu Tiệp. Machine learning cơ bản, 2020. URL: <https://machinelearningcoban.com>.
- [9] Đỗ Minh Hải. Blog, 2018. URL: <https://dominhhai.github.io>.