Trần Minh Tâm

XÂY DỰNG MÔ HÌNH KINH TẾ LƯỢNG HIỆN ĐẠI VỚI PYTHON

TP. Hồ Chí Minh - 2025

Mục lục

Phần I.

Tổng quan về kinh tế lượng và Python

Chương 1.

Giới thiệu kinh tế lượng và vai trò của Python

1.1. Kinh tế lượng là gì?

Kinh tế lượng là lĩnh vực kết hợp giữa kinh tế học, thống kê và toán học để phân tích dữ liệu kinh tế. Kinh tế lượng đóng vai trò quan trọng trong nghiên cứu dữ liệu **?**. Phương pháp thực nghiệm trong kinh tế lượng được đề xuất trong **?**.

1.2. Tại sao sử dụng Python trong kinh tế lượng?

Python ngày càng phổ biến trong nghiên cứu kinh tế lượng vì cú pháp dễ hiểu và hệ sinh thái phong phú.

1.3. Các thư viện quan trọng

Dưới đây là một số thư viện quan trọng trong Python cho kinh tế lượng:

- **NumPy**: Xử lý ma trận và tính toán số học.
- **Pandas**: Xử lý dữ liệu dạng bảng.
- **Statsmodels**: Thực hiện các mô hình kinh tế lượng.

1.4. Cài đặt môi trường làm việc

1.4.1. Cài đặt Python và Anaconda

Python là một ngôn ngữ lập trình mạnh mẽ và dễ dàng sử dụng trong kinh tế lượng **?**. Để cài đặt Python, ta nên sử dụng Anaconda, một phần phối chứa sẵn nhiều thư viện Python hữu ích cho phân tích dữ liệu và kinh tế lượng **?**.

Các bước cài đặt Anaconda

1. Truy cập trang chủ Anaconda: `https://www.anaconda.com/`

2. Tải và cài đặt phiên bản phù hợp với hệ điều hành.
3. Kiểm tra cài đặt bằng lệnh:

```
conda --version
python --version
```

1.4.2. Thiết lập môi trường làm việc bằng Jupyter Notebook

Jupyter Notebook là công cụ hữu ích cho phân tích dữ liệu, lập trình và trình bày kết quả khoa học **?**. **Cài đặt Jupyter Notebook** Sau khi cài đặt Anaconda, có thể chạy Jupyter Notebook bằng lệnh:

```
jupyter notebook
```

Hoặc cài đặt riêng:

```
pip install notebook
```

1.4.3. Hướng dẫn cài đặt thư viện

Các thư viện quan trọng trong kinh tế lượng bao gồm:

- **numpy**: Toán học ma trận.
- **pandas**: Xử lý dữ liệu dạng bảng.
- **statsmodels**: Hồi quy và phân tích dữ liệu.
- **scipy**: Các công cụ toán học.
- **matplotlib**: Vẽ đồ thị.
- **seaborn**: Trực quan hóa dữ liệu.

Chương 2.

Xử lý dữ liệu trong kinh tế lượng

2.1. Tổng quan về dữ liệu

2.1.1. Khái niệm

2.1.2. Phân loại dữ liệu

2.1.3. Dữ liệu trong kinh tế lượng hiện đại

- **Dữ liệu chéo (Cross Sectional Data)**
- **Dữ liệu chuỗi thời gian (Time Series Data)**
- **Dữ liệu chéo gộp (Pooled Cross Sectional Data)**
- **Dữ liệu bảng (Panel Data)**
- **Dữ liệu không gian (Spatial Data)**
- **Dữ liệu tần số cao (High-Frequency Data)**
- **Dữ liệu văn bản (Text Data)**

2.2. Các phương pháp đo lường dữ liệu

2.2.1. Đo lường mức độ tập trung

a. Trung bình (Mean)

Định nghĩa: Trung bình (Mean) là một đại lượng đo lường xu hướng trung tâm của dữ liệu. Nó cho biết giá trị đại diện của một tập hợp dữ liệu bằng cách lấy tổng tất cả các giá trị chia cho số lượng phần tử.

Công thức tính trung bình:

- **Trung bình số học (Arithmetic Mean)**

  Trung bình số học của một tập dữ liệu gồm n quan sát $x_1, x_2, \ldots, x_n$ được tính bằng công thức:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2..1}$$

  Ví dụ: Tập dữ liệu: $\{10, 20, 30, 40, 50\}$

$$\bar{x} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30 \tag{2..2}$$

- **Trung bình có trọng số (Weighted Mean)**

  Nếu mỗi giá trị x_i có trọng số tương ứng w_i, thì trung bình có trọng số được tính bằng:

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \tag{2..3}$$

  Ví dụ: Điểm của sinh viên:

  – Toán (trọng số 4, điểm 8)

  – Lý (trọng số 3, điểm 7)

  – Hóa (trọng số 2, điểm 6)

$$\bar{x}_w = \frac{(4 \times 8) + (3 \times 7) + (2 \times 6)}{4 + 3 + 2} = \frac{32 + 21 + 12}{9} = \frac{65}{9} \approx 7.22 \tag{2..4}$$

- **Trung bình hình học (Geometric Mean)**

  Dùng khi dữ liệu có dạng tăng trưởng theo cấp số nhân:

$$GM = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = (x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}} \tag{2..5}$$

  Ví dụ: Tăng trưởng doanh thu qua 3 năm là 5%, 10%, 15%, trung bình hình học là:

$$GM = (1.05 \times 1.10 \times 1.15)^{\frac{1}{3}} \approx 1.096 \tag{2..6}$$

  Tức là mức tăng trung bình mỗi năm khoảng 9.6%.

- **Trung bình điều hòa (Harmonic Mean)**

  Dùng khi dữ liệu là tốc độ hoặc tỷ lệ:

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{2..7}$$

  Ví dụ: Nếu một ô tô đi 60 km/h trong 1 giờ và 40 km/h trong 1 giờ:

$$HM = \frac{2}{\frac{1}{60} + \frac{1}{40}} = \frac{2}{\frac{2}{120} + \frac{3}{120}} = \frac{2}{\frac{5}{120}} = \frac{240}{5} = 48 \text{ km/h} \tag{2..8}$$

b. Trung vị (Median)

Định nghĩa

Trung vị (Median) là giá trị nằm ở giữa một tập hợp dữ liệu đã được sắp xếp theo thứ tự tăng dần hoặc giảm dần. Nó chia tập dữ liệu thành hai phần bằng nhau: 50% giá trị nhỏ hơn trung vị và 50% giá trị lớn hơn trung vị.

-Ưu điểm:

- Ít bị ảnh hưởng bởi giá trị ngoại lai (outliers).
- Phù hợp khi dữ liệu có phân phối lệch.

Cách tính trung vị

*** Dữ liệu rời rạc**

- Nếu số lượng quan sát N là số lẻ:

$$\tilde{x} = X_{\frac{N+1}{2}}$$

- Nếu số lượng quan sát N là số chẵn:

$$\tilde{x} = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2}$$

*** Dữ liệu nhóm (có bảng tần số)**

Trung vị được tính theo công thức:

$$\tilde{x} = L + \frac{\frac{N}{2} - F}{f} \times h$$

Trong đó:

- L: Cận dưới của lớp chứa trung vị.
- N: Tổng số quan sát.
- F: Tần số tích lũy trước lớp chứa trung vị.
- f: Tần số của lớp chứa trung vị.
- h: Độ rộng lớp chứa trung vị.

*** Dữ liệu phân phối liên tục (sử dụng CDF)**

Trung vị là giá trị x sao cho:

$$F(\tilde{x}) = 0.5$$

Tức là điểm mà 50% dữ liệu nằm dưới nó trong hàm phân phối tích lũy.

Khi nào nên dùng trung vị thay vì trung bình?

- Khi dữ liệu có ngoại lai, trung vị ít bị ảnh hưởng hơn.
- Khi dữ liệu có phân phối lệch, trung vị thể hiện xu hướng trung tâm tốt hơn.
- Khi dữ liệu có dạng phân phối log-normal, chẳng hạn như bất động sản, thu nhập, giá cổ phiếu, v.v.

c. Mode

Định nghĩa: Mode (Yếu vị) là giá trị xuất hiện nhiều nhất trong một tập dữ liệu. Đây là một trong ba thước đo xu hướng trung tâm chính, bên cạnh Mean (trung bình) và Median (trung vị).

- Nếu một tập dữ liệu có một giá trị xuất hiện nhiều nhất, nó được gọi là **unimodal** (đơn mode).
- Nếu có hai giá trị cùng xuất hiện với tần suất cao nhất, tập dữ liệu được gọi là **bimodal** (hai mode).
- Nếu có nhiều hơn hai giá trị xuất hiện với tần suất cao nhất, tập dữ liệu được gọi là **multimodal** (đa mode).

*** Công thức xác định mode:** Mode không có công thức cố định như mean hay median. Nó đơn giản là giá trị có tần suất xuất hiện cao nhất trong tập dữ liệu.

Ví dụ:

- Dữ liệu: {2, 3, 5, 3, 3, 6, 7, 2, 2, 3}
- Mode = **3** (vì số 3 xuất hiện 4 lần, nhiều nhất trong tập dữ liệu).

*** Cách xác định mode trong phân bố tần suất** Với dữ liệu nhóm trong bảng tần suất, mode có thể được ước lượng bằng công thức:

$$\text{Mode} = L + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times h \tag{2..9}$$

Trong đó:

- L là cận dưới của lớp có tần suất cao nhất (lớp modal),
- f_1 là tần suất của lớp modal,
- f_0 là tần suất của lớp trước lớp modal,
- f_2 là tần suất của lớp sau lớp modal,
- h là độ rộng của lớp.

Ví dụ: Nếu có bảng tần suất như sau:

Khoảng lớp	Tần suất
10 - 20	5
20 - 30	8
30 - 40	12
40 - 50	9
50 - 60	6

- Lớp có tần suất cao nhất là **30 - 40** với $f_1 = 12$, - $L = 30$, $f_0 = 8$, $f_2 = 9$, - $h = 10$.

Áp dụng công thức:

$$\text{Mode} = 30 + \frac{(12-8)}{(2\times 12 - 8 - 9)} \times 10 = 30 + \frac{4}{7} \times 10 = 30 + 5.71 = 35.71 \quad (2..10)$$

*** Cách xác định Mode bằng đạo hàm**

Mode của một phân bố liên tục là giá trị x sao cho hàm mật độ xác suất $f(x)$ đạt cực đại, tức là:

**** Bước 1: Lấy đạo hàm bậc nhất**

Lấy đạo hàm bậc nhất của $f(x)$ và giải phương trình:

$$f'(x) = 0 \quad (2..11)$$

Đây là điều kiện cần để tìm điểm cực trị.

**** Bước 2: Kiểm tra đạo hàm bậc hai**

- Nếu $f''(x) < 0$, thì x là điểm cực đại và là Mode.
- Nếu $f''(x) > 0$, thì x là điểm cực tiểu (không phải Mode).

=> Ví dụ: Mode của phân bố chuẩn

Xét phân bố chuẩn $N(\mu, \sigma^2)$ có hàm mật độ xác suất:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2..12)$$

- Bước 1: Lấy đạo hàm

$$f'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \left(-\frac{2(x-\mu)}{2\sigma^2}\right) \tag{2..13}$$

$$= f(x) \cdot \left(-\frac{x-\mu}{\sigma^2}\right) \tag{2..14}$$

Đặt $f'(x) = 0$, ta có $x - \mu = 0$ hay Mode $= \mu$.

- Bước 2: Kiểm tra đạo hàm bậc hai

$$f''(x) = f(x) \cdot \left(-\frac{1}{\sigma^2}\right) + f(x) \cdot \left(-\frac{x-\mu}{\sigma^2}\right)^2 \tag{2..15}$$

Tại $x = \mu$, ta có $f''(x) < 0$, suy ra đây là điểm cực đại.

**** Kết luận:**

Mode của phân bố chuẩn chính là trung bình μ.

Đặc điểm của mode:

- Mode có thể không tồn tại hoặc có nhiều hơn một giá trị trong tập dữ liệu.
- Mode có thể bị ảnh hưởng bởi sự thay đổi nhỏ trong tần suất của dữ liệu.
- Đối với dữ liệu định tính (categorical data), mode là thước đo trung tâm phù hợp nhất.

Ví dụ:

- Dữ liệu màu sắc yêu thích của 100 người: {Đỏ, Xanh, Xanh, Xanh, Đỏ, Đỏ, Đỏ, Xanh, Xanh, Đỏ, Đỏ, Xanh}
- Mode = **"Xanh"** (vì xuất hiện nhiều nhất).

Thuộc tính	**Mode**	**Mean (Trung bình)**	**Median (Trung vị)**
Định nghĩa	Giá trị xuất hiện nhiều nhất	Trung bình số học của tất cả giá trị	Giá trị chính giữa của tập dữ liệu
Khi nào dùng?	Khi dữ liệu có giá trị lặp lại hoặc là dữ liệu định tính	Khi dữ liệu phân bố đều, không bị lệch	Khi dữ liệu có giá trị ngoại lai
Bị ảnh hưởng bởi ngoại lai?	Không	Có	Ít bị ảnh hưởng

Bảng 2..1: So sánh Mode với Mean và Median

Ứng Dụng của Mode

- **Thống kê kinh doanh**: Xác định sản phẩm bán chạy nhất.
- **Giáo dục**: Xác định điểm số phổ biến nhất trong lớp học.
- **Tiếp thị**: Tìm màu sắc, kích cỡ hoặc mẫu mã sản phẩm được ưa chuộng nhất.

Mode là một thước đo quan trọng trong thống kê, giúp hiểu rõ hơn về xu hướng trung tâm của dữ liệu. Trong nhiều trường hợp, nó là công cụ hữu ích hơn mean và median, đặc biệt đối với dữ liệu phân loại hoặc dữ liệu có phân phối không chuẩn.

2.2.2. Đo lường mức độ phân tán

a. Tứ phân vị (Quartiles)

Tứ phân vị là các giá trị chia một tập dữ liệu đã được sắp xếp thành bốn phần bằng nhau. Các giá trị này giúp chúng ta hiểu rõ hơn về sự phân bố của dữ liệu.

*** Các loại tứ phân vị**

- **Tứ phân vị thứ nhất** (Q_1): Là phần tử nằm ở vị trí 25% của dữ liệu đã sắp xếp. Đây là trung vị của nửa dưới của dữ liệu.
- **Tứ phân vị thứ hai** (Q_2): Là trung vị (median) của toàn bộ dữ liệu, chia dữ liệu thành hai phần bằng nhau (50%).
- **Tứ phân vị thứ ba** (Q_3): Là phần tử nằm ở vị trí 75% của dữ liệu. Đây là trung vị của nửa trên của dữ liệu.

*** Cách tính tứ phân vị**

1. Sắp xếp dữ liệu theo thứ tự tăng dần.
2. Tìm Q_2 (trung vị của toàn bộ dữ liệu).
3. Tìm Q_1 (trung vị của nửa dưới) và Q_3 (trung vị của nửa trên).

Ví dụ

Xét dãy số:

$$2, 4, 7, 10, 12, 15, 18, 22, 25, 30$$

- Q_2 **(Median)**: Trung vị là giá trị nằm giữa dãy số. Ở đây có 10 số, trung vị là:

$$Q_2 = \frac{12 + 15}{2} = 13.5$$

- Q_1 **(Tứ phân vị thứ nhất)**: Trung vị của nửa dưới:

$$Q_1 = \frac{4+7}{2} = 5.5$$

- Q_3 **(Tứ phân vị thứ ba)**: Trung vị của nửa trên:

$$Q_3 = \frac{22+25}{2} = 23.5$$

*** Ý nghĩa của tứ phân vị**

- Giúp xác định vị trí trung tâm và mức độ phân tán của dữ liệu.
- Dùng để tính khoảng biến thiên liên tứ phân vị (IQR) nhằm đo lường độ phân tán.

b. Khoảng biến thiên liên tứ phân vị (IQR - Interquartile Range)

Khoảng biến thiên liên tứ phân vị (IQR) đo độ phân tán của 50% dữ liệu trung tâm bằng cách tính hiệu giữa tứ phân vị thứ ba và tứ phân vị thứ nhất.

Công thức tính IQR

$$IQR = Q_3 - Q_1 \tag{2..16}$$

Trong đó:

- Q_1 là tứ phân vị thứ nhất (25%).
- Q_3 là tứ phân vị thứ ba (75%).

Ví dụ

Từ ví dụ trước với:

$$Q_1 = 5.5,$$
$$Q_3 = 23.5$$

Ta có:

$$IQR = 23.5 - 5.5 = 18 \tag{2..17}$$

→ **Khoảng 50% dữ liệu trung tâm nằm trong khoảng từ 5.5 đến 23.5.**

*** Ý nghĩa của IQR**

- ✓ Không bị ảnh hưởng bởi ngoại lệ, vì chỉ xét khoảng giữa 50% dữ liệu.
- ✓ Giúp phát hiện giá trị ngoại lệ, dựa vào ngưỡng ngoài:

Giới hạn dưới và giới hạn trên

$$\text{Giới hạn dưới} = Q_1 - 1.5 \times IQR \tag{2..18}$$

$$\text{Giới hạn trên} = Q_3 + 1.5 \times IQR \tag{2..19}$$

Nếu một điểm dữ liệu nằm ngoài khoảng này, nó có thể là ngoại lệ.

*** Ví dụ về phát hiện ngoại lệ**

Với $Q_1 = 5.5$, $Q_3 = 23.5$, và $IQR = 18$:

$$\text{Giới hạn dưới} = 5.5 - 1.5 \times 18 = -21.5$$

$$\text{Giới hạn trên} = 23.5 + 1.5 \times 18 = 50.5$$

→ Nếu một giá trị nhỏ hơn -21.5 hoặc lớn hơn 50.5, nó có thể là ngoại lệ.

c. Phương sai

*** Định nghĩa**

Phương sai thể hiện mức độ chênh lệch của các giá trị dữ liệu so với giá trị trung bình. Nếu phương sai lớn, dữ liệu có mức độ phân tán cao; ngược lại, nếu phương sai nhỏ, các giá trị dữ liệu tập trung quanh giá trị trung bình.

Phương sai thường được ký hiệu là:

- σ^2 (sigma bình phương) cho tổng thể.
- s^2 cho mẫu thống kê.

*** Công thức tính phương sai**

**** Phương sai của tổng thể**

Khi có toàn bộ dữ liệu trong tổng thể, phương sai được tính theo công thức:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{2..20}$$

Trong đó:

- σ^2 là phương sai của tổng thể.
- N là số lượng phần tử trong tổng thể.
- x_i là từng giá trị dữ liệu.
- μ là giá trị trung bình của tổng thể:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2..21}$$

**** Phương sai của mẫu**

Khi chỉ có một mẫu từ tổng thể, phương sai được ước lượng bằng công thức:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2..22}$$

Trong đó:

- s^2 là phương sai của mẫu.
- n là số lượng phần tử trong mẫu.
- x_i là từng giá trị trong mẫu.
- $\bar{x}$ là trung bình của mẫu:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2..23}$$

Lưu ý rằng trong công thức phương sai mẫu, mẫu số là $n-1$ thay vì n để bù trừ độ chệch khi ước lượng phương sai của tổng thể từ mẫu nhỏ.

*** Ý nghĩa của phương sai**

- **Đo lường mức độ phân tán**: Nếu phương sai lớn, dữ liệu phân tán rộng; nếu phương sai nhỏ, dữ liệu tập trung gần giá trị trung bình.
- **Quan trọng trong thống kê và học máy**: Phương sai được sử dụng rộng rãi trong kiểm định giả thuyết, hồi quy tuyến tính, và các thuật toán học máy để đánh giá mức độ biến động của dữ liệu.
- **So sánh độ biến động giữa các tập dữ liệu**: Ví dụ, phương sai giá cổ phiếu cao cho thấy biến động lớn, trong khi phương sai nhiệt độ môi trường thấp cho thấy nhiệt độ ổn định.

d. Độ lệch chuẩn (Standard Deviation)

*** Định Nghĩa Độ lệch chuẩn**

Độ lệch chuẩn là một thước đo phản ánh mức độ phân tán của tập dữ liệu so với giá trị trung bình. Nếu độ lệch chuẩn lớn, dữ liệu có xu hướng phân tán rộng; nếu nhỏ, dữ liệu tập trung quanh giá trị trung bình.

*** Công Thức Tính Độ lệch chuẩn**

**** Độ lệch chuẩn của Tổng thể**

Công thức tính độ lệch chuẩn của tổng thể:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2..24}$$

Trong đó:

- σ là độ lệch chuẩn của tổng thể.
- N là số phần tử trong tổng thể.
- x_i là từng giá trị dữ liệu.
- μ là giá trị trung bình của tổng thể:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2..25}$$

**** Độ lệch chuẩn của Mẫu**

Khi chỉ có một mẫu từ tổng thể, công thức tính độ lệch chuẩn mẫu là:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2..26}$$

Trong đó:

- s là độ lệch chuẩn của mẫu.
- n là số phần tử trong mẫu.
- x_i là từng giá trị trong mẫu.
- $\bar{x}$ là trung bình của mẫu:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2..27}$$

Lưu ý rằng trong công thức độ lệch chuẩn mẫu, mẫu số là $n-1$ thay vì n để bù trừ độ chệch khi ước lượng độ lệch chuẩn của tổng thể từ mẫu nhỏ.

*** Ý Nghĩa của Độ lệch chuẩn**

- **Đo lường mức độ phân tán:** Nếu độ lệch chuẩn lớn, dữ liệu phân tán rộng; nếu nhỏ, dữ liệu tập trung gần giá trị trung bình.
- **Quan trọng trong thống kê và học máy:** Độ lệch chuẩn được sử dụng trong kiểm định giả thuyết, hồi quy tuyến tính, và các thuật toán học máy.
- **So sánh độ biến động giữa các tập dữ liệu:** Ví dụ, độ lệch chuẩn giá cổ phiếu cao cho thấy biến động lớn, trong khi độ lệch chuẩn nhiệt độ môi trường thấp cho thấy nhiệt độ ổn định.

2.2.3. Đo lường hình dạng phân phối dữ liệu

Hình dạng phân phối mô tả cách dữ liệu được sắp xếp xung quanh giá trị trung tâm.

a. Độ lệch (Skewness)

Định nghĩa: Độ lệch đo lường mức độ đối xứng của phân phối dữ liệu.

Công thức:

$$\text{Skewness} = \frac{\sum(x_i - \bar{x})^3}{(n-1)s^3} \tag{2..28}$$

Diễn giải:

- Skewness = 0: Phân phối đối xứng.
- Skewness > 0: Phân phối lệch phải (đuôi dài bên phải).
- Skewness < 0: Phân phối lệch trái (đuôi dài bên trái).

Ví dụ: Thu nhập của dân số thường có phân phối lệch phải vì có ít người có thu nhập rất cao.

b. Độ nhọn (Kurtosis)

Định nghĩa: Độ nhọn đo mức độ "tập trung"của dữ liệu quanh trung bình so với phân phối chuẩn.

Công thức:

$$Kurtosis = \frac{\sum(x_i - \bar{x})^4}{(n-1)s^4} \tag{2..29}$$

Diễn giải:

- $Kurtosis = 3$: Phân phối chuẩn (mesokurtic).
- $Kurtosis > 3$: Phân phối có đỉnh nhọn (leptokurtic), có nhiều ngoại lai.
- $Kurtosis < 3$: Phân phối có đỉnh thấp, dẹt hơn (platykurtic).

Ví dụ: Giá cổ phiếu có thể có kurtosis cao vì có nhiều biến động lớn bất thường.

2.2.4. Đo lường mối quan hệ giữa các biến

Các thước đo này giúp đánh giá mối quan hệ giữa hai biến số

a. Hiệp phương sai (Covariance)

Định nghĩa: Hiệp phương sai đo mức độ thay đổi cùng nhau của hai biến số.

Công thức:

$$Cov(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{2..30}$$

Diễn giải:

- Nếu $Cov(X,Y) > 0$, hai biến có xu hướng tăng hoặc giảm cùng nhau.

- Nếu $Cov(X, Y) < 0$, một biến tăng thì biến kia giảm.
- Nếu $Cov(X, Y) = 0$, hai biến không liên hệ tuyến tính với nhau.

Ví dụ: Hiệp phương sai giữa thu nhập và chi tiêu của một hộ gia đình thường là dương.

b. Hệ số tương quan Pearson (Pearson Correlation)

Định nghĩa: Đo lường mức độ tuyến tính của mối quan hệ giữa hai biến.

Công thức:

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y} \tag{2..31}$$

Diễn giải:

- $r = 1$: Mối quan hệ tuyến tính hoàn hảo dương.
- $r = -1$: Mối quan hệ tuyến tính hoàn hảo âm.
- $r = 0$: Không có tương quan tuyến tính.

Ví dụ: Tương quan giữa số giờ học và điểm thi thường dương, nhưng không phải lúc nào cũng là 1.

c. Hệ số tương quan Spearman (Spearman Correlation)

* Giới thiệu

Hệ số tương quan Spearman (**Spearman's rank correlation coefficient**), ký hiệu là ρ hoặc r_s, đo mức độ tương quan giữa hai tập hợp dữ liệu dựa trên **thứ hạng** thay vì giá trị thực tế. Nó được sử dụng khi dữ liệu không tuân theo phân phối chuẩn hoặc khi mối quan hệ giữa hai biến không hoàn toàn tuyến tính.

* Công thức tính

Hệ số Spearman được tính theo công thức:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2..32}$$

Trong đó:

- d_i là hiệu giữa thứ hạng của từng cặp dữ liệu: $d_i = \text{rank}(x_i) - \text{rank}(y_i)$.
- n là số lượng quan sát (cặp dữ liệu).

X	Y	Rank(X)	Rank(Y)	$d_i = \text{Rank}(X) - \text{Rank}(Y)$
10	200	1	2	-1
20	180	2	1	1
30	220	3	4	-1
40	240	4	5	-1
50	210	5	3	2

Bảng 2..2: Ví dụ về tính hệ số Spearman

**** Bảng tính toán ví dụ**

Tính tổng bình phương sai lệch:

$$\sum d_i^2 = 1^2 + (-1)^2 + (-1)^2 + (-1)^2 + 2^2 = 1 + 1 + 1 + 1 + 4 = 8 \quad (2..33)$$

Thay vào công thức tính Spearman:

$$r_s = 1 - \frac{6(8)}{5(25-1)} = 1 - \frac{48}{120} = 1 - 0.4 = 0.6 \quad (2..34)$$

Kết quả $r_s = 0.6$ cho thấy mối quan hệ tương quan dương vừa phải giữa X và Y.

*** So sánh với Pearson**

Tiêu chí	Pearson (r)	Spearman (r_s)
Dữ liệu yêu cầu	Phân phối chuẩn	Không yêu cầu
Tính toán dựa trên	Giá trị thực	Thứ hạng
Đo lường quan hệ	Tuyến tính	Phi tuyến đơn điệu
Nhạy cảm với ngoại lệ	Có	Ít hơn

Bảng 2..3: So sánh hệ số Pearson và Spearman

*** Khi nào nên sử dụng Spearman?**

- Khi dữ liệu không tuân theo phân phối chuẩn.
- Khi dữ liệu có quan hệ phi tuyến nhưng đơn điệu (tăng hoặc giảm liên tục).
- Khi có nhiều ngoại lệ ảnh hưởng đến phân phối của dữ liệu.
- Khi làm việc với dữ liệu xếp hạng (ordinal data).

Hệ số tương quan Spearman là một công cụ hữu ích để đo lường mối quan hệ giữa hai biến trong trường hợp dữ liệu không tuyến tính hoặc không có phân phối chuẩn. Nó có tính ứng dụng cao trong phân tích dữ liệu xã hội, tài chính, và khoa học tự nhiên.

2.3. Xử lý dữ liệu trong kinh tế lượng

2.3.1. Định nghĩa bài toán

2.3.2. Thu thập dữ liệu

2.3.3. Xử lý dữ liệu

2.3.4. Kết luận

Chương 3.

Luật phân bố xác suất

3.1. Giới thiệu

Xác suất là một công cụ quan trọng trong toán học và thống kê để mô tả sự không chắc chắn của các hiện tượng ngẫu nhiên. Trong chương này, chúng ta sẽ trình bày các luật phân bố xác suất, bao gồm phân bố rời rạc và liên tục, cùng với các định lý quan trọng.

3.2. Các Định Nghĩa Cơ Bản

3.2.1. Biến ngẫu nhiên

Định nghĩa: Một biến ngẫu nhiên là một hàm số ánh xạ từ không gian mẫu (tập hợp tất cả các kết quả có thể của một thí nghiệm ngẫu nhiên) vào tập số thực $\mathbb{R}$. Nói cách khác, biến ngẫu nhiên là một đại lượng số học có thể nhận các giá trị khác nhau do yếu tố ngẫu nhiên.

Ví dụ minh họa: Giả sử tung một con xúc xắc.

- Không gian mẫu: $S = \{1, 2, 3, 4, 5, 6\}$.
- Định nghĩa biến ngẫu nhiên X là "số chấm xuất hiện trên mặt ngửa của xúc xắc".
- Khi đó, X có thể nhận các giá trị $1, 2, 3, 4, 5, 6$, mỗi giá trị này tương ứng với một khả năng xảy ra.

3.2.2. Hàm phân bố xác suất (CDF - Cumulative Distribution Function)

Định nghĩa: Hàm phân bố xác suất của một biến ngẫu nhiên X được định nghĩa là:

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

Hàm phân bố xác suất giúp mô tả cách xác suất được phân bố trên tập giá trị của biến ngẫu nhiên.

Ví dụ minh họa (Biến ngẫu nhiên rời rạc): Xét biến ngẫu nhiên X là số chấm trên mặt ngửa của một xúc xắc 6 mặt cân bằng. Khi đó, ta có:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{6}, & 1 \le x < 2 \\ \frac{2}{6}, & 2 \le x < 3 \\ \frac{3}{6}, & 3 \le x < 4 \\ \frac{4}{6}, & 4 \le x < 5 \\ \frac{5}{6}, & 5 \le x < 6 \\ 1, & x \ge 6 \end{cases}$$

3.2.3. Hàm mật độ xác suất (PDF - Probability Density Function)

Định nghĩa: Nếu X là một biến ngẫu nhiên liên tục, thì xác suất để X nằm trong khoảng $[a, b]$ được xác định bằng tích phân:

$$P(a \le X \le b) = \int_a^b f_X(x)\, dx.$$

Ví dụ minh họa: Xét biến ngẫu nhiên X có phân bố chuẩn (Gaussian) với kỳ vọng $\mu = 0$ và phương sai $\sigma^2 = 1$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Xác suất X nằm trong khoảng từ -1 đến 1:

$$P(-1 \le X \le 1) = \int_{-1}^{1} f_X(x)\, dx \approx 0.6826.$$

3.2.4. Hàm khối xác suất (PMF - Probability Mass Function)

Định nghĩa: Nếu X là một biến ngẫu nhiên rời rạc, thì xác suất để X nhận giá trị x_i được xác định bằng hàm khối xác suất:

$$P(X = x_i) = p_X(x_i), \quad \sum_i p_X(x_i) = 1.$$

Ví dụ minh họa: Xét biến ngẫu nhiên X biểu diễn số lần xuất hiện mặt ngửa khi tung 2 đồng xu cân bằng. Khi đó, X có thể nhận các giá trị 0, 1, hoặc 2 với xác suất:

$$p_X(0) = P(X = 0) = \frac{1}{4}, \quad p_X(1) = P(X = 1) = \frac{2}{4}, \quad p_X(2) = P(X = 2) = \frac{1}{4}.$$

Tổng tất cả các xác suất:

$$\sum_i p_X(x_i) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1.$$

Điều này xác nhận rằng tổng xác suất của tất cả giá trị có thể xảy ra bằng 1.

3.3. Luật Số Lớn

Luật số lớn (LLN) là một định lý cơ bản trong xác suất thống kê, mô tả xu hướng hội tụ của trung bình mẫu về giá trị kỳ vọng khi kích thước mẫu tăng. LLN đóng vai trò quan trọng trong thống kê và nhiều ứng dụng thực tế như kinh tế, khoa học dữ liệu.

Có hai dạng của định lý Luật số lớn:

3.3.1. Luật số lớn yếu (Weak Law of Large Numbers - WLLN)

Giả sử $X_1, X_2, \ldots, X_n$ là một dãy các biến ngẫu nhiên độc lập và cùng phân phối (i.i.d) với kỳ vọng hữu hạn $E[X_i] = \mu$. Khi đó, với mọi $\varepsilon > 0$, ta có:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) = 0 \tag{3..1}$$

Điều này có nghĩa là khi kích thước mẫu n đủ lớn, xác suất để trung bình mẫu khác xa kỳ vọng thực tế sẽ tiến về 0.

Giải thích các ký hiệu:

- $X_1, X_2, \ldots, X_n$: Các biến ngẫu nhiên độc lập, cùng phân phối.
- $E[X_i]$: Kỳ vọng của biến ngẫu nhiên X_i, ký hiệu là μ.
- $\frac{1}{n}\sum_{i=1}^{n} X_i$: Trung bình mẫu.
- $P(A)$: Xác suất xảy ra của biến cố A.
- $\lim_{n\to\infty}$: Giới hạn khi kích thước mẫu tiến đến vô cùng.
- ε: Một số dương nhỏ tùy ý.

3.3.2. Luật số lớn mạnh (Strong Law of Large Numbers - SLLN)

Với cùng điều kiện như trên, Luật số lớn mạnh phát biểu rằng:

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1 \tag{3..2}$$

Tức là trung bình mẫu sẽ hội tụ chắc chắn (almost surely) về kỳ vọng μ khi $n \to \infty$.

Giải thích các ký hiệu:

- Các ký hiệu tương tự như Luật số lớn yếu.
- $P(A) = 1$: Sự kiện A xảy ra với xác suất chắc chắn.
- $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$: Trung bình mẫu hội tụ về kỳ vọng μ khi $n \to \infty$.

3.3.3. Ví dụ minh họa

Giả sử ta có một đồng xu không cân bằng với xác suất xuất hiện mặt ngửa là $p = 0.6$. Gieo đồng xu n lần và tính xác suất trung bình của số lần xuất hiện mặt ngửa:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3..3}$$

Theo Luật số lớn, khi n tăng, $\bar{X}_n$ sẽ hội tụ về $p = 0.6$.

Giải thích các ký hiệu:

- X_i: Biến ngẫu nhiên nhận giá trị 1 nếu lần gieo thứ i ra mặt ngửa, và 0 nếu ra mặt sấp.
- $\bar{X}_n$: Trung bình của các lần thử, tức là tỷ lệ số lần xuất hiện mặt ngửa trong n lần thử.
- Khi n càng lớn, $\bar{X}_n$ sẽ tiến gần về giá trị kỳ vọng $p = 0.6$, theo Luật số lớn.

Luật số lớn cho thấy khi thu thập nhiều dữ liệu hơn, giá trị trung bình của mẫu sẽ gần hơn với giá trị kỳ vọng thực tế. Đây là cơ sở lý thuyết quan trọng trong thống kê, tài chính, trí tuệ nhân tạo và nhiều lĩnh vực khác.

3.4. Các luật phân bố xác suất quan trọng

3.4.1. Phân bố nhị thức

Định nghĩa Phân bố nhị thức mô tả số lần xảy ra của một sự kiện trong một số lần thử độc lập, khi mỗi lần thử chỉ có hai kết quả: **thành công** hoặc **thất bại**.

Một biến ngẫu nhiên X tuân theo phân bố nhị thức với các tham số n (số lần thử) và p (xác suất thành công trong mỗi lần thử) nếu xác suất để X nhận giá trị k (tức là có đúng k lần thành công trong n phép thử) được tính theo công thức:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n. \tag{3..4}$$

Trong đó:

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ là hệ số nhị thức (binomial coefficient).
- p^k là xác suất có đúng k lần thành công.
- $(1-p)^{n-k}$ là xác suất có $(n-k)$ lần thất bại.

Kỳ vọng và Phương sai

$$E(X) = np, \quad \text{Var}(X) = np(1-p). \tag{3..5}$$

Ví Dụ Giả sử một bài kiểm tra trắc nghiệm có 10 câu hỏi, mỗi câu có 4 đáp án nhưng chỉ có 1 đáp án đúng. Một học sinh chọn đáp án ngẫu nhiên cho mỗi câu. Gọi X là số câu trả lời đúng, thì X tuân theo phân bố nhị thức $B(10, 0.25)$ vì xác suất chọn đúng một đáp án là $p = 0.25$.

Ứng dụng thực tế Phân bố nhị thức có nhiều ứng dụng trong thực tế, bao gồm:

- Xác suất một sản phẩm bị lỗi khi lấy mẫu kiểm tra trong dây chuyền sản xuất.
- Dự đoán số lượng khách hàng tiềm năng sẽ mua sản phẩm sau khi quảng cáo.
- Xác suất thắng một trò chơi nếu người chơi có một tỷ lệ chiến thắng cố định.

3.4.2. Phân Bố Poisson

Định nghĩa : Phân bố Poisson là một phân bố xác suất rời rạc mô tả số lần xảy ra của một sự kiện trong một khoảng thời gian (hoặc không gian) nhất định khi các sự kiện đó xảy ra độc lập với nhau và có tỷ lệ trung bình không đổi.

Một biến ngẫu nhiên X tuân theo phân bố Poisson với tham số $\lambda > 0$ nếu nó có xác suất:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{3..6}$$

trong đó:

- λ là số lần xảy ra trung bình của sự kiện trong khoảng thời gian hoặc không gian xác định.
- $k!$ là giai thừa của k với quy ước $0! = 1$.
- $e \approx 2.718$ là hằng số Euler.

Ý nghĩa và ứng dụng Phân bố Poisson được sử dụng để mô tả số lần xảy ra của các sự kiện hiếm gặp trong một khoảng thời gian hoặc không gian cố định, chẳng hạn như:

- Số cuộc gọi đến tổng đài trong một giờ.
- Số lỗi xảy ra trong một hệ thống máy tính trong một ngày.
- Số tai nạn giao thông trên một đoạn đường trong một tuần.

- Số khách hàng đến một cửa hàng trong một khoảng thời gian nhất định.

Các đặc trưng của phân bố Poisson

- **Kỳ vọng (trung bình)**: $E(X) = \lambda$
- **Phương sai**: $\text{Var}(X) = \lambda$
- **Độ lệch chuẩn**: $\sigma = \sqrt{\lambda}$

Một số tính chất quan trọng:

- Phân bố Poisson có thể được sử dụng để xấp xỉ phân bố nhị thức $B(n, p)$ khi n lớn và p nhỏ sao cho $\lambda = np$.
- Nếu $X_1 \sim \text{Poisson}(\lambda_1)$ và $X_2 \sim \text{Poisson}(\lambda_2)$ độc lập, thì tổng của chúng cũng tuân theo phân bố Poisson:

$$X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2). \tag{3..7}$$

Ví dụ minh họa

- **Ví dụ 1: Số cuộc gọi đến tổng đài** Giả sử một tổng đài nhận trung bình 4 cuộc gọi mỗi phút. Hỏi xác suất để trong một phút có đúng 2 cuộc gọi đến là bao nhiêu?

  Áp dụng công thức phân bố Poisson với $\lambda = 4$, $k = 2$:

$$P(X = 2) = \frac{4^2 e^{-4}}{2!} = \frac{16e^{-4}}{2} \approx 0.1465. \tag{3..8}$$

  Vậy xác suất nhận đúng 2 cuộc gọi trong một phút là khoảng **14.65%**.

- **Ví dụ 2: Số lỗi phần mềm** Một phần mềm có trung bình 3 lỗi xảy ra mỗi ngày. Xác suất để hôm nay có **không có lỗi nào** là bao nhiêu?

  Dùng công thức với $\lambda = 3$, $k = 0$:

$$P(X = 0) = \frac{3^0 e^{-3}}{0!} = e^{-3} \approx 0.0498. \tag{3..9}$$

  Vậy xác suất không có lỗi nào trong ngày hôm nay là **4.98%**.

Mối liên hệ với các phân bố khác

- Khi $n \to \infty$, $p \to 0$ nhưng $np = \lambda$ cố định, phân bố nhị thức $B(n, p)$ xấp xỉ phân bố Poisson với tham số λ.
- Khi λ lớn, phân bố Poisson có thể được xấp xỉ bằng phân bố chuẩn:

$$X \approx N(\lambda, \lambda). \tag{3..10}$$

3.4.3. Phân Bố Chuẩn (Gauss)

Phân bố chuẩn, còn gọi là phân bố Gauss, là một trong những phân bố quan trọng nhất trong thống kê và xác suất. Nó được sử dụng rộng rãi trong nhiều lĩnh vực như tài chính, khoa học dữ liệu, kỹ thuật và kinh tế.

Định nghĩa Phân bố chuẩn có dạng hàm mật độ xác suất (PDF) như sau:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3..11}$$

trong đó:

- μ là kỳ vọng (trung bình) của phân bố.
- σ là độ lệch chuẩn.
- σ^2 là phương sai.
- x là biến ngẫu nhiên tuân theo phân bố chuẩn.

Đặc điểm của Phân Bố Chuẩn

Phân bố chuẩn có một số đặc điểm quan trọng:

1. Đối xứng quanh giá trị trung bình μ.
2. Đường cong hình chuông với đỉnh tại $x = \mu$.
3. Tổng diện tích dưới đường cong bằng 1.
4. Khoảng $\mu \pm \sigma$ chứa khoảng 68.27% dữ liệu.
5. Khoảng $\mu \pm 2\sigma$ chứa khoảng 95.45% dữ liệu.
6. Khoảng $\mu \pm 3\sigma$ chứa khoảng 99.73% dữ liệu.

Phân bố chuẩn tắc

Phân bố chuẩn tắc (standard normal distribution) là trường hợp đặc biệt của phân bố chuẩn với:

- $\mu = 0$
- $\sigma = 1$

Trong trường hợp này, công thức phân bố chuẩn trở thành:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{3..12}$$

Khi một biến ngẫu nhiên X tuân theo phân bố chuẩn với kỳ vọng μ và độ lệch chuẩn σ, ta có thể chuẩn hóa về phân bố chuẩn tắc bằng công thức:

$$Z = \frac{X - \mu}{\sigma} \tag{3..13}$$

Biến đổi này giúp ta dễ dàng tra cứu bảng phân bố chuẩn và tính toán xác suất.

Ứng dụng của Phân bố chuẩn

Phân bố chuẩn có rất nhiều ứng dụng trong thực tế:

- Kiểm định giả thuyết thống kê.
- Mô hình hóa dữ liệu thực tế trong nhiều lĩnh vực.
- Dùng trong kiểm soát chất lượng sản xuất.
- Ước lượng khoảng tin cậy trong thống kê.
- Dự báo và phân tích rủi ro trong tài chính.

3.5. Bậc tự do (Degrees of Freedom - DoF)

Trong thống kê, bậc tự do liên quan đến số lượng giá trị có thể thay đổi tự do trong một phép tính, thường xuất hiện trong kiểm định giả thuyết và phân bố xác suất.

3.5.1. Định nghĩa toán học của bậc tự do

Trong thống kê, **bậc tự do** của một phép tính là số lượng giá trị có thể thay đổi tự do mà không bị ràng buộc bởi các điều kiện hoặc mối quan hệ toán học khác.

Nếu có n quan sát nhưng một số quan sát bị ràng buộc bởi một hoặc nhiều điều kiện, thì bậc tự do là số lượng giá trị có thể thay đổi một cách độc lập.

Công thức tổng quát của bậc tự do trong thống kê:

$$df = n - k \tag{3..14}$$

trong đó:

- n là tổng số quan sát,
- k là số lượng tham số ước lượng từ dữ liệu.

Ví dụ: Nếu bạn có 5 số và biết trung bình của chúng, thì chỉ cần biết 4 số đầu tiên là có thể suy ra số thứ 5, nghĩa là chỉ có 4 bậc tự do.

3.5.2. Ý nghĩa trong ước lượng thống kê

Trong thống kê, khi tính toán các đặc trưng của mẫu (ví dụ: phương sai, độ lệch chuẩn), bậc tự do ảnh hưởng trực tiếp đến độ chính xác của ước lượng.

Phương sai mẫu s^2 Khi tính phương sai của mẫu, ta sử dụng công thức:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{3..15}$$

Ở đây, $n-1$ là số bậc tự do, vì ta đã sử dụng một quan sát để tính giá trị trung bình $\bar{x}$, làm giảm số lượng giá trị có thể thay đổi độc lập.

Nếu dùng n thay vì $n-1$, ước lượng phương sai sẽ bị lệch (underestimate).

Ứng dụng thực tế: Khi tính phương sai của một tập dữ liệu nhỏ, việc sử dụng bậc tự do $n-1$ giúp tạo ra một ước lượng không thiên lệch cho phương sai tổng thể.

3.5.3. Bậc tự do trong kiểm định giả thuyết

Bậc tự do rất quan trọng trong các kiểm định thống kê như kiểm định t-test, kiểm định χ^2, và ANOVA.

*** Kiểm định t-test**

Kiểm định t-test được sử dụng để so sánh trung bình của hai nhóm.

Công thức bậc tự do trong kiểm định t-test một mẫu:

$$df = n - 1 \tag{3..16}$$

Trong kiểm định t-test hai mẫu độc lập:

$$df = n_1 + n_2 - 2 \tag{3..17}$$

trong đó n_1, n_2 là kích thước mẫu của hai nhóm.

Ứng dụng thực tế:

- So sánh điểm thi giữa hai lớp học.
- Đánh giá hiệu quả của một loại thuốc giữa hai nhóm bệnh nhân.

*** Kiểm định χ^2 (Kiểm định phù hợp và kiểm định độc lập)**

Kiểm định χ^2 giúp xác định sự khác biệt giữa các nhóm danh mục (categorical data).

Công thức bậc tự do trong bảng tần suất:

$$df = (r-1) \times (c-1) \tag{3..18}$$

trong đó r là số hàng, c là số cột.

Ứng dụng thực tế:

- Kiểm tra xem giới tính có ảnh hưởng đến sở thích mua sắm hay không.
- Đánh giá mối quan hệ giữa thói quen ăn uống và tình trạng sức khỏe.

*** Phân tích phương sai (ANOVA)**

Trong ANOVA, bậc tự do giúp xác định nguồn biến thiên giữa các nhóm và bên trong nhóm.

Công thức:

$$df_{between} = k - 1 \tag{3..19}$$

$$df_{within} = N - k \tag{3..20}$$

trong đó k là số nhóm và N là tổng số quan sát.

Ứng dụng thực tế:

- So sánh hiệu suất của ba phương pháp giảng dạy khác nhau.
- Đánh giá hiệu quả của ba chiến lược tiếp thị.

3.5.4. Bậc tự do trong hồi quy tuyến tính

Bậc tự do cũng quan trọng trong hồi quy tuyến tính vì nó ảnh hưởng đến chất lượng mô hình dự báo.

Trong mô hình hồi quy tuyến tính có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{3..21}$$

Bậc tự do được tính là:

$$df = n - (k + 1) \tag{3..22}$$

trong đó:

- n là số quan sát.
- k là số biến độc lập.

Ứng dụng thực tế

- Dự đoán giá bất động sản dựa trên diện tích, số phòng ngủ, và vị trí.
- Phân tích các yếu tố ảnh hưởng đến doanh thu doanh nghiệp.

3.5.5. Tác động của bậc tự do đến phân phối xác suất

Bậc tự do cũng ảnh hưởng đến hình dạng của một số phân phối xác suất như phân phối t-Student, phân phối χ^2, và phân phối F.

- Khi bậc tự do tăng, phân phối t-Student dần tiến gần đến phân phối chuẩn.
- Trong phân phối χ^2, bậc tự do ảnh hưởng đến mức độ phân tán của phân phối.
- Trong kiểm định F, bậc tự do ảnh hưởng đến xác suất từ chối giả thuyết không.

Ứng dụng thực tế

- Khi kiểm tra giả thuyết với số lượng mẫu nhỏ, ta sử dụng phân phối t-Student thay vì phân phối chuẩn.
- Trong kiểm định phương sai, số bậc tự do quyết định xác suất sai lầm loại I.

Chương 4.

Các phương pháp phân tích dữ liệu bằng mô hình thống kê

Trong kinh tế lượng, hai phương pháp tiếp cận quan trọng cần đề cập là ước tính tham số và kiểm định giả thuyết, vì chúng là nền tảng để xây dựng và đánh giá các mô hình kinh tế lượng.

4.1. Phương pháp ước lượng tham số (Parameter Estimation)

4.1.1. Phương pháp bình quân nhỏ nhất (OLS - Ordinary Least Squares)

a. Giới thiệu về phương pháp OLS

Phương pháp bình phương nhỏ nhất (OLS) là một trong những phương pháp phổ biến nhất trong kinh tế lượng và thống kê để ước lượng các tham số của mô hình hồi quy tuyến tính. Mục tiêu của OLS là tìm ra các hệ số hồi quy sao cho tổng bình phương phần dư (sai số giữa giá trị thực tế và giá trị dự báo) là nhỏ nhất.

b. Mô hình hồi quy tuyến tính tổng quát

Giả sử mô hình hồi quy tuyến tính có dạng:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{4..1}$$

trong đó:

- Y_i là biến phụ thuộc (biến kết quả)
- X_{ij} là các biến độc lập (biến giải thích)
- $\beta_0, \beta_1, \ldots, \beta_k$ là các hệ số hồi quy cần ước lượng
- ε_i là sai số ngẫu nhiên

c. Nguyên lý của phương pháp OLS

Phương pháp OLS tìm kiếm các hệ số β bằng cách cực tiểu hóa tổng bình phương sai số:

$$S(\beta) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}))^2 \tag{4..2}$$

Để tìm các hệ số β, ta giải hệ phương trình bình thường (normal equations):

$$(X'X)\hat{\beta} = X'Y \tag{4..3}$$

trong đó:

- X là ma trận dữ liệu của các biến độc lập $(n \times k)$
- Y là vector của biến phụ thuộc $(n \times 1)$
- $\hat{\beta}$ là vector hệ số hồi quy $(k \times 1)$
- X' là ma trận chuyển vị của X

d. Các giả định của OLS

Phương pháp OLS hoạt động tốt khi các giả định sau được thỏa mãn:

1. **Tuyến tính**: Mô hình phải có dạng tuyến tính đối với các tham số.
2. **Kỳ vọng bằng 0 của sai số**: $E(\varepsilon_i) = 0$.
3. **Độc lập của sai số**: Sai số không có tương quan với nhau (không có tự tương quan).
4. **Phương sai đồng nhất**: Sai số có phương sai không đổi (không có hiện tượng phương sai thay đổi).
5. **Không có đa cộng tuyến hoàn hảo**: Các biến độc lập không được có tương quan tuyến tính hoàn hảo.
6. **Phân phối chuẩn của sai số (nếu mẫu nhỏ)**: Giả định này giúp kiểm định giả thuyết và tính khoảng tin cậy chính xác hơn.

e. Ước lượng và kiểm định ý nghĩa của hệ số hồi quy

Sau khi ước lượng các hệ số hồi quy bằng OLS, ta kiểm định ý nghĩa thống kê của chúng bằng kiểm định t (t-test). Giả thuyết kiểm định cho mỗi hệ số β_j:

- $H_0 : \beta_j = 0$ (hệ số không có ý nghĩa thống kê)
- $H_1 : \beta_j \neq 0$ (hệ số có ý nghĩa thống kê)

Chỉ số thống kê t được tính bằng:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{4..4}$$

trong đó $SE(\hat{\beta}_j)$ là sai số chuẩn của hệ số ước lượng β_j. Nếu giá trị p-value của kiểm định nhỏ hơn mức ý nghĩa α (thường là 0.05), ta bác bỏ H_0 và kết luận rằng hệ số có ý nghĩa thống kê.

f. Đánh giá chất lượng mô hình hồi quy

*** Hệ số xác định R^2**

Hệ số xác định R^2 đo lường mức độ giải thích của mô hình đối với biến phụ thuộc:

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST} \tag{4..5}$$

trong đó:

- SSR là tổng bình phương sai số (Sum of Squared Residuals)
- SST là tổng bình phương tổng thể (Total Sum of Squares)
- SSE là tổng bình phương hồi quy (Sum of Squares for Regression)

*** Kiểm định F**

Kiểm định F đánh giá xem mô hình hồi quy có phù hợp hay không:

$$F = \frac{(SST - SSR)/k}{SSR/(n-k-1)} \tag{4..6}$$

Nếu p-value của kiểm định F nhỏ hơn mức ý nghĩa α, mô hình được xem là có ý nghĩa tổng thể.

Phương pháp bình phương nhỏ nhất (OLS) là một kỹ thuật phổ biến để ước lượng mô hình hồi quy tuyến tính. Khi các giả định của OLS được thỏa mãn, phương pháp này giúp chúng ta có được các ước lượng không chệch, hiệu quả và tối ưu. Tuy nhiên, nếu các giả định bị vi phạm, có thể cần đến các phương pháp hồi quy khác như hồi quy tổng quát (GLS), hồi quy Ridge hoặc hồi quy LASSO để khắc phục.

4.1.2. Phương pháp hợp lý tối đa (MLE - Maximum Likelihood Estimation)

a. Giới thiệu về phương pháp MLE

Phương pháp hợp lý tối đa (MLE - Maximum Likelihood Estimation) là một phương pháp thống kê dùng để ước lượng các tham số của một mô hình xác

suất dựa trên dữ liệu quan sát. Nguyên tắc cơ bản của MLE là tìm giá trị của các tham số sao cho xác suất quan sát được tập dữ liệu hiện có là lớn nhất.

Nếu mô hình xác suất có phân phối xác suất $f(Y \mid \theta)$, trong đó:

- $Y = (Y_1, Y_2, ..., Y_n)$ là tập dữ liệu quan sát
- θ là vector tham số cần ước lượng

Thì MLE sẽ tìm giá trị $\hat{\theta}$ sao cho hàm hợp lý $L(\theta)$ đạt cực đại:

$$\hat{\theta} = \arg\max_{\theta} L(\theta) \tag{4..7}$$

trong đó $L(\theta)$ là hàm hợp lý của dữ liệu:

$$L(\theta) = P(Y \mid \theta) = \prod_{i=1}^{n} f(Y_i \mid \theta) \tag{4..8}$$

b. Hàm hợp lý và hàm log-hợp lý

Do tích của nhiều xác suất nhỏ có thể dẫn đến vấn đề số học, ta thường sử dụng hàm log-hợp lý:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(Y_i \mid \theta) \tag{4..9}$$

Bài toán tối đa hóa hàm hợp lý chuyển thành:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta) \tag{4..10}$$

c. Ví dụ: Ước lượng tham số trong phân phối chuẩn

Giả sử dữ liệu $Y_1, Y_2, ..., Y_n$ được lấy mẫu từ phân phối chuẩn:

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \tag{4..11}$$

Hàm mật độ xác suất của phân phối chuẩn là:

$$f(Y_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \mu)^2}{2\sigma^2}\right) \tag{4..12}$$

Hàm log-hợp lý:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y_i - \mu)^2}{2\sigma^2}\right] \tag{4..13}$$

Tính đạo hàm theo μ:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma^2} \tag{4..14}$$

Giải phương trình $\frac{\partial \ell}{\partial \mu} = 0$ ta được:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{4..15}$$

Tương tự, ước lượng của σ^2 là:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\mu})^2 \tag{4..16}$$

d. Các tính chất của ước lượng MLE

* Tính nhất quán

Ước lượng MLE là nhất quán, nghĩa là khi $n \to \infty$, giá trị ước lượng $\hat{\theta}$ hội tụ về giá trị thật θ_0.

* Tính không chệch và hiệu quả

Dưới các điều kiện thông thường, MLE gần như không chệch và đạt được giới hạn Cramér-Rao.

* Phân phối tiệm cận

Khi kích thước mẫu đủ lớn:

$$\hat{\theta} \sim \mathcal{N}(\theta_0, I(\theta_0)^{-1}) \tag{4..17}$$

trong đó $I(\theta)$ là ma trận thông tin Fisher:

$$I(\theta) = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right] \tag{4..18}$$

e. Kiểm định giả thuyết với MLE

Sau khi ước lượng θ, ta có thể kiểm định giả thuyết bằng kiểm định Wald:

$$z = \frac{\hat{\theta} - \theta_0}{\text{SE}(\hat{\theta})} \sim \mathcal{N}(0, 1) \tag{4..19}$$

trong đó $\text{SE}(\hat{\theta})$ là sai số chuẩn của $\hat{\theta}$.

Phương pháp hợp lý tối đa (MLE) là một phương pháp mạnh mẽ để ước lượng tham số của mô hình xác suất. MLE có nhiều tính chất quan trọng như tính nhất quán, hiệu quả và không chệch tiệm cận. Trong thực tế, MLE được

áp dụng rộng rãi trong thống kê, kinh tế lượng, machine learning và nhiều lĩnh vực khác.

4.1.3. Ước lượng Hậu nghiệm Tối đa (Maximum A Posteriori - MAP)

a. Giới thiệu

Ước lượng MAP là một phương pháp thống kê trong khuôn khổ Bayesian, được sử dụng để tìm tham số θ sao cho xác suất hậu nghiệm $P(\theta|D)$ đạt giá trị lớn nhất.

MAP là một mở rộng của Ước lượng hợp lý tối đa (MLE) bằng cách đưa thêm phân bố tiên nghiệm $P(\theta)$ vào mô hình, giúp kiểm soát nhiễu và tránh hiện tượng quá khớp.

b. Công thức toán học

Theo định lý Bayes, xác suất hậu nghiệm được tính bởi:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{4..20}$$

Trong đó:

- $P(\theta|D)$ là xác suất hậu nghiệm của tham số θ.
- $P(D|\theta)$ là hàm hợp lý (likelihood) – xác suất của dữ liệu quan sát được khi biết tham số θ.
- $P(\theta)$ là phân bố tiên nghiệm (prior) của θ.
- $P(D)$ là hàm bằng chứng:

$$P(D) = \int P(D|\theta)P(\theta)d\theta \tag{4..21}$$

Vì $P(D)$ là một hằng số, nên việc tối đa hóa $P(\theta|D)$ tương đương với tối đa hóa tử số:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) \tag{4..22}$$

c. So sánh với MLE

Nếu phân bố tiên nghiệm $P(\theta)$ là đều (Uniform prior), ta có:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(D|\theta) \tag{4..23}$$

Điều này chính là ước lượng MLE.

*** Ví dụ cụ thể**

=> MAP với phân phối Gaussian

Giả sử ta muốn ước lượng tham số θ từ dữ liệu $D = \{x_1, x_2, ..., x_n\}$ theo mô hình:

$$x_i \sim \mathcal{N}(\theta, \sigma^2)$$

với tiên nghiệm:

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

**** Hàm hợp lý:**

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right)$$

**** Phân bố tiên nghiệm:**

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right)$$

**** Xác suất hậu nghiệm:**

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Lấy log của biểu thức trên và tối đa hóa theo θ:

$$\log P(\theta|D) = -\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2} - \frac{(\theta - \mu_0)^2}{2\sigma_0^2} + C$$

Lấy đạo hàm theo θ và đặt bằng 0:

$$\sum_{i=1}^{n} \frac{x_i - \theta}{\sigma^2} + \frac{\mu_0 - \theta}{\sigma_0^2} = 0$$

Giải ra được:

$$\hat{\theta}_{MAP} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

Ta thấy rằng MAP là trung bình có trọng số giữa trung bình dữ liệu và giá trị tiên nghiệm.

- Nếu $\sigma_0^2 \to \infty$ (tức là không có prior), ta thu được MLE:

$$\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Nếu $n \to \infty$, dữ liệu thống trị prior, nên MAP gần với MLE.

d. Ứng dụng của MAP

- Học máy (Machine Learning): MAP được dùng trong Hồi quy Bayesian và Phân loại Naïve Bayes.
- Xử lý ngôn ngữ tự nhiên (NLP): Smoothing trong mô hình Markov ẩn (HMM).
- Xử lý ảnh (Computer Vision): Khử nhiễu hình ảnh bằng cách thêm prior vào mô hình.
- Kinh tế lượng: MAP giúp giảm hiện tượng đa cộng tuyến trong hồi quy tuyến tính.

Ước lượng MAP giúp ước lượng tham số bằng cách kết hợp thông tin từ dữ liệu và thông tin tiên nghiệm. Nó mở rộng MLE bằng cách thêm prior vào mô hình, giúp ổn định ước lượng trong trường hợp dữ liệu ít hoặc có nhiễu cao.

4.1.4. Ước lượng Bayes đầy đủ (Bayesian Estimation)

a. Giới thiệu

Ước lượng Bayes đầy đủ là một phương pháp suy luận thống kê dựa trên lý thuyết Bayes, kết hợp thông tin từ dữ liệu quan sát với thông tin tiên nghiệm (prior) để đưa ra ước lượng xác suất của tham số cần ước lượng.

b. Công thức Bayes cho ước lượng tham số

Giả sử ta có dữ liệu quan sát $D = \{x_1, x_2, \ldots, x_n\}$ và cần ước lượng tham số θ. Theo định lý Bayes, xác suất hậu nghiệm của tham số θ được tính bằng:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{4..24}$$

trong đó:

- $P(D|\theta)$ là **hàm hợp lý (likelihood)**, thể hiện xác suất quan sát dữ liệu D khi biết tham số θ.
- $P(\theta)$ là **phân bố tiên nghiệm (prior distribution)** của θ.
- $P(D)$ là **bằng chứng (evidence)**, được tính bằng:

$$P(D) = \int P(D|\theta)P(\theta)d\theta \tag{4..25}$$

c. Ước lượng Bayes đầy đủ và kỳ vọng hậu nghiệm

Ước lượng Bayes đầy đủ của tham số θ thường được lấy là **kỳ vọng hậu nghiệm (posterior mean)**:

$$\hat{\theta}_{\text{Bayes}} = E[\theta|D] = \int \theta P(\theta|D) d\theta \tag{4..26}$$

Ngoài ra, có thể chọn **trung vị hậu nghiệm** hoặc **chế độ hậu nghiệm (MAP - Maximum A Posteriori)**:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\theta|D) \tag{4..27}$$

Nếu prior là phân bố không thông tin (non-informative prior) hoặc khi kích thước mẫu lớn, thì ước lượng Bayes thường hội tụ về Ước lượng hợp lý tối đa (MLE).

* Ví dụ: Ước lượng Bayes với tham số của phân phối Gaussian

Giả sử dữ liệu $D = \{x_1, x_2, ..., x_n\}$ được lấy mẫu từ phân phối Gaussian:

$$x_i \sim \mathcal{N}(\theta, \sigma^2) \tag{4..28}$$

với phương sai σ^2 đã biết. Ta muốn ước lượng tham số θ theo phương pháp Bayes.

* Bước 1: Chọn phân bố tiên nghiệm

Giả sử ta chọn prior của θ là một phân phối Gaussian:

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{4..29}$$

* Bước 2: Tính xác suất hậu nghiệm

Hàm hợp lý của dữ liệu là:

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \theta)^2}{2\sigma^2}\right) \tag{4..30}$$

Theo định lý Bayes, xác suất hậu nghiệm của θ là:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{4..31}$$

Lấy log của xác suất hậu nghiệm và tối đa hóa theo θ, ta thu được:

$$\hat{\theta}_{\text{Bayes}} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{4..32}$$

**** Nhận xét:**

- $\hat{\theta}_{\text{Bayes}}$ là trung bình có trọng số giữa giá trị tiên nghiệm μ_0 và trung bình mẫu của dữ liệu.
- Nếu prior không có thông tin (tức là $\sigma_0^2 \to \infty$), thì $\hat{\theta}_{\text{Bayes}}$ hội tụ về MLE:

$$\hat{\theta}_{\text{MLE}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4..33}$$

d. So sánh với các phương pháp khác

Tiêu chí	MAP	Bayes (Kỳ vọng hậu nghiệm)	MLE
Cách chọn	$\arg\max P(\theta\|D)$	$E[\theta\|D]$	$\arg\max P(D\|\theta)$
Ảnh hưởng của prior	Có	Có (mạnh hơn MAP)	Không
Khi $n \to \infty$	Hội tụ về MLE	Hội tụ về MLE	Chính là MLE

Bảng 4..1: So sánh các phương pháp ước lượng

e. Ứng dụng thực tế

- **Học máy**: Naive Bayes, Gaussian Process Regression.
- **Khoa học dữ liệu**: Phân tích dữ liệu với bất định cao.
- **Y học**: Ước lượng hiệu quả của thuốc.
- **Kinh tế lượng**: Dự báo tài chính bằng mô hình Bayes.

Ước lượng Bayes đầy đủ kết hợp thông tin tiên nghiệm và dữ liệu quan sát để đưa ra kết quả chính xác hơn so với phương pháp MLE. Tuy nhiên, tính toán có thể phức tạp và cần sử dụng phương pháp xấp xỉ như MCMC.

4.2. Kiểm định giả thuyết thống kê (Hypothesis Testing)

4.2.1. Giá trị p (p-value)

a. Định nghĩa p-value

Trong thống kê, p-value (giá trị p) là xác suất thu được kết quả ít nhất cực đoan như quan sát thực tế, giả sử rằng giả thuyết không (null hypothesis, H_0) là đúng.

Công thức toán học: p-value được định nghĩa là:

$$p = P(T \geq T_{obs}|H_0) \tag{4..34}$$

trong đó:

- T là thống kê kiểm định,
- T_{obs} là giá trị thống kê kiểm định tính toán từ mẫu dữ liệu,
- H_0 là giả thuyết không.

b. Cách tính p-value

Cách tính p-value phụ thuộc vào loại kiểm định:

- **Kiểm định một phía:** p-value được tính bằng xác suất của phần đuôi của phân phối thống kê kiểm định vượt quá giá trị quan sát được.

$$p = P(T \geq T_{obs}|H_0) \quad \text{hoặc} \quad p = P(T \leq T_{obs}|H_0) \tag{4..35}$$

- **Kiểm định hai phía:** p-value là tổng của hai xác suất đuôi của phân phối thống kê kiểm định:

$$p = 2\min\{P(T \geq T_{obs}|H_0), P(T \leq T_{obs}|H_0)\} \tag{4..36}$$

c. Ý nghĩa của p-value

- Nếu p-value nhỏ hơn mức ý nghĩa α (thường là 0.05 hoặc 0.01), bác bỏ giả thuyết không H_0.
- Nếu p-value lớn hơn mức ý nghĩa α, không đủ bằng chứng để bác bỏ H_0.

*** Ví dụ minh họa**

Giả sử chúng ta thực hiện kiểm định giả thuyết với thống kê kiểm định tuân theo phân phối chuẩn chuẩn hóa $N(0, 1)$ và giá trị quan sát được là $T_{obs} = 2.1$. Khi đó:

$$p = P(Z \geq 2.1) = 1 - \Phi(2.1) \approx 0.0179 \tag{4..37}$$

Nếu chọn mức ý nghĩa $\alpha = 0.05$, ta bác bỏ giả thuyết không.

4.2.2. Kiểm định giả thuyết về hệ số hồi quy/Kiểm định t (t-test)

a. Giới thiệu về kiểm định t trong hồi quy

Trong mô hình hồi quy tuyến tính tổng quát:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{4..38}$$

Phương pháp bình phương nhỏ nhất (OLS) được sử dụng để ước lượng các hệ số hồi quy $\hat{\beta}_j$. Sau khi ước lượng, ta cần kiểm định xem các hệ số này có ý nghĩa thống kê hay không. Kiểm định t (t-test) được sử dụng để đánh giá xem một hệ số hồi quy β_j có khác 0 một cách có ý nghĩa thống kê hay không.

b. Xây dựng giả thuyết kiểm định

Với mỗi hệ số hồi quy β_j, ta có giả thuyết kiểm định:

- **Giả thuyết không (H_0)**: Hệ số hồi quy không có ý nghĩa thống kê.

$$H_0 : \beta_j = 0 \tag{4..39}$$

- **Giả thuyết đối (H_1)**: Hệ số hồi quy có ý nghĩa thống kê.

$$H_1 : \beta_j \neq 0 \tag{4..40}$$

c. Thống kê kiểm định t

Thống kê kiểm định t được tính theo công thức:

$$t_j = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \tag{4..41}$$

trong đó:

- $\hat{\beta}_j$ là ước lượng của hệ số hồi quy β_j.
- $SE(\hat{\beta}_j)$ là sai số chuẩn của $\hat{\beta}_j$, được tính bởi:

$$SE(\hat{\beta}_j) = \sqrt{\sigma^2 (X'X)^{-1}_{jj}} \tag{4..42}$$

với σ^2 là phương sai của sai số ngẫu nhiên, ước lượng bởi:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n - k - 1} \tag{4..43}$$

trong đó:

  - $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ là phần dư của mô hình hồi quy.
  - n là số quan sát.
  - k là số biến độc lập trong mô hình (không tính hằng số).

d. Phân phối của thống kê t

Thống kê kiểm định t tuân theo phân phối **Student t** với $n-k-1$ **bậc tự do**.

- Nếu kích thước mẫu n lớn ($n > 30$), phân phối t gần với phân phối chuẩn $N(0, 1)$.
- Nếu n nhỏ, ta phải sử dụng bảng phân phối t để tìm giá trị tới hạn.

e. Quy tắc ra quyết định

Với mức ý nghĩa α (thường chọn $\alpha = 0.05$ hoặc $\alpha = 0.01$), ta xác định giá trị tới hạn $t_{\alpha/2,n-k-1}$ từ bảng phân phối t.

- Nếu $|t_j| > t_{\alpha/2,n-k-1}$, ta bác bỏ giả thuyết $H_0 \Rightarrow$ Kết luận rằng β_j có ý nghĩa thống kê.
- Nếu $|t_j| \leq t_{\alpha/2,n-k-1}$, ta không đủ cơ sở bác bỏ $H_0 \Rightarrow$ Kết luận rằng không có đủ bằng chứng để khẳng định β_j khác 0.

Ngoài ra, ta cũng có thể sử dụng **p-value**:

- Nếu **p-value** $< \alpha \Rightarrow$ bác bỏ H_0, hệ số có ý nghĩa.
- Nếu **p-value** $\geq \alpha \Rightarrow$ không bác bỏ H_0, hệ số không có ý nghĩa.

*** Ví dụ minh họa**

Giả sử ta có mô hình hồi quy:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{4..44}$$

với ước lượng OLS thu được:

$$\hat{\beta}_1 = 2.5, \quad SE(\hat{\beta}_1) = 0.8 \tag{4..45}$$

Số quan sát $n = 25$, số biến $k = 1$, nên bậc tự do $df = 25 - 1 - 1 = 23$.
Thống kê t:

$$t_1 = \frac{2.5}{0.8} = 3.125 \tag{4..46}$$

Tra bảng phân phối t với $df = 23$ và $\alpha = 0.05$, ta có giá trị tới hạn:

$$t_{0.025,23} \approx 2.069 \tag{4..47}$$

Vì $3.125 > 2.069$, ta bác bỏ $H_0 \Rightarrow$ Kết luận rằng β_1 có ý nghĩa thống kê ở mức $\alpha = 0.05$.

Kiểm định t giúp đánh giá mức độ ảnh hưởng của từng biến độc lập trong mô hình hồi quy tuyến tính. Khi sử dụng kiểm định này, cần đảm bảo các giả định của OLS được thỏa mãn để đảm bảo tính chính xác của kết quả kiểm định.

4.2.3. Kiểm định F

a. Giới thiệu về kiểm định F

Kiểm định F được sử dụng để kiểm tra xem toàn bộ mô hình hồi quy có ý nghĩa thống kê hay không. Cụ thể, nó kiểm định giả thuyết rằng tất cả các hệ số hồi quy (ngoại trừ hằng số) đều bằng 0.
Giả thuyết kiểm định:

- H_0: Các hệ số hồi quy không có ý nghĩa thống kê, tức là $\beta_1 = \beta_2 = \cdots = \beta_k = 0$.
- H_1: Ít nhất một trong các hệ số hồi quy khác 0.

Nếu bác bỏ H_0, ta kết luận rằng ít nhất một biến độc lập có ảnh hưởng đáng kể đến biến phụ thuộc.

b. Công thức kiểm định F

Thống kê F được tính bằng công thức:

$$F = \frac{\left(\frac{SST-SSR}{k}\right)}{\left(\frac{SSR}{n-k-1}\right)} \tag{4..48}$$

Trong đó:

- SST (Total Sum of Squares) là tổng bình phương tổng thể:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{4..49}$$

- SSR (Sum of Squared Residuals) là tổng bình phương phần dư:

$$SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4..50}$$

- SSE (Sum of Squares for Regression) là tổng bình phương hồi quy:

$$SSE = SST - SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{4..51}$$

Mô hình có:

- n là số quan sát.
- k là số biến độc lập.
- $n - k - 1$ là bậc tự do của phần dư.

c. Phân phối của thống kê F

Thống kê F tuân theo phân phối F với hai bậc tự do:

- $df_1 = k$ (số biến độc lập).
- $df_2 = n - k - 1$ (số quan sát trừ đi số tham số cần ước lượng).

Nếu giá trị **p-value** nhỏ hơn mức ý nghĩa α (thường là 0.05), ta bác bỏ giả thuyết H_0 và kết luận rằng mô hình có ý nghĩa thống kê.

d. Ý nghĩa của kiểm định F

- Nếu giá trị F lớn và p-value nhỏ, mô hình có ý nghĩa tổng thể.
- Nếu giá trị F nhỏ và p-value lớn, mô hình không có ý nghĩa, tức là biến độc lập không giải thích được biến phụ thuộc.

4.2.4. Kiểm định hiện tượng phương sai sai số thay đổi (heteroskedasticity)

a. Giới thiệu

Trong mô hình hồi quy tuyến tính, giả định phương sai của sai số là không đổi (homoskedasticity). Tuy nhiên, nếu phương sai của sai số thay đổi theo biến độc lập, ta có hiện tượng phương sai sai số thay đổi (heteroskedasticity). Điều này có thể dẫn đến ước lượng không hiệu quả trong hồi quy OLS và ảnh hưởng đến kiểm định giả thuyết.

b. Mô hình toán học

Xét mô hình hồi quy tuyến tính:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \tag{4..52}$$

với giả định:

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma_i^2. \tag{4..53}$$

Nếu σ_i^2 không phải là hằng số mà phụ thuộc vào một hoặc nhiều biến độc lập, ta có heteroskedasticity:

$$\sigma_i^2 = g(x_{i1}, x_{i2}, \ldots, x_{ik}). \tag{4..54}$$

c. Các kiểm định phương sai sai số thay đổi

*** Kiểm định Breusch-Pagan**

Kiểm định Breusch-Pagan kiểm tra xem phương sai của sai số có phụ thuộc vào biến độc lập không. Cụ thể, hồi quy phần dư bình phương $\hat{\varepsilon}_i^2$ theo biến độc lập:

$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_k x_{ik} + u_i. \tag{4..55}$$

Sau đó, sử dụng thống kê kiểm định LM:

$$LM = nR^2 \sim \chi_k^2, \tag{4..56}$$

trong đó R^2 là hệ số xác định từ hồi quy trên. Nếu giá trị LM lớn hơn giá trị tới hạn, bác bỏ giả thuyết không có heteroskedasticity.

*** Kiểm định White**

Kiểm định White kiểm tra heteroskedasticity mà không giả định cấu trúc cụ thể của phương sai sai số. Hồi quy phần dư bình phương theo tất cả các biến độc lập và bình phương của chúng:

$$\hat{\varepsilon}_i^2 = \gamma_0 + \sum_{j=1}^{k} \gamma_j x_{ij} + \sum_{j=1}^{k} \sum_{m=j}^{k} \gamma_{jm} x_{ij} x_{im} + u_i. \tag{4..57}$$

Thống kê kiểm định tương tự như Breusch-Pagan:

$$LM = nR^2 \sim \chi_m^2, \tag{4..58}$$

trong đó m là số bậc tự do.

Hiện tượng phương sai sai số thay đổi làm ảnh hưởng đến tính hiệu quả của ước lượng OLS. Các kiểm định như Breusch-Pagan và White giúp phát hiện heteroskedasticity để điều chỉnh mô hình phù hợp.

4.2.5. Kiểm định tự tương quan

a. Giới thiệu

Tự tương quan (autocorrelation) là hiện tượng khi các sai số trong mô hình hồi quy không độc lập với nhau. Điều này thường xuất hiện trong dữ liệu chuỗi thời gian hoặc khi có yếu tố hệ thống chưa được mô hình hóa đúng.

b. Mô hình hồi quy tuyến tính

Giả sử mô hình hồi quy tuyến tính:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t, \tag{4..59}$$

trong đó ε_t là sai số ngẫu nhiên.

Nếu tồn tại tự tương quan, ta có:

$$Cov(\varepsilon_t, \varepsilon_{t-1}) \neq 0. \tag{4..60}$$

c. Kiểm định Durbin-Watson

Một trong những kiểm định phổ biến để phát hiện tự tương quan bậc nhất là kiểm định Durbin-Watson (DW). Thống kê kiểm định DW được tính như sau:

$$DW = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}. \tag{4..61}$$

Giá trị DW nằm trong khoảng $[0, 4]$ và được diễn giải như sau: - $DW \approx 2$: Không có tự tương quan bậc nhất. - $DW < 2$: Có tự tương quan dương. - $DW > 2$: Có tự tương quan âm.

d. Kiểm định Breusch-Godfrey

Kiểm định Durbin-Watson chỉ áp dụng cho tự tương quan bậc nhất. Để kiểm định tự tương quan bậc cao hơn, ta sử dụng kiểm định Breusch-Godfrey:

1. Hồi quy mô hình gốc và lấy phần dư $\hat{\varepsilon}_t$. 2. Hồi quy phần dư theo chính nó:

$$\hat{\varepsilon}_t = \rho_1 \hat{\varepsilon}_{t-1} + \rho_2 \hat{\varepsilon}_{t-2} + \cdots + \rho_p \hat{\varepsilon}_{t-p} + \eta_t. \tag{4..62}$$

3. Kiểm định giả thuyết:

- H_0: Không có tự tương quan bậc p.
- H_1: Có tự tương quan bậc p.

Kiểm định dựa trên thống kê LM:

$$LM = nR^2 \sim \chi^2_p, \tag{4..63}$$

trong đó R^2 là hệ số xác định từ hồi quy phụ.

Nếu phát hiện tự tương quan, có thể sử dụng các phương pháp như mô hình sai số tổng quát (GLS) hoặc ước lượng Newey-West để hiệu chỉnh phương sai của ước lượng.

Phần II.

Mô hình hồi quy tuyến tính

Chương 5.

Mô hình hồi quy tuyến tinh đơn giản

Chương 6.

Mô hình hồi quy tuyến tính với biến tiên lượng phân nhóm

Chương 7.

Mô hình hồi quy đa biến

Chương 8.

Mô hình hồi quy đa thức

Chương 9.

Mô hình hồi quy vững chắc (Robust Regression)

Chương 10.

Mô hình hồi quy đa biến đa thức (Multivariate Polynomial Regression)

Phần III.

Mô hình hồi quy phi tuyến

Chương 11.

Mô hình hồi quy hàm mũ (Exponential Regression)

Chương 12.

Mô hình hồi quy logarit (Logarithmic Regression)

Chương 13.

Mô hình hồi quy hàm Cobb-Douglas (Cobb-Douglas Regression)

Chương 14.

Mô hình hồi quy Logistic (Logistic Regression)

Chương 15.

Mô hình hồi quy Probit (Probit Regression)

Chương 16.

Mô hình hồi quy Tobit (Tobit Regression)

Chương 17.

Mô hình hồi quy Poisson (Poisson Regression)

Phần IV.

Hồi quy sống còn (Survival Regression)

Chương 18.

Mô hình hồi quy Cox (Cox Proportional Hazards Model)

Chương 19.

Mô hình Weibull (Weibull Regression Model)

Chương 20.

Mô hình hồi quy Log-logistic (Log-logistic Regression Model)

Chương 21.

Mô hình hồi quy Gamma (Gamma Regression Model)

Chương 22.

Mô hình hồi quy hỗn hợp (Frailty Models)

Phần V.

Ước lượng Bayesian

Chương 23.

Giới thiệu về Ước lượng Bayesian

Chương 24.

Phân phối trước, phân phối hậu nghiệm và quy trình tính toán

Chương 25.

Ước lượng Bayesian trong mô hình hồi quy

Chương 26.

MCMC và ứng dụng trong mô hình kinh tế lượng

Chương 27.

Các ứng dụng thực tế và ví dụ minh họa

Phần VI.

Phân tích dữ liệu chuỗi thời gian (Time Series Data)

Chương 28.

Tổng quan về chuỗi thời gian

28.1. Các khái niệm cơ bản: tính dừng, tự tương quan, mùa vụ

28.2. Biểu diễn và phân tích dữ liệu chuỗi thời gian

Chương 29.

Mô hình ARIMA và các biến thể

29.1. Mô hình AR, MA, ARMA, ARIMA

29.2. Phương pháp chọn bậc mô hình tối ưu (AIC, BIC)

29.3. Dự báo bằng ARIMA

Chương 30.

Mô hình ARCH/GARCH

30.1. Biến động tài chính và mô hình ARCH/GARCH

30.2. Ước lượng tham số và dự báo biến động

Phần VII.

Kinh tế lượng không gian (Spatial Econometrics)

Chương 31.

Tổng quan về kinh tế lượng không gian

Chương 32.

Mô hình hồi quy không gian

Phần VIII.

Machine Learning trong kinh tế lượng

Chương 33.

Các phương pháp Machine Learning trong Kinh tế lượng

33.1. Giới thiệu về Machine Learning trong Kinh tế lượng

Machine Learning (ML) ngày càng được ứng dụng rộng rãi trong kinh tế lượng nhằm nâng cao độ chính xác của mô hình dự báo, kiểm định mô hình và xử lý dữ liệu lớn. Sự khác biệt chính giữa ML và các phương pháp kinh tế lượng truyền thống là cách tiếp cận dữ liệu: ML tập trung vào tính linh hoạt và tối ưu hóa dự báo, trong khi kinh tế lượng truyền thống thường dựa trên lý thuyết kinh tế và kiểm định giả thuyết.

33.2. Hồi quy tuyến tính mở rộng: Ridge, Lasso, Elastic Net

33.3. Mô hình cây quyết định và boosting (Random Forest, XGBoost

33.4. Machine Learning nhân quả: Double ML, Causal Inference

33.5. Deep Learning trong phân tích kinh tế

Chương 34.

Xử lý dữ liệu lớn trong Kinh tế lượng

34.1. Tiền xử lý dữ liệu kinh tế (missing data, outliers, scaling)

34.2. Chọn biến và giảm chiều dữ liệu (PCA, Feature Selection)

34.3. Xử lý dữ liệu bảng (panel data) bằng ML

34.4. Dữ liệu thời gian thực và vấn đề xử lý dữ liệu lớn

Chương 35.

Hồi quy và Dự báo với Machine Learning

35.1. So sánh hồi quy truyền thống với ML

35.2. Mô hình XGBoost, Random Forest và hồi quy phi tuyến

35.3. Đánh giá mô hình dự báo (MAPE, RMSE, R-squared)

35.4. ML và các phương pháp Bayes trong dự báo kinh tế lượng

Chương 36.

Machine Learning trong phân tích nhân quả và chính sách

36.1. Machine Learning nhân quả (Causal Forest, Double ML)

36.2. Xác định tác động chính sách bằng ML

36.3. Kiểm định giả thuyết và ML trong phân tích chính sách

Chương 37.

Machine Learning trong phân tích dữ liệu bảng (Panel Data)

37.1. Xử lý dữ liệu bảng lớn với ML

37.2. So sánh Fixed Effects, Random Effects với ML

37.3. Ứng dụng ML vào phân tích tác động theo thời gian

Chương 38.

Machine Learning trong phân tích dữ liệu chuỗi thời gian

38.1. Mô hình hóa dữ liệu chuỗi thời gian bằng ML

38.2. So sánh ARIMA, VAR với ML (XGBoost, LSTM)

38.3. Phát hiện xu hướng và cú sốc kinh tế bằng ML

Phần IX.

Phương pháp Monte Carlo và mô phỏng

Chương 39.

Ứng dụng mô phỏng trong kinh tế

39.1. Mô phỏng dữ liệu kinh tế lượng

39.1.1. Mô hình tổng quát

Chúng ta xét một mô hình hồi quy tuyến tính bội như sau:

$$GDP_t = \beta_0 + \beta_1 Investment_t + \beta_2 Consumption_t + \varepsilon_t \tag{39..1}$$

với:

- GDP_t là tổng sản phẩm quốc nội tại thời điểm t,
- $Investment_t$ là đầu tư tại thời điểm t,
- $Consumption_t$ là tiêu dùng tại thời điểm t,
- $\varepsilon_t \sim N(0, \sigma^2)$ là nhiễu ngẫu nhiên có phân phối chuẩn với trung bình bằng 0 và phương sai σ^2,
- $\beta_0, \beta_1, \beta_2$ là các tham số cần ước lượng.

39.1.2. Mô phỏng dữ liệu bằng phương pháp Monte Carlo

Bước 1: Xác định số lượng quan sát
Chọn số lượng quan sát T, ví dụ $T = 1000$, để mô phỏng dữ liệu.

Bước 2: Xác định giá trị thực của các tham số
Chọn giá trị thực của các tham số:

$$\beta_0 = 5, \quad \beta_1 = 2, \quad \beta_2 = 3 \tag{39..2}$$

Bước 3: Sinh dữ liệu giả lập cho biến độc lập
Sinh các biến $Investment_t$ và $Consumption_t$ từ phân phối chuẩn hoặc phân phối khác phù hợp với thực tế. Ví dụ:

$$Investment_t \sim N(10, 5^2), \quad Consumption_t \sim N(20, 10^2) \tag{39..3}$$

Bước 4: Sinh nhiễu ngẫu nhiên ε_t
Sinh ε_t từ phân phối chuẩn:

$$\varepsilon_t \sim N(0, 2^2) \tag{39..4}$$

Bước 5: Tính giá trị của GDP_t theo phương trình hồi quy

$$GDP_t = 5 + 2Investment_t + 3Consumption_t + \varepsilon_t \tag{39..5}$$

Bước 6: Ước lượng tham số hồi quy
Sử dụng phương pháp bình phương nhỏ nhất (OLS - Ordinary Least Squares) để ước lượng các tham số β. Hệ số ước lượng được tính theo công thức:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{39..6}$$

trong đó:

- X là ma trận thiết kế:

$$X = \begin{bmatrix} 1 & Investment_1 & Consumption_1 \\ 1 & Investment_2 & Consumption_2 \\ \vdots & \vdots & \vdots \\ 1 & Investment_T & Consumption_T \end{bmatrix} \tag{39..7}$$

- Y là vector kết quả:

$$Y = \begin{bmatrix} GDP_1 \\ GDP_2 \\ \vdots \\ GDP_T \end{bmatrix} \tag{39..8}$$

- $\hat{\beta}$ là vector các ước lượng:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \tag{39..9}$$

Bước 7: Lặp lại quy trình
Lặp lại các bước trên N lần (ví dụ $N = 1000$) để tạo ra nhiều bộ dữ liệu khác nhau và phân tích sự ổn định của các ước lượng $\hat{\beta}$.

39.1.3. Đánh giá tính ổn định của ước lượng

Sau khi thực hiện mô phỏng Monte Carlo, ta có thể phân tích sự ổn định của các ước lượng bằng cách:

- Tính **trung bình của các ước lượng** $\hat{\beta}$ trên toàn bộ các lần lặp. Nếu $E[\hat{\beta}] \approx \beta$, thì ước lượng không chệch.

- Tính **phương sai của các ước lượng** để đánh giá độ tin cậy của mô hình. Nếu phương sai nhỏ, mô hình có độ tin cậy cao.
- Vẽ **biểu đồ phân phối của** $\hat{\beta}$ để kiểm tra tính chuẩn của ước lượng.

Tóm lại:

- Mô phỏng dữ liệu kinh tế lượng giúp kiểm tra tính ổn định của ước lượng bằng phương pháp Monte Carlo.
- Ta có thể sinh ngẫu nhiên các biến đầu vào, sau đó sử dụng OLS để ước lượng tham số.
- Việc lặp lại nhiều lần giúp đánh giá tính ổn định của các ước lượng β.

Dưới đây là mã Python để thực hiện mô phỏng Monte Carlo theo mô hình kinh tế lượng đã đề cập:

```
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt

def monte_carlo_simulation(n_simulations=1000, n_samples=100):
    beta_0, beta_1, beta_2 = 2, 0.5, 0.3
    beta_estimates = []

    for _ in range(n_simulations):
        investment = np.random.normal(50, 10, n_samples)
        consumption = np.random.normal(30, 5, n_samples)
        epsilon = np.random.normal(0, 5, n_samples)

        gdp = beta_0 + beta_1 * investment + beta_2 * consumption + epsilon

        X = np.column_stack((np.ones(n_samples), investment, consumption))
        y = gdp

        model = sm.OLS(y, X).fit()
        beta_estimates.append(model.params)

    beta_estimates = np.array(beta_estimates)

    plt.figure(figsize=(10, 5))
```

```
25      plt.hist(beta_estimates[:, 1], bins=30, alpha=0.6,
    label="Beta 1")
26      plt.hist(beta_estimates[:, 2], bins=30, alpha=0.6,
    label="Beta 2")
27      plt.axvline(beta_1, color='r', linestyle='dashed',
    linewidth=2, label='True Beta 1')
28      plt.axvline(beta_2, color='g', linestyle='dashed',
    linewidth=2, label='True Beta 2")
29      plt.xlabel("Gia tri uoc luong")
30      plt.ylabel("Tan suat")
31      plt.title("Phan phoi cua uoc luong Beta")
32      plt.legend()
33      plt.show()
34
35      return beta_estimates
36
37  beta_estimates = monte_carlo_simulation()
38
```

Mã Python trên thực hiện mô phỏng Monte Carlo với các tham số đã nêu, sử dụng hồi quy OLS để ước lượng hệ số và trực quan hóa phân phối của các ước lượng.

39.2. Mô hình dự báo tài chính bằng mô phỏng - Minh họa với trường hợp mô phỏng giá cổ phiếu bằng ARIMA + Monte Carlo

39.2.1. Giới thiệu phương pháp

Phương pháp này kết hợp mô hình **ARIMA** để ước lượng chuỗi thời gian giá cổ phiếu và phương pháp **Monte Carlo** để mô phỏng nhiều kịch bản giá trong tương lai với các sai số ngẫu nhiên.

- **ARIMA (AutoRegressive Integrated Moving Average)**: Dự báo giá cổ phiếu dựa trên cấu trúc tự hồi quy và trung bình trượt của chuỗi thời gian.

- **Monte Carlo Simulation**: Dự báo bằng cách tạo nhiều đường giá tương lai với nhiễu ngẫu nhiên để đánh giá sự không chắc chắn.

39.2.2. Mô hình ARIMA

Mô hình ARIMA(p, d, q) được xác định bởi:

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d Y_t = (1 + \sum_{j=1}^{q} \theta_j L^j)\epsilon_t \tag{39..10}$$

Trong đó:

- p là số bậc tự hồi quy (AR).
- d là số lần lấy sai phân (I).
- q là số bậc trung bình trượt (MA).
- $\epsilon_t \sim N(0, \sigma^2)$ là nhiễu trắng.

Sau khi ước lượng mô hình ARIMA từ dữ liệu lịch sử, ta sẽ sử dụng nó để dự báo xu hướng giá tương lai.

39.2.3. Mô phỏng Monte Carlo dựa trên ARIMA

Sau khi có mô hình ARIMA, ta mô phỏng Monte Carlo bằng cách:

1. Lấy **giá dự báo từ ARIMA**.
2. Thêm **nhiễu ngẫu nhiên** $\epsilon_t \sim N(0, \hat{\sigma}^2)$.
3. Lặp lại nhiều lần để tạo các kịch bản khác nhau.

Công thức dự báo giá cổ phiếu:

$$S_{t+1} = \hat{S}_{t+1}^{ARIMA} + \sigma \cdot Z_t \tag{39..11}$$

với $Z_t \sim N(0, 1)$, mô phỏng nhiều kịch bản giá khác nhau.

Dưới đây là mã Python để thực hiện mô phỏng giá cổ phiếu bằng mô hình ARIMA kết hợp với Monte Carlo.

Cài đặt thư viện

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA


num_simulations = 1000
forecast_days = 30
```

Đọc dữ liệu và ước lượng mô hình ARIMA

```

np.random.seed(42)
returns = np.random.normal(0, 1, 500)
prices = 100 + np.cumsum(returns)

df = pd.DataFrame({'Price': prices})


model = ARIMA(df['Price'], order=(1, 1, 1))
model_fit = model.fit()
```

Mô phỏng Monte Carlo

```

simulated_paths = np.zeros((num_simulations, forecast_days)
    )

for i in range(num_simulations):
forecast = model_fit.forecast(steps=forecast_days)
noise = np.random.normal(0, np.std(forecast), forecast_days
    )
simulated_paths[i, :] = forecast + noise
```

Hiển thị kết quả

```

plt.figure(figsize=(10, 5))
plt.plot(simulated_paths.T, color='lightblue', alpha=0.2)
plt.plot(np.mean(simulated_paths, axis=0), color='red',
    label='Trung binh du bao')
plt.legend()
plt.title('Mo phong gia co phieu bang ARIMA + Monte Carlo')
plt.xlabel('Ngay')
plt.ylabel('Gia co phieu')
plt.show()
```

Phương pháp kết hợp ARIMA và mô phỏng Monte Carlo giúp dự báo kịch bản giá cổ phiếu với độ không chắc chắn được mô hình hóa bằng nhiễu ngẫu nhiên.

Chương 40.

Giới thiệu phương pháp Monte Carlo

40.1. Tổng quan về phương pháp Monte Carlo

Phương pháp Monte Carlo là một kỹ thuật tính toán dựa trên mô phỏng ngẫu nhiên để tìm nghiệm số của các bài toán phức tạp. Trong kinh tế lượng, Monte Carlo được sử dụng để:

- Kiểm định tính chính xác của các ước lượng thống kê.
- Phân tích tính chất phân phối của các thống kê kiểm định.
- Đánh giá hiệu suất của các mô hình kinh tế lượng.

Cho một mô hình kinh tế lượng:

$$Y = f(X, \theta) + \varepsilon \tag{40..1}$$

trong đó:

- Y là biến phụ thuộc,
- X là ma trận biến độc lập,
- θ là vector tham số cần ước lượng,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ là nhiễu ngẫu nhiên.

Monte Carlo có thể được sử dụng để phân tích các tính chất của ước lượng $\hat{\theta}$.

40.2. Mô phỏng Monte Carlo trong kiểm định giả thuyết kinh tế lượng

40.2.1. Mô tả quy trình Monte Carlo

Để đánh giá một ước lượng $\hat{\theta}$, ta sử dụng quy trình mô phỏng Monte Carlo sau:

1. Chọn một giá trị thật của tham số θ_0 và xác định mô hình kinh tế lượng.
2. Sinh ngẫu nhiên tập dữ liệu giả lập (X, Y) theo mô hình:

$$Y_i = X_i\theta_0 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{40..2}$$

3. Ước lượng tham số $\hat{\theta}$ bằng phương pháp ước lượng như OLS:

$$\hat{\theta} = (X'X)^{-1}X'Y \tag{40..3}$$

4. Lặp lại bước 2 và 3 với nhiều bộ dữ liệu giả lập (X, Y) để có các ước lượng $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(M)}$.
5. Phân tích phân phối của $\hat{\theta}$ để kiểm tra tính chệch, phương sai và hiệu suất của phương pháp ước lượng.

40.2.2. Tính chệch và phương sai của ước lượng

Từ các lần lặp Monte Carlo, ta có thể tính kỳ vọng và phương sai của ước lượng:

$$E[\hat{\theta}] \approx \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{(m)} \tag{40..4}$$

$$\mathrm{Var}(\hat{\theta}) \approx \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}^{(m)} - E[\hat{\theta}])^2 \tag{40..5}$$

- Nếu $E[\hat{\theta}] = \theta_0$, phương pháp ước lượng là không chệch.
- Nếu phương sai nhỏ, phương pháp có độ chính xác cao.

40.2.3. Ứng dụng trong kiểm định giả thuyết

Trong kiểm định giả thuyết, Monte Carlo được dùng để kiểm tra phân phối của các thống kê kiểm định. Ví dụ, với kiểm định giả thuyết:

$$H_0 : \theta = \theta_0 \tag{40..6}$$

ta có thể sử dụng thống kê t:

$$t = \frac{\hat{\theta} - \theta_0}{s_{\hat{\theta}}} \tag{40..7}$$

- $s_{\hat{\theta}}$ là độ lệch chuẩn của $\hat{\theta}$.
- Nếu Monte Carlo cho thấy phân phối của t không tuân theo phân phối chuẩn t, thì kiểm định truyền thống có thể không đáng tin cậy.