Trần Minh Tâm

XÂY DỰNG MÔ HÌNH KINH TẾ LƯỢNG HIỆN ĐẠI VỚI PYTHON

TP. Hồ Chí Minh - 2025

Mục lục

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:

$$\sum d_i^2 = 1^2 + (-1)^2 + (-1)^2 + (-1)^2 + 2^2 = 1 + 1 + 1 + 1 + 4 = 8 \tag{2.33}$$

Thay vào công thức tính Spearman:

$$r_s = 1 - \frac{6(8)}{5(25-1)} = 1 - \frac{48}{120} = 1 - 0.4 = 0.6 \tag{2.34}$$

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.

3.6 Kết Luận

Trong chương này, chúng ta đã tìm hiểu về các phân bố xác suất quan trọng, các công thức liên quan và một số định lý quan trọng. Những kiến thức này sẽ là nền tảng cho các ứng dụng trong thống kê và khoa học dữ liệu.