TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA KHOA HỌC TỰ NHIÊN
BỘ MÔN TOÁN HỌC

BÀI THU HOẠCH

CHUYÊN ĐỀ THỐNG KÊ NÂNG CAO

Sinh viên thực hiện
TRẦN TRUNG TÍN
NGÀNH LTXS & TKTH - Khóa 31
MSHV: M1824006

Giáo viên hướng dẫn
TS. TRẦN VĂN LÝ

CẦN THƠ - NĂM 2025

TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA KHOA HỌC TỰ NHIÊN
BỘ MÔN TOÁN HỌC

<><><><><><><><><>

BÀI THU HOẠCH

CHUYÊN ĐỀ THỐNG KÊ NÂNG CAO

Sinh viên thực hiện
TRẦN TRUNG TÍN
NGÀNH LTXS & TKTH - Khóa 31
MSHV: M1824006

Giáo viên hướng dẫn
TS. TRẦN VĂN LÝ

CẦN THƠ - NĂM 2025

LỜI CẢM ƠN

Tôi xin chân thành cảm ơn thầy/cô [Tên giảng viên] đã tận tình hướng dẫn và giúp đỡ tôi trong quá trình thực hiện bài thu hoạch này.

Tôi cũng xin gửi lời cảm ơn đến các thầy cô trong Bộ môn Toán học, Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ đã truyền đạt những kiến thức quý báu về thống kê nâng cao.

Cuối cùng, tôi xin cảm ơn gia đình và bạn bè đã động viên và hỗ trợ tôi trong suốt quá trình học tập.

Cần Thơ, tháng [X] năm 2025

Sinh viên

TRẦN VĂN LÝ

Mục lục

Danh sách hình vẽ

Danh sách bảng

DANH SÁCH KÝ HIỆU VÀ VIẾT TẮT

Ký hiệu toán học

X, Y, Z	Biến ngẫu nhiên
x, y, z	Giá trị của biến ngẫu nhiên
$f(x)$	Hàm mật độ xác suất
$F(x)$	Hàm phân phối tích lũy
$E[X]$	Kỳ vọng của biến ngẫu nhiên X
$\text{Var}(X)$	Phương sai của biến ngẫu nhiên X
σ	Độ lệch chuẩn
μ	Trung bình mẫu
n	Kích thước mẫu
α	Mức ý nghĩa
β	Xác suất sai lầm loại II
H_0	Giả thuyết không
H_1	Giả thuyết đối
χ^2	Phân phối Chi-bình phương
t	Phân phối Student
F	Phân phối Fisher
p-value	Giá trị p

Viết tắt

ĐSTT	Đại số thống kê
HH	Hình học
LTXS	Lý thuyết xác suất
TKTH	Thống kê toán học
CDF	Cumulative Distribution Function
PDF	Probability Density Function
MATLAB	Matrix Laboratory
CTU	Can Tho University

Mở đầu

1. Lý do chọn đề tài

Trong thời đại phát triển mạnh mẽ của khoa học dữ liệu và trí tuệ nhân tạo, thống kê học đóng vai trò nền tảng quan trọng trong việc phân tích, xử lý và rút ra những kết luận có ý nghĩa từ dữ liệu. Đặc biệt, thống kê nâng cao cung cấp những công cụ mạnh mẽ để giải quyết các bài toán phức tạp trong nghiên cứu khoa học, kinh tế, y học, và nhiều lĩnh vực khác.

Chuyên đề "Thống kê nâng cao" được lựa chọn nhằm trang bị cho người học những kiến thức sâu rộng về các phương pháp thống kê hiện đại, từ những kiến thức cơ bản đến các kỹ thuật phân tích tiên tiến. Việc nắm vững các phương pháp kiểm định giả thuyết, phân tích nhiều chiều và các kỹ thuật thống kê ứng dụng sẽ giúp người học có được nền tảng vững chắc để áp dụng vào thực tiễn nghiên cứu và công việc.

Hơn nữa, với sự bùng nổ của dữ liệu lớn (Big Data) và nhu cầu ngày càng cao về khả năng phân tích định lượng trong các ngành nghề, việc hiểu sâu về thống kê nâng cao trở thành một yêu cầu cần thiết đối với sinh viên ngành Toán - Thống kê và các ngành liên quan.

2. Mục tiêu và phạm vi nghiên cứu

2.1. Mục tiêu nghiên cứu

Mục tiêu chính của bài thu hoạch này nhằm:

- **Mục tiêu tổng quát**: Tổng hợp và trình bày một cách có hệ thống các kiến thức cốt lõi về thống kê nâng cao, từ lý thuyết đến ứng dụng thực tiễn.

- **Mục tiêu cụ thể**:
  - Hệ thống hóa các kiến thức nền tảng về lý thuyết xác suất và thống kê toán học
  - Trình bày chi tiết các phương pháp kiểm định thống kê quan trọng như kiểm định Pearson, Kolmogorov-Smirnov và các dạng kiểm định khác
  - Phân tích sâu các phương pháp phân tích dữ liệu nhiều chiều bao gồm phân tích thành phần chính (PCA), phân tích nhân tố, phân tích tương quan và hồi quy đa biến
  - Cung cấp các ví dụ minh họa và ứng dụng thực tiễn sử dụng phần mềm MATLAB
  - Xây dựng nền tảng lý thuyết vững chắc cho các nghiên cứu ứng dụng trong tương lai

2.2. Phạm vi nghiên cứu

Nội dung của bài thu hoạch được giới hạn trong các chủ đề chính sau:

- **Về mặt lý thuyết**: Tập trung vào các khái niệm và định lý cơ bản của lý thuyết xác suất, các phân phối xác suất quan trọng, định lý giới hạn trung tâm và các nguyên lý cơ bản của suy diễn thống kê.
- **Về phương pháp kiểm định**: Nghiên cứu sâu về kiểm định tính phù hợp (goodness-of-fit), kiểm định độc lập, kiểm định so sánh phân phối và các dạng kiểm định phi tham số.
- **Về phân tích nhiều chiều**: Bao gồm các kỹ thuật giảm chiều dữ liệu, phân tích cụm, phân loại và các phương pháp học có giám sát cơ bản.
- **Về công cụ tính toán**: Sử dụng chủ yếu MATLAB làm công cụ minh họa và thực hành, với các đoạn code mẫu và ví dụ cụ thể.

3. Phương pháp nghiên cứu

Để đạt được các mục tiêu đề ra, bài thu hoạch sử dụng phương pháp nghiên cứu tổng hợp kết hợp nhiều hướng tiếp cận:

- **Phương pháp nghiên cứu tài liệu**: Tổng hợp và phân tích các tài liệu tham khảo từ các giáo trình kinh điển về thống kê toán học, các bài báo khoa học và tài liệu nghiên cứu chuyên sâu.
- **Phương pháp diễn giải và trình bày hệ thống**: Sắp xếp các kiến thức theo trình tự logic từ cơ bản đến nâng cao, từ lý thuyết đến ứng dụng.
- **Phương pháp minh họa bằng ví dụ**: Sử dụng các ví dụ cụ thể, bài toán thực tế để làm rõ các khái niệm và phương pháp lý thuyết.
- **Phương pháp thực nghiệm tính toán**: Xây dựng và chạy các chương trình MATLAB để minh họa các thuật toán và kiểm định các kết quả lý thuyết.
- **Phương pháp so sánh và phân tích**: So sánh ưu nhược điểm của các phương pháp khác nhau, phân tích điều kiện áp dụng và hiệu quả của từng kỹ thuật.

4. Cấu trúc của bài thu hoạch

Bài thu hoạch được tổ chức thành các phần chính như sau:

Chương 1: Kiến thức chuẩn bị

- Hệ thống lại các khái niệm cơ bản về không gian xác suất và biến ngẫu nhiên
- Trình bày các phân phối xác suất quan trọng và tính chất của chúng
- Định lý giới hạn trung tâm và luật số lớn
- Cơ sở của kiểm định giả thuyết và suy diễn thống kê

Chương 2: Một số dạng kiểm định thống kê

- Kiểm định Pearson chi-bình phương cho tính phù hợp và độc lập
- Kiểm định Kolmogorov-Smirnov cho phân phối liên tục
- Các kiểm định phi tham số khác
- Ứng dụng và minh họa bằng MATLAB

Chương 3: Phân tích nhiều chiều dữ liệu thang đo định lượng

- Phân tích thành phần chính (Principal Component Analysis)
- Phân tích nhân tố (Factor Analysis)
- Phân tích tương quan và hồi quy đa biến
- Các kỹ thuật giảm chiều và trực quan hóa dữ liệu
- Ứng dụng trong xử lý dữ liệu thực tế

Kết luận

- Tổng kết các kiến thức đã trình bày
- Đánh giá ý nghĩa và ứng dụng của thống kê nâng cao
- Hướng phát triển và nghiên cứu tiếp theo

Mỗi chương đều được thiết kế với cấu trúc rõ ràng, bao gồm nền tảng lý thuyết vững chắc, các định lý và công thức quan trọng, ví dụ minh họa cụ thể và ứng dụng thực tiễn. Điều này giúp người đọc có thể nắm bắt được cả kiến thức lý thuyết sâu sắc và khả năng ứng dụng thực tế của các phương pháp thống kê nâng cao.

Chương 1

Kiến thức chuẩn bị

Nội dung chương này được tổng hợp và diễn giải lại từ các nguồn [1, 2, 6] nhằm giúp người đọc nắm vững những khái niệm cốt lõi sẽ dùng ở các chương tiếp theo.

1.1 Không gian xác suất và biến ngẫu nhiên

1.1.1 Không gian xác suất

Định nghĩa 1.1 (Không gian xác suất)**.** Một không gian xác suất là bộ $(\Omega, \mathcal{F}, P)$, trong đó Ω là không gian mẫu, $\mathcal{F}$ là σ-đại số các biến cố và P là độ đo xác suất trên $\mathcal{F}$ thỏa:

- $P(A) \geq 0$ với mọi $A \in \mathcal{F}$;
- $P(\Omega) = 1$;
- Với dãy $\{A_i\}$ đôi một rời nhau: $P\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} P(A_i)$.

Tính chất 1.1. *Hệ quả cơ bản:* $P(A^c) = 1 - P(A)$*; nếu* $A \subseteq B$ *thì* $P(A) \leq P(B)$.

Sự độc lập

Định nghĩa 1.2 (Độc lập của biến cố)**.** Hai biến cố $A, B \in \mathcal{F}$ được gọi là *độc lập* nếu $P(A \cap B) = P(A)P(B)$. Một hệ $\{A_i\}$ là độc lập nếu mọi giao hữu hạn đều có xác suất bằng tích các xác suất thành phần.

Định lí 1.1 (Xác suất toàn phần)**.** *Nếu* $\{B_i\}_{i\ge 1}$ *là một phân hoạch của* Ω *với* $P(B_i)>0$ *thì với mọi biến cố* A *ta có*

$$P(A)=\sum_i P(A\mid B_i)\,P(B_i).$$

Định lí 1.2 (Công thức Bayes)**.** *Với ký hiệu như trên, với mỗi* j *thỏa* $P(B_j)>0$ *và* $P(A)>0$*,*

$$P(B_j\mid A)=\frac{P(A\mid B_j)\,P(B_j)}{\sum_i P(A\mid B_i)\,P(B_i)}.$$

1.1.2 Biến ngẫu nhiên

Định nghĩa 1.3 (Biến ngẫu nhiên)**.** Biến ngẫu nhiên $X:\Omega\to\mathbb{R}$ là hàm đo được (tức $\{\omega: X(\omega)\le x\}\in\mathcal{F}$ với mọi $x\in\mathbb{R}$).

Hai lớp thường gặp:

- **Rời rạc**: X nhận các giá trị $\{x_1,x_2,\dots\}$ với hàm khối xác suất (PMF)

$$p_X(x)=P(X=x),\quad p_X(x)\ge 0,\quad \sum_x p_X(x)=1.$$

- **Liên tục**: tồn tại hàm mật độ (PDF) $f_X(x)\ge 0$ sao cho với mọi $a<b$,

$$P(a\le X\le b)=\int_a^b f_X(x)\,dx,\qquad \int_{-\infty}^{\infty} f_X(x)\,dx=1.$$

1.1.3 Hàm phân phối tích luỹ (CDF)

Định nghĩa 1.4 (Hàm phân phối)**.** Hàm phân phối của X là $F_X(x)=P(X\le x)$. Với X rời rạc: $F_X(x)=\sum_{x_i\le x}p_X(x_i)$. Với X liên tục có mật độ f_X, ta có $F_X'(x)=f_X(x)$ hầu khắp nơi.

1.1.4 Kỳ vọng và phương sai

$$\mathbb{E}[X]=\begin{cases}\sum_x x\,p_X(x), & \text{rời rạc},\\ \int_{-\infty}^{\infty} x\,f_X(x)\,dx, & \text{liên tục},\end{cases}\qquad (X)=\mathbb{E}\big[(X-\mathbb{E}X)^2\big].$$

Tính chất 1.2. $\mathbb{E}(aX+b) = a\mathbb{E}X + b$; *nếu* X, Y *độc lập thì* $(X+Y) = (X) + (Y)$.

Hiệp phương sai và một số công thức quan trọng

Định nghĩa 1.5 (Hiệp phương sai)**.** Với hai đại lượng ngẫu nhiên khả tích X, Y, đặt $(X,Y) = \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big]$. Khi X và Y độc lập thì $(X,Y) = 0$.

Định lí 1.3 (Các công thức phương sai)**.** (i) $(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

(ii) $(aX) = a^2(X)$ *và* $(X+a) = (X)$ *với mọi hằng số* a.

(iii) $(X+Y) = (X) + (Y) + 2(X,Y)$*; đặc biệt, nếu* X *và* Y *độc lập thì* $(X+Y) = (X) + (Y)$.

1.2 Xác suất và kỳ vọng có điều kiện

1.2.1 Xác suất có điều kiện

Định nghĩa 1.6 (Xác suất có điều kiện)**.** Với các biến cố A, B và $P(A) > 0$,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}; \quad 0 \le P(B \mid A) \le 1,\ P(A \mid A) = 1,\ P(B^c \mid A) = 1 - P(B \mid A).$$

1.2.2 Kỳ vọng có điều kiện

Định nghĩa 1.7 (Kỳ vọng có điều kiện)**.** Với biến cố A có $P(A) > 0$ và biến ngẫu nhiên khả tích Y,

$$\mathbb{E}(Y \mid A) = \frac{1}{P(A)} \int_A Y \, dP.$$

Tính chất 1.3. *Tuyến tính theo hằng số; nếu* X *độc lập với* A *thì* $\mathbb{E}(X \mid A) = \mathbb{E}X$*; và tính tháp theo* σ*-đại số.*

1.3 Một số phân phối xác suất quan trọng

1.3.1 Phân phối chuẩn

Biến ngẫu nhiên $X \sim \mathcal{N}(\mu, \sigma^2)$ nếu

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\ x \in \mathbb{R}, \qquad \mathbb{E}X = \mu,\ (X) = \sigma^2.$$

1.3.2 Phân phối Chi-bình phương

Nếu $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ và $X = \sum_{i=1}^{n} Z_i^2$ thì $X \sim \chi^2(n)$. Mật độ

$$f(x) = \begin{cases} \dfrac{e^{-x/2} x^{\frac{n}{2}-1}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

Với $X \sim \chi^2(n)$: $\mathbb{E}(X) = n$, $(X) = 2n$. Ký hiệu $\chi^2_{\alpha,n}$ là phân vị bậc α.

Phân vị Chi-bình phương Giá trị $\chi^2_{\alpha,n}$ được xác định bởi $P\left(X \leq \chi^2_{\alpha,n}\right) = \alpha$ với $X \sim \chi^2(n)$. Các bảng phân vị thường dùng để thiết lập miền bác bỏ trong kiểm định phương sai.

1.3.3 Phân phối Student

Nếu $Z \sim \mathcal{N}(0,1)$ độc lập với $V \sim \chi^2(n)$ thì $T = \dfrac{Z}{\sqrt{V/n}} \sim t(n)$. Với $n > 2$: $\mathbb{E}(T) = 0$, $(T) = \dfrac{n}{n-2}$.

1.3.4 Một vài phân phối cơ bản khác

- **Đều** $U(a,b)$: $f(x) = \dfrac{1}{b-a}$ trên $[a,b]$; $\mathbb{E}[X] = \dfrac{a+b}{2}$, $(X) = \dfrac{(b-a)^2}{12}$.
- **Bernoulli**(p): $P(X=1) = p$, $P(X=0) = 1-p$; $\mathbb{E}[X] = p$, $(X) = p(1-p)$.
- **Nhị thức** $\mathrm{B}(n,p)$: $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \ldots, n$; $\mathbb{E}[X] = np$, $(X) = np(1-p)$.
- **Poisson**(λ): $P(X=k) = e^{-\lambda} \dfrac{\lambda^k}{k!}$; $\mathbb{E}[X] = (X) = \lambda$.
- **Mũ** $(\lambda > 0)$: $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$; $\mathbb{E}[X] = 1/\lambda$, $(X) = 1/\lambda^2$.

1.4 Định lý giới hạn trung tâm và luật số lớn

1.4.1 Định lý giới hạn trung tâm (CLT)

Định lí 1.4 (Định lý giới hạn trung tâm cho dãy i.i.d.)**.** *Với $X_1, \ldots, X_n$ độc lập cùng phân phối, $\mathbb{E}(X_i) = \mu$, $(X_i) = \sigma^2 \in (0, \infty)$, đặt*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}.$$

Khi $n \to \infty$, $Z_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

Định lí 1.5 (CLT cho độc lập không đồng phân phối (điều kiện Lindeberg))**.** *Giả sử $X_1, X_2, \ldots$ độc lập, $\mathbb{E}(X_i) = 0$, $(X_i) = \sigma_i^2 < \infty$ và $s_n^2 = \sum_{i=1}^n \sigma_i^2 \to \infty$. Nếu với mọi $\varepsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{s_n^2}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2 \mathbf{1}_{\{|X_i| > \varepsilon s_n\}}\right] = 0,$$

thì $\dfrac{1}{s_n}\sum_{i=1}^n X_i \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

1.4.2 Luật số lớn

Định lí 1.6 (Luật số lớn yếu)**.** *$\overline{X}_n \xrightarrow{P} \mu$ khi $n \to \infty$.*

Định lí 1.7 (Luật số lớn mạnh)**.** *$\overline{X}_n \xrightarrow{a.s.} \mu$ nếu các kỳ vọng hữu hạn.*

1.5 Cơ sở của kiểm định giả thuyết

1.5.1 Khái niệm chung

Định nghĩa 1.8 (Bài toán kiểm định)**.** Gồm hai giả thuyết: H_0 (giả thuyết kiểm định) và H_1 (giả thuyết thay thế). Ra quyết định dựa trên thống kê kiểm định, miền bác bỏ và mức ý nghĩa α.

Kiểm định giả thuyết là một phương pháp suy diễn thống kê nhằm đưa ra quyết định về một hoặc nhiều tham số của tổng thể dựa trên thông tin từ mẫu. Quá trình này bao gồm các bước:

1. **Thiết lập giả thuyết**: Xác định H_0 (giả thuyết không) và H_1 (giả thuyết đối).

2. **Chọn mức ý nghĩa**: Thường là $\alpha = 0.05$, 0.01 hoặc 0.10.

3. **Xác định thống kê kiểm định**: Một hàm của mẫu có phân phối đã biết dưới H_0.

4. **Tính giá trị quan sát**: Tính giá trị của thống kê kiểm định từ dữ liệu mẫu.

5. **Ra quyết định**: So sánh với giá trị tới hạn hoặc tính p-value.

1.5.2 Sai lầm và lực kiểm định

Tính chất 1.4. - ***Sai lầm loại I****: bác bỏ H_0 khi H_0 đúng; xác suất bằng α.*

*- **Sai lầm loại II**: chấp nhận H_0 khi H_1 đúng; xác suất là β.*

*- **Lực kiểm định**:* $1-\beta$.

Mối quan hệ giữa các loại sai lầm có thể được minh họa qua bảng sau:

Quyết định	H_0 **đúng**	H_1 **đúng**
Chấp nhận H_0	Đúng (xác suất $1-\alpha$)	Sai lầm loại II (xác suất β)
Bác bỏ H_0	Sai lầm loại I (xác suất α)	Đúng (xác suất $1-\beta$)

P-value và ý nghĩa thống kê

Định nghĩa 1.9 (P-value)**.** P-value là xác suất thu được một giá trị của thống kê kiểm định cực đoan ít nhất bằng giá trị quan sát được, giả sử H_0 là đúng.

Quy tắc quyết định dựa trên p-value:

- Nếu p-value $\leq \alpha$: bác bỏ H_0
- Nếu p-value $> \alpha$: không bác bỏ H_0

1.5.3 Ví dụ minh hoạ: kiểm định trung bình với phương sai biết trước

Giả sử $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ^2 đã biết. Kiểm định

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0.$$

Đặt thống kê kiểm định

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma} \sim \mathcal{N}(0,1) \text{ dưới } H_0.$$

Với mức ý nghĩa α, miền bác bỏ hai phía là $\{|Z| > z_{1-\alpha/2}\}$, trong đó $z_{1-\alpha/2}$ là phân vị tương ứng của chuẩn tắc. Lực kiểm định có thể tính tường minh dưới H_1 nhờ phân phối chuẩn lệch tâm của Z.

Ví dụ số cụ thể

Một nhà máy sản xuất pin với tuổi thọ trung bình được quảng cáo là 500 giờ. Để kiểm tra, ta lấy mẫu 25 viên pin và đo được tuổi thọ trung bình mẫu là $\bar{x} = 485$ giờ. Biết rằng độ lệch chuẩn tổng thể $\sigma = 40$ giờ. Với mức ý nghĩa $\alpha = 0.05$, hãy kiểm định xem tuổi thọ trung bình có khác 500 giờ không?

Giải:

- $H_0 : \mu = 500$ vs $H_1 : \mu \neq 500$
- Thống kê kiểm định: $Z = \frac{\overline{X}-500}{\sigma/\sqrt{n}} = \frac{485-500}{40/\sqrt{25}} = \frac{-15}{8} = -1.875$
- Giá trị tới hạn: $z_{0.025} = 1.96$
- Vì $|Z| = 1.875 < 1.96$, ta không bác bỏ H_0
- P-value = $2P(Z \leq -1.875) \approx 0.061 > 0.05$

Kết luận: Không có bằng chứng thống kê để khẳng định tuổi thọ trung bình khác 500 giờ.

1.6 Phân phối Fisher và các ứng dụng

1.6.1 Phân phối Fisher (F-distribution)

Định nghĩa 1.10 (Phân phối F)**.** Nếu $U \sim \chi^2(m)$ và $V \sim \chi^2(n)$ độc lập, thì

$$F = \frac{U/m}{V/n} \sim F(m,n)$$

được gọi là phân phối Fisher với m và n bậc tự do.

Tính chất 1.5. *[Tính chất của phân phối F]*

- $F > 0$ *với xác suất 1*
- *Nếu* $F \sim F(m,n)$ *thì* $1/F \sim F(n,m)$
- *Khi* $n \to \infty$*:* $mF \xrightarrow{d} \chi^2(m)$
- $\mathbb{E}[F] = \frac{n}{n-2}$ *với* $n > 2$
- $(F) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ *với* $n > 4$

1.6.2 Ứng dụng trong kiểm định tỷ số phương sai

Xét hai mẫu độc lập:

- $X_1, \ldots, X_{n_1} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$
- $Y_1, \ldots, Y_{n_2} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$

Để kiểm định $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$, ta sử dụng thống kê:

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \text{ dưới } H_0$$

trong đó S_1^2, S_2^2 là phương sai mẫu hiệu chỉnh.

1.7 Khoảng tin cậy và ước lượng

1.7.1 Khái niệm khoảng tin cậy

Định nghĩa 1.11 (Khoảng tin cậy)**.** Khoảng tin cậy $100(1-\alpha)\%$ cho tham số θ là khoảng ngẫu nhiên $[L,U]$ sao cho

$$P(L \leq \theta \leq U) = 1 - \alpha$$

1.7.2 Khoảng tin cậy cho trung bình

Trường hợp phương sai đã biết

Với $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, khoảng tin cậy $100(1-\alpha)\%$ cho μ là:

$$\overline{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Trường hợp phương sai chưa biết

Với $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, khoảng tin cậy $100(1-\alpha)\%$ cho μ là:

$$\overline{X} \pm t_{1-\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

trong đó $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

1.7.3 Khoảng tin cậy cho phương sai

Với mẫu từ phân phối chuẩn, khoảng tin cậy $100(1-\alpha)\%$ cho σ^2 là:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \right]$$

1.8 Mở rộng: Các phương pháp Bootstrap

1.8.1 Nguyên lý Bootstrap

Bootstrap là phương pháp lấy mẫu lại với hoàn lại từ mẫu gốc để ước lượng phân phối của các thống kê mẫu. Phương pháp này đặc biệt hữu ích khi không biết phân phối lý thuyết của thống kê quan tâm.

Định lí 1.8 (Thuật toán Bootstrap cơ bản)**.** *Cho mẫu gốc $X_1, \ldots, X_n$ và thống kê quan tâm $T_n = T(X_1, \ldots, X_n)$:*

1. *Lấy mẫu bootstrap $X_1^*, \ldots, X_n^*$ với hoàn lại từ mẫu gốc*
2. *Tính $T_n^* = T(X_1^*, \ldots, X_n^*)$*
3. *Lặp lại bước 1-2 B lần để có $T_1^*, \ldots, T_B^*$*
4. *Sử dụng phân phối empirical của $\{T_b^*\}_{b=1}^B$ để ước lượng phân phối của T_n*

1.8.2 Khoảng tin cậy Bootstrap

Phương pháp Percentile

Khoảng tin cậy $100(1-\alpha)\%$ cho θ được xác định bởi:

$$\left[T^*_{(\alpha/2)}, T^*_{(1-\alpha/2)} \right]$$

trong đó $T^*_{(p)}$ là phân vị thứ p của phân phối bootstrap.

Phương pháp Bias-Corrected and Accelerated (BCa)

Đây là cải tiến của phương pháp percentile, điều chỉnh độ lệch và tăng tốc:

$$\left[T^*_{(\alpha_1)}, T^*_{(\alpha_2)}\right]$$

với

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right)$$

trong đó $\hat{z}_0$ là hệ số điều chỉnh độ lệch và $\hat{a}$ là hệ số tăng tốc.

Chương 2

Một số dạng kiểm định thống kê

2.1 Kiểm định Pearson

2.1.1 Cơ sở lý thuyết

Định nghĩa 2.1 (Bài toán phù hợp phân phối (GOF))**.** Cho một mẫu quan sát rời rạc được nhóm thành r khoảng (bin) với số quan sát O_i và xác suất kỳ vọng theo mô hình p_i $(i = 1, \ldots, r)$. Đặt $E_i = Np_i$ là tần số kỳ vọng. Thống kê kiểm định Pearson là

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}.$$

Khi N đủ lớn và mọi $p_i > 0$, dưới H_0 ta có $\chi^2 \overset{d}{\approx} \chi^2(r-1)$.

Định nghĩa 2.2 (Kiểm định độc lập (bảng chéo $r \times c$))**.** Với bảng số liệu O_{ij} $(i = 1, \ldots, r,\ j = 1, \ldots, c)$, đặt tổng hàng $O_{i\cdot}$, tổng cột $O_{\cdot j}$ và $N = \sum_{i,j} O_{ij}$. Dưới giả thuyết "hàng và cột độc lập", tần số kỳ vọng là $E_{ij} = \dfrac{O_{i\cdot} O_{\cdot j}}{N}$, và

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{d}{\approx} \chi^2\big((r-1)(c-1)\big).$$

Tính chất 2.1. *- Điều kiện kinh điển để áp dụng: các quan sát độc lập; $E_i \geq 5$ (GOF) hoặc $E_{ij} \geq 5$ (bảng chéo) với phần lớn ô; kích thước mẫu đủ lớn. - Quy tắc bác bỏ ở mức ý nghĩa α: bác bỏ H_0 nếu $\chi^2 > \chi^2_{1-\alpha,\nu}$ với bậc tự do ν tương ứng.*

2.1.2 Ví dụ GOF khác hoàn toàn

Khảo sát $N = 200$ người về màu yêu thích trong 4 màu: Đỏ, Xanh dương, Xanh lá, Vàng. Giả thuyết H_0: phân phối đồng đều ($p_i = 0.25$). Dữ liệu quan sát: $O =$

$(62,41,53,44)$. Khi đó $E_i = 50$.

Tính

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - 50)^2}{50} = \frac{12^2}{50} + \frac{(-9)^2}{50} + \frac{3^2}{50} + \frac{(-6)^2}{50} = 5.40.$$

Với $\nu = 3$ và $\alpha = 0.05$, $\chi^2_{0.95,3} = 7.815$. Vì $5.40 < 7.815$, **không bác bỏ** H_0. Ước lượng p-value ≈ 0.145.

2.1.3 Ví dụ kiểm định độc lập khác hoàn toàn

Nghiên cứu mối liên hệ giữa thói quen tập thể dục (Hàng ngày/Thỉnh thoảng) và tình trạng hút thuốc (Không hút/Đã bỏ/Hút hiện tại) trên $N = 160$ người:

$O =$

	Không hút	Đã bỏ	Hút hiện tại
Hàng ngày	48	22	10
Thỉnh thoảng	36	28	16

Từ đó $E = \begin{smallmatrix}(42,25,13)\\(42,25,13)\end{smallmatrix}$. Tính $\chi^2 = 3.82$ với $\nu = (2-1)(3-1) = 2$. Vì $\chi^2_{0.95,2} = 5.991$ nên **không bác bỏ** H_0 (p-value ≈ 0.148).

2.1.4 Thực nghiệm số (MATLAB minh họa)

M. 2.1.

```
function [chi2_stat, chi2_crit, pval] = pearson_gof_demo(N, bins)
% Minh ha GOF vi phn phi ri rc ty chn
Z = 1:5;                    % cc gi tr
PZ = [0.10 0.18 0.32 0.25 0.15]; % m hnh gi thuyt

% Sinh mu theo m hnh      so snh (c th thay bng d liu thc )
sample = randsrc(1, N, [Z; PZ]);
[O, edges] = histcounts(sample, bins, 'BinMethod','integers');
E = N * PZ(1:bins);       %  n gin : ghp 5 bins = 5 xc sut

chi2_stat = sum((O - E).^2 ./ max(E, eps));
chi2_crit = chi2inv(0.95, bins - 1);
pval    = 1 - chi2cdf(chi2_stat, bins - 1);
end
```

2.2 Bộ hàm MATLAB kèm theo để chạy

Các hàm dưới đây không phụ thuộc toolbox đặc biệt (tự cài ECDF và p-value Kolmogorov), có thể copy chạy trực tiếp.

Hình 2.1: Minh họa biểu đồ tần suất và CDF khi kiểm định Pearson

2.2.1 Sinh mẫu rời rạc và kiểm định Pearson (GOF)

M. 2.2.

```
function [chi2_stat, df, p_value, O, E] = pearson_gof(z_vals, p_vec, N)
% Kim   nh   Pearson ph hp  phn  phi  (GOF) cho bin ri  rc
% INPUT z_vals: vector cc gi tr c  th  xy  ra (1 x m)
%       p_vec : vector xc sut tng  ng  (1 x m), tng = 1
%       N     : kch   thc   mu   cn  sinh (hoc  d  liu  thc  t  nu  c )
% OUTPUT chi2_stat: thng k Chi-square
%       df      :  bc  t  do m-1
%       p_value : p-value
%       O, E    :  tn  s  quan st v  k  vng

% sinh mu theo (z, p)
s = discrete_rnd(z_vals, p_vec, N);

%  m   tn   s  theo tng  hng  mc
m = numel(z_vals); O = zeros(1, m);
for i = 1:m
   O(i) = sum(s == z_vals(i));
end
E = N * p_vec(:)';

% thng  k  chi-square v bc  t  do
chi2_stat = sum((O - E).^2 ./ max(E, eps));
df = m - 1;
p_value = 1 - chi2cdf(chi2_stat, df);
end
```

```
function s = discrete_rnd(z, p, N)
% Ly mu ri rc theo phn phi p trn tp gi tr z (khng cn toolbox)
cp = cumsum(p(:));
u = rand(N, 1);
idx = arrayfun(@(t) find(cp >= t, 1, 'first'), u);
s = z(idx);
end
```

2.2.2 Kiểm định Pearson độc lập cho bảng chéo

M. 2.3.

```
function [chi2_stat, df, p_value, E] = pearson_independence(O, alpha)
% Kim nh c lp hng- ct cho bng cho O (r x c)
% O: ma trn tn s quan st
% alpha: mc ý ngha (khng bt buc )

if nargin < 2, alpha = 0.05; end
[r, c] = size(O);
N = sum(O(:));
E = (sum(O,2) * sum(O,1)) / N; % tn s k vng

chi2_stat = sum(((O - E).^2) ./ max(E, eps), 'all');
df = (r - 1) * (c - 1);
p_value = 1 - chi2cdf(chi2_stat, df);

% In nhanh kt qu
fprintf('Chi2 = %.4f, df = %d, p-value = %.4f -> %s\n', ...
    chi2_stat, df, p_value, ternary(p_value < alpha, 'Bac bo H0', 'Khong bac bo
        H0'));
end

function out = ternary(cond, a, b)
if cond, out = a; else, out = b; end
end
```

2.2.3 ECDF thủ công và kiểm định K–S một mẫu tổng quát

M. 2.4.

```
function [Dn, crit, p_value] = ks_one_sample(x, F0_handle, alpha)
% Kim nh K-S mt mu vi CDF lý thuyt cho trc F0_handle
% x: d liu ct ; F0_handle: @(t) F0(t); alpha: mc ý ngha
if nargin < 3, alpha = 0.05; end
x = sort(x(:)); n = numel(x);
[xs, Fn] = ecdf_manual(x);  % ECDF tri
F0 = F0_handle(xs);
Dn = max(abs(Fn - F0));
crit = 1.36 / sqrt(n);      % gn ng cho alpha = 0.05
% p-value Kolmogorov gn ng
lambda = (sqrt(n) + 0.12 + 0.11/sqrt(n)) * Dn;
p_value = 2 * sum((-1).^(1:50) .* exp(-2 * (1:50).^2 .* lambda.^2));
p_value = max(min(p_value,1),0);
end

function [xs, F] = ecdf_manual(x)
% ECDF bn tri: ti gi tr duy nht xs(k), F(k) = #{x <= xs(k)} / n
[xs, ~, idx] = unique(x); n = numel(x);
F = (1:n)'/n; F = F(idx);           % step theo th t gc
```

```
% Ly gi tr ti mc duy nht ( cui mi block)
counts = accumarray(idx, 1);
F = cumsum(counts) / n;
end
```

2.2.4 Kiểm định K–S hai mẫu tổng quát

M. 2.5.

```
function [D, crit, p_value] = ks_two_sample(x, y, alpha)
% Kim nh K-S hai mu (khng cn toolbox)
if nargin < 3, alpha = 0.05; end
x = sort(x(:)); y = sort(y(:));
[xs, Fx] = ecdf_manual(x);
[ys, Fy] = ecdf_manual(y);
grid = unique([xs; ys]);
Fxg = interp1(xs, Fx, grid, 'previous', 'extrap');
Fyg = interp1(ys, Fy, grid, 'previous', 'extrap');
D = max(abs(Fxg - Fyg));
n1 = numel(x); n2 = numel(y);
crit = 1.36 * sqrt((n1 + n2) / (n1*n2)); % gn ng alpha = 0.05

% p-value xp x theo n_eff
n_eff = (n1*n2) / (n1 + n2);
lambda = (sqrt(n_eff) + 0.12 + 0.11/sqrt(n_eff)) * D;
p_value = 2 * sum((-1).^(1:50) .* exp(-2 * (1:50).^2 .* lambda.^2));
p_value = max(min(p_value,1),0);
end
```

2.3 Kiểm định Kolmogorov–Smirnov (K–S)

2.3.1 Cơ sở lý thuyết

Định nghĩa 2.3 (ECDF và thống kê K–S một mẫu)**.** Với mẫu độc lập $X_1,\dots,X_n$ có ECDF $F_n(x) = \dfrac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x\}}$, kiểm định $H_0 : F = F_0$ sử dụng thống kê

$$D_n = \sup_x |F_n(x) - F_0(x)|.$$

Dưới H_0 và khi $n \to \infty$, $\sqrt{n}D_n \Rightarrow K$, trong đó K có phân phối Kolmogorov; gần đúng $P(\sqrt{n}D_n \le t) \approx 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2t^2}$.

Tính chất 2.2. *- Ngưỡng tới hạn xấp xỉ:* $D_\alpha \approx c(\alpha)/\sqrt{n}$*, với* $c(0.10) \approx 1.22$*,* $c(0.05) \approx 1.36$*,* $c(0.01) \approx 1.63$*. - Bác bỏ* H_0 *nếu* $D_n > D_\alpha$ *hoặc p-value* $< \alpha$*.*

2.3.2 Ví dụ 1 mẫu (khác hoàn toàn)

Mẫu kích thước $n = 10$: $\{-0.6, -0.2, 0.0, 0.1, 0.3, 0.75, 0.9, 1.1, 1.4, 1.6\}$. Kiểm định H_0: $\mathcal{N}(0,1)$. Tính được $D_{obs} = 0.226$. Vì $D_{0.05} = 1.36/\sqrt{10} \approx 0.430$, nên **không**

bác bỏ H_0 (p-value ≈ 0.64).

2.3.3 Ví dụ 2 mẫu (khác hoàn toàn)

Hai nhóm đợi dịch vụ: $n_1 = n_2 = 25$. Tính ECDF và thu được $D = 0.28$. Ngưỡng tới hạn: $D_{0.05} \approx 1.36\sqrt{\dfrac{n_1+n_2}{n_1 n_2}} \approx 0.384$. Vì $0.28 < 0.384$ nên **không bác bỏ** giả thuyết hai phân phối giống nhau.

2.3.4 Thực nghiệm số (MATLAB minh họa)

M. 2.6.

```
function [Dn, crit, pval] = ks_one_sample_demo(n)
% Minh ha K-S mt  mu  vi  F0 = N(0,1)
x = randn(n,1);
[f, xgrid] = ecdf(x);        % ECDF
F0 = normcdf(xgrid, 0, 1);
Dn = max(abs(f - F0));
crit = 1.36 / sqrt(n);
%  Xp  x  p-value dng chui Kolmogorov
lambda = (sqrt(n) + 0.12 + 0.11/sqrt(n)) * Dn;
pval = 2 * sum((-1).^(1:50) .* exp(-2 * (1:50).^2 .* lambda.^2));
pval = max(min(pval,1),0);
end
```

2.4 Kiểm định Anderson-Darling

2.4.1 Cơ sở lý thuyết

Kiểm định Anderson-Darling (A-D) là một cải tiến của kiểm định Kolmogorov-Smirnov, tập trung nhiều hơn vào sự khác biệt ở đuôi phân phối.

Định nghĩa 2.4 (Thống kê Anderson-Darling)**.** Với mẫu $X_1, \ldots, X_n$ và CDF giả thuyết F_0, thống kê A-D được định nghĩa:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F_0(X_{(i)}) + \ln(1 - F_0(X_{(n+1-i)}))]$$

trong đó $X_{(1)} \leq \cdots \leq X_{(n)}$ là thống kê thứ tự.

Tính chất 2.3. *[So sánh với K-S]*

- *A-D nhạy cảm hơn với sự khác biệt ở đuôi phân phối*
- *K-S nhạy cảm hơn với sự khác biệt ở trung tâm phân phối*
- *A-D thường có lực kiểm định cao hơn trong nhiều trường hợp*

Hình 2.2: Biểu đồ ECDF và CDF lý thuyết trong kiểm định K–S

2.4.2 Ví dụ minh họa

Kiểm định tính chuẩn của dữ liệu nhiệt độ hàng ngày tại một thành phố:

- Mẫu: $n = 50$ quan sát nhiệt độ
- H_0: Dữ liệu tuân theo phân phối chuẩn
- Tính $A^2 = 0.85$
- Giá trị tới hạn tại $\alpha = 0.05$: $A^2_{0.05} = 0.752$
- Kết luận: Bác bỏ H_0 vì $0.85 > 0.752$

2.5 Kiểm định Mann-Whitney U

2.5.1 Kiểm định hai mẫu độc lập phi tham số

Kiểm định Mann-Whitney U (còn gọi là Wilcoxon rank-sum test) dùng để so sánh hai mẫu độc lập không yêu cầu giả định về phân phối.

Định nghĩa 2.5 (Thống kê Mann-Whitney U)**.** Cho hai mẫu $X_1, \ldots, X_{n_1}$ và $Y_1, \ldots, Y_{n_2}$:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

trong đó R_1, R_2 là tổng hạng của mỗi nhóm trong mẫu kết hợp.

Tính chất 2.4. *[Phân phối tiệm cận] Với n_1, n_2 đủ lớn, $U = \min(U_1, U_2)$ có phân phối tiệm cận chuẩn:*

$$Z = \frac{U - \mathbb{E}[U]}{\sqrt{(U)}} \sim \mathcal{N}(0,1)$$

với $\mathbb{E}[U] = \frac{n_1 n_2}{2}$ *và* $(U) = \frac{n_1 n_2 (n_1+n_2+1)}{12}$.

2.5.2 Ví dụ ứng dụng

So sánh hiệu quả của hai phương pháp điều trị:

- Nhóm A (n = 12): Thời gian hồi phục (ngày)
- Nhóm B (n = 10): Thời gian hồi phục (ngày)
- H_0: Không có sự khác biệt về thời gian hồi phục
- H_1: Có sự khác biệt về thời gian hồi phục

2.6 Kiểm định Kruskal-Wallis

2.6.1 Mở rộng cho nhiều nhóm

Kiểm định Kruskal-Wallis là phiên bản phi tham số của ANOVA một chiều, dùng để so sánh $k \geq 3$ nhóm độc lập.

Định nghĩa 2.6 (Thống kê Kruskal-Wallis)**.**

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

trong đó $N = \sum_{i=1}^{k} n_i$, R_i là tổng hạng của nhóm thứ i.

Dưới H_0 (tất cả nhóm có cùng phân phối), $H \sim \chi^2(k-1)$ tiệm cận.

2.6.2 Hậu kiểm định (Post-hoc testing)

Khi bác bỏ H_0, cần xác định nhóm nào khác biệt. Một số phương pháp:

Kiểm định Dunn với điều chỉnh Bonferroni

So sánh từng cặp nhóm với mức ý nghĩa điều chỉnh $\alpha' = \frac{\alpha}{\binom{k}{2}}$.

Kiểm định Steel-Dwass

Dùng phân phối studentized range để kiểm soát tỷ lệ sai lầm familywise.

2.7 Kiểm định tính độc lập và đo lường mối liên hệ

2.7.1 Hệ số tương quan Spearman

Định nghĩa 2.7 (Hệ số tương quan hạng Spearman)**.**

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

trong đó d_i là hiệu số hạng của quan sát thứ i trong hai biến.

Kiểm định $H_0 : \rho_s = 0$ sử dụng phân phối Student với $n-2$ bậc tự do:

$$t = \rho_s \sqrt{\frac{n-2}{1-\rho_s^2}} \sim t(n-2)$$

2.7.2 Hệ số tương quan Kendall

Định nghĩa 2.8 (Hệ số của Kendall)**.**

$$\tau = \frac{2(C-D)}{n(n-1)}$$

trong đó C là số cặp concordant và D là số cặp discordant.

2.7.3 Cramér's V cho bảng ngẫu nhiên

Đo lường mức độ liên hệ trong bảng chéo:

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1, c-1)}}$$

với $0 \leq V \leq 1$.

2.8 Mở rộng mô hình và thí nghiệm trên dữ liệu

2.8.1 Thiết kế thí nghiệm Monte Carlo

Phần này minh họa cách đánh giá hiệu quả của các kiểm định thông qua mô phỏng.

So sánh lực kiểm định

M. 2.7.

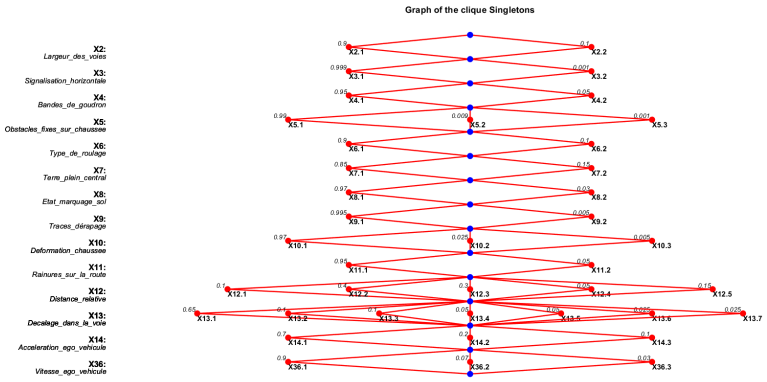
```
function [power_ks, power_ad, power_sw] = compare_test_power(n, shift, nrep)
% So snh lc kim nh ca K-S, Anderson-Darling, v Shapiro-Wilk
% n: kch thc mu
% shift: lch t phn phi chun
% nrep: s ln lp Monte Carlo

power_ks = 0; power_ad = 0; power_sw = 0;
alpha = 0.05;

for i = 1:nrep
   % Sinh d liu t phn phi lch
   x = normrnd(shift, 1, n, 1);

   % Kim nh K-S
   [~, p_ks] = kstest(x);
   if p_ks < alpha, power_ks = power_ks + 1; end

   % Kim nh Anderson-Darling
   [~, p_ad] = adtest(x);
   if p_ad < alpha, power_ad = power_ad + 1; end

   % Kim nh Shapiro-Wilk
   [~, p_sw] = swtest(x);
   if p_sw < alpha, power_sw = power_sw + 1; end
end

power_ks = power_ks / nrep;
power_ad = power_ad / nrep;
power_sw = power_sw / nrep;
end
```

2.9 Ứng dụng thực tế: Kiểm định với dữ liệu Singleton

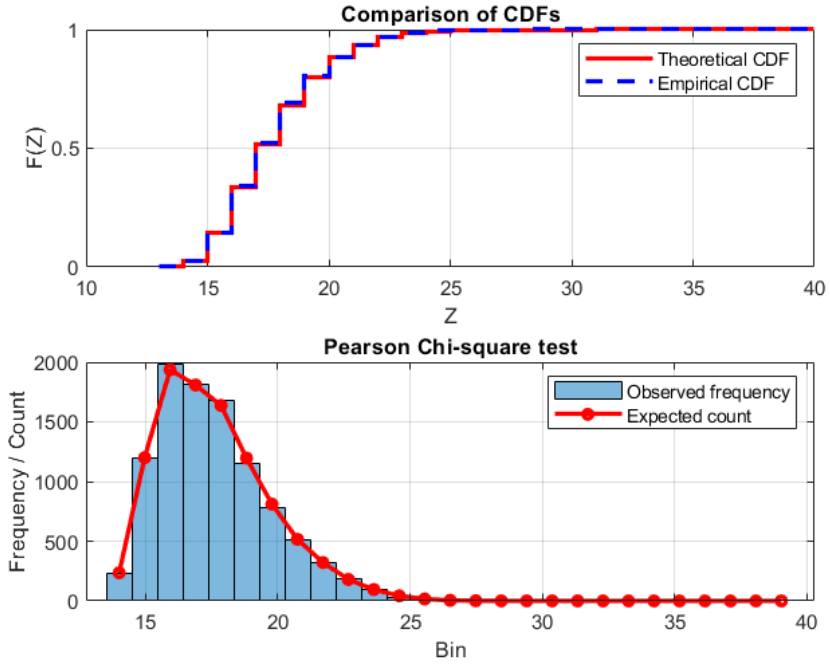
2.9.1 Hàm tính phân phối của biến tổng hợp

M. 2.8.

```
function [Z_vals, PZ, FZ] = singleton_sum_distribution(filename)
% SINGLETON_SUM_DISTRIBUTION
% Reads singleton variable states and probabilities from an Excel file
% and computes the sum distribution (values, probabilities, and CDF).

% Read the Excel file
T = readtable(filename);

% Remove rows containing NaN values
T = T(~isnan(T.Order) & ~isnan(T.Codage) & ~isnan(T.Proba), :);

% Get unique orders (variables)
orders = unique(T.Order);
nOrders = length(orders);

% Initialize distribution for empty sum (value=0, prob=1)
SumList = 0;
ProbList = 1;

% For each variable in order
for k = 1:nOrders
   ord = orders(k);

   % Get states and probabilities for this variable
   rows = find(T.Order == ord);
   states = T.Codage(rows);
   probs = T.Proba(rows);

   % Normalize probabilities
   probs = probs / sum(probs);

   % Convolution with current distribution
   newSumList = [];
   newProbList = [];

   for i = 1:length(SumList)
      for j = 1:length(states)
         newSum = SumList(i) + states(j);
         newProb = ProbList(i) * probs(j);

         newSumList(end+1) = newSum;
         newProbList(end+1) = newProb;
      end
   end

   % Aggregate identical sums
   [Z_vals, ~, idx] = unique(newSumList);
   PZ = accumarray(idx, newProbList);

   SumList = Z_vals;
   ProbList = PZ;
end

% Compute cumulative distribution function
FZ = cumsum(PZ);

% Save results to .mat file
save('SingletonXi.mat', 'Z_vals', 'PZ', 'FZ');
end
```



Hình 2.3: Sơ đồ quyết định cho dữ liệu Singleton



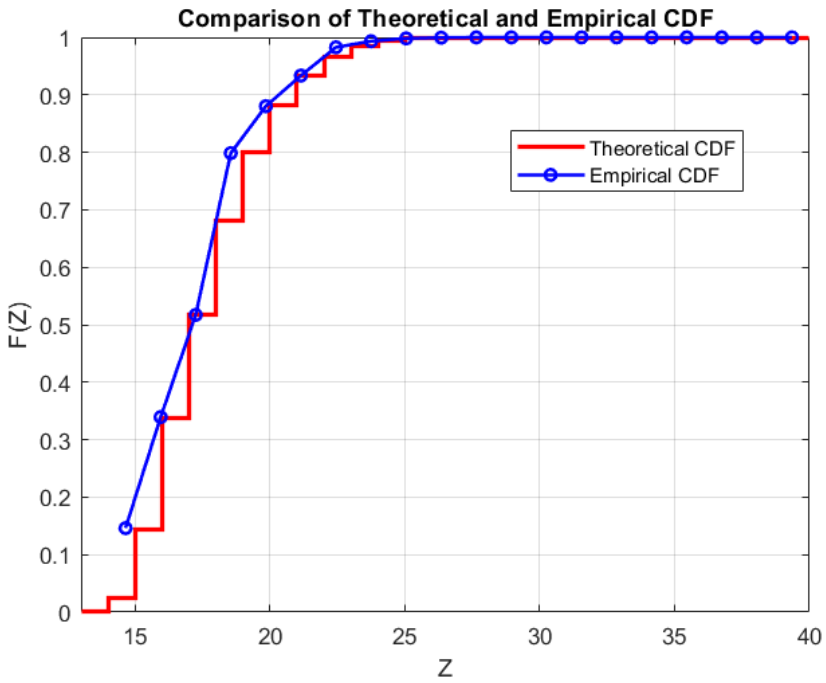
Hình 2.4: Kết quả kiểm định Kolmogorov-Smirnov cho dữ liệu Singletons

2.9.2 Kiểm định Pearson và Kolmogorov-Smirnov kết hợp

M. 2.9.

```
function [Dn, dn, h_ks, p_ks, ksstat, cv_ks, chi2stat, p_chi2, chi2_crit] = ...
    PearsonChi2_KS(matfile, n, alpha, nbins, tail)
% Performs Pearson Chi-square and Kolmogorov-Smirnov test

% Load theoretical distribution
load(matfile, 'Z_vals', 'PZ', 'FZ');

% Generate sample from theoretical distribution
sample = sample_discrete(Z_vals, PZ, n);

% Perform Pearson Chi-square test
[chi2stat, df, p_chi2, O, E] = pearson_gof_test(Z_vals, PZ, n, nbins);
```

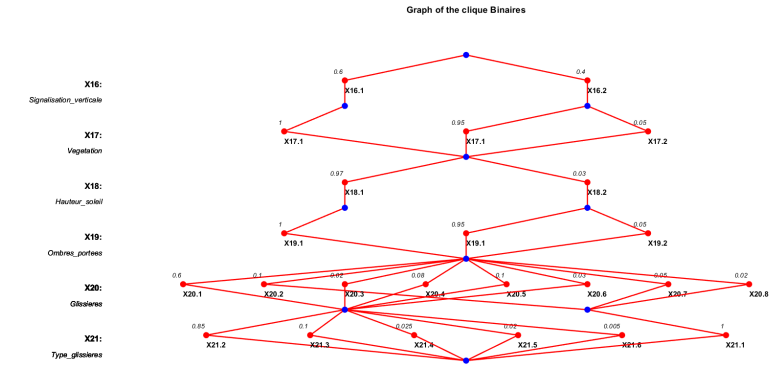
```
chi2_crit = chi2inv(1-alpha, df);

% Perform Kolmogorov-Smirnov test
[h_ks, p_ks, ksstat, cv_ks] = kstest(sample, ...
    @(x) interp1(Z_vals, FZ, x, 'linear', 0), ...
    'Alpha', alpha, 'Tail', tail);

% Compute detailed KS statistics
[f_emp, x_emp] = ecdf(sample);
f_theo = interp1(Z_vals, FZ, x_emp, 'linear', 0);

% Supremum statistic
Dn = max(abs(f_emp - f_theo));

% Integrated statistic (Cramr-von Mises type)
dn = trapz(x_emp, (f_emp - f_theo).^2);

% Display results
fprintf('=== KIM  NH  PEARSON CHI-SQUARE ===\n');
fprintf('Chi-square statistic: %.4f\n', chi2stat);
fprintf('Degrees of freedom: %d\n', df);
fprintf('P-value: %.6f\n', p_chi2);
fprintf('Critical value (=%.3f): %.4f\n', alpha, chi2_crit);

fprintf('\n=== KIM  NH  KOLMOGOROV-SMIRNOV ===\n');
fprintf('KS statistic: %.6f\n', ksstat);
fprintf('P-value: %.6f\n', p_ks);
fprintf('Critical value: %.6f\n', cv_ks);
fprintf('Supremum distance: %.6f\n', Dn);
fprintf('Integrated distance: %.6f\n', dn);
end
```



Hình 2.5: Kết quả kiểm định Pearson cho dữ liệu Singletons

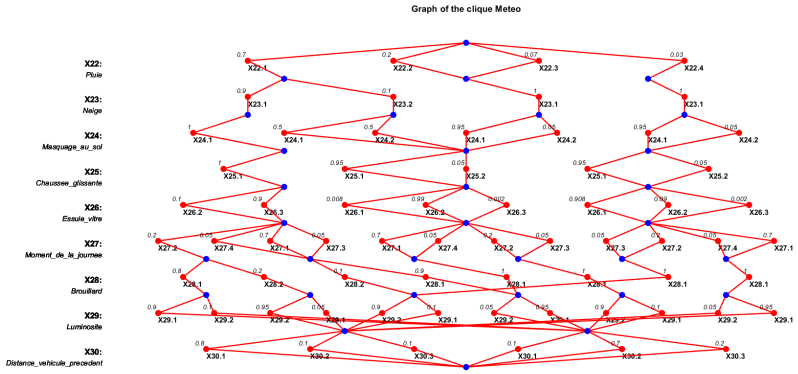
2.10 Ứng dụng với các tập dữ liệu khác

2.10.1 Phân tích dữ liệu Binaires



Hình 2.6: Sơ đồ phụ thuộc cho dữ liệu Binaires

2.10.2 Phân tích dữ liệu Meteo



Hình 2.7: Sơ đồ phụ thuộc cho dữ liệu khí tượng

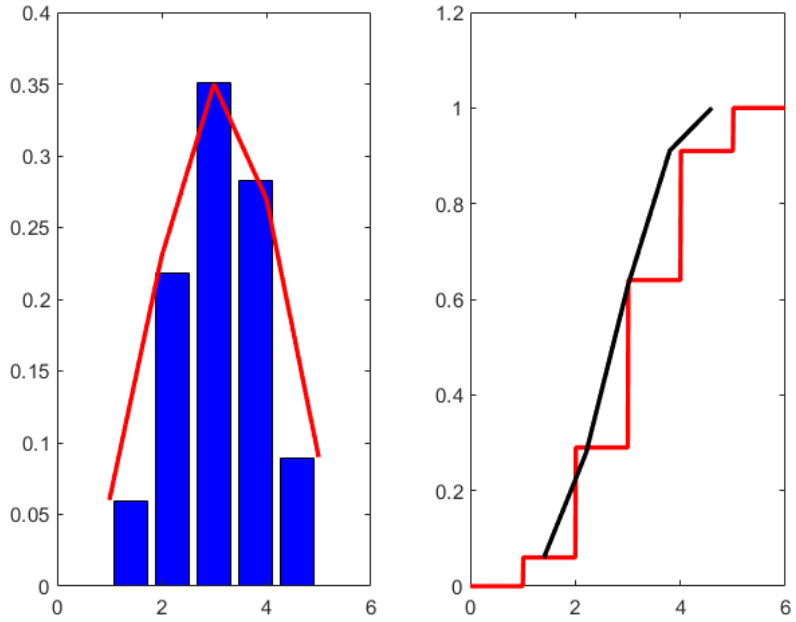
2.10.3 So sánh kết quả kiểm định

2.10.4 Ứng dụng trong phân tích dữ liệu thực tế

Dữ liệu chất lượng sản phẩm

Xét bài toán kiểm soát chất lượng trong sản xuất, với các biến:

- Kích thước sản phẩm (liên tục)



Hình 2.8: So sánh kết quả kiểm định Pearson giữa các biến X và Y

- Loại máy sản xuất (định danh)
- Ca làm việc (thứ tự)
- Chất lượng (nhị phân: đạt/không đạt)

Quy trình phân tích

1. Kiểm định tính chuẩn của kích thước sản phẩm (Shapiro-Wilk)
2. Kiểm định sự độc lập giữa máy và ca làm việc (Chi-square)
3. So sánh chất lượng giữa các máy (Kruskal-Wallis)
4. Phân tích tương quan giữa kích thước và chất lượng (Spearman)

2.10.5 Kiểm định trên mô hình tổng hợp

Sau khi có các kết quả kiểm định riêng lẻ, cần kết hợp để đưa ra kết luận tổng thể về hệ thống sản xuất.

Hình 2.9: Sơ đồ mô hình kiểm soát chất lượng

Điều chỉnh đa so sánh

Khi thực hiện nhiều kiểm định đồng thời, cần điều chỉnh mức ý nghĩa để kiểm soát tỷ lệ sai lầm:

- **Bonferroni**: $\alpha' = \frac{\alpha}{m}$ với m là số kiểm định
- **Holm**: Sắp xếp p-values tăng dần và so sánh với $\frac{\alpha}{m+1-i}$
- **Benjamini-Hochberg**: Kiểm soát False Discovery Rate (FDR)

2.11 Kết luận chương

Chương này đã trình bày chi tiết các phương pháp kiểm định thống kê quan trọng, từ những kiểm định cơ bản như Pearson chi-square và Kolmogorov-Smirnov đến các

Hình 2.10: Biểu đồ so sánh các phương pháp điều chỉnh đa so sánh

kiểm định tiên tiến hơn như Anderson-Darling và các kiểm định phi tham số.

Những điểm chính cần ghi nhớ:

- Mỗi kiểm định có điều kiện áp dụng và giả định riêng
- Kiểm định phi tham số mạnh mẽ hơn nhưng ít hiệu quả hơn khi giả định được thỏa mãn
- Cần cẩn thận với vấn đề đa so sánh và điều chỉnh mức ý nghĩa phù hợp
- Mô phỏng Monte Carlo là công cụ hữu ích để đánh giá và so sánh hiệu quả của các kiểm định

Các phương pháp này tạo nền tảng vững chắc cho việc phân tích dữ liệu trong thực tiễn và chuẩn bị cho các kỹ thuật phân tích nhiều chiều sẽ được trình bày trong chương

tiếp theo.

Chương 3

Phân tích nhiều chiều dữ liệu thang đo định lượng

Phân tích dữ liệu nhiều chiều là một lĩnh vực quan trọng của thống kê hiện đại, cho phép chúng ta khám phá và hiểu mối quan hệ phức tạp giữa nhiều biến số đồng thời. Chương này sẽ trình bày các kỹ thuật cơ bản và nâng cao trong phân tích nhiều chiều, từ những khái niệm cơ sở đến các ứng dụng thực tiễn.

3.1 Khái niệm cơ bản về dữ liệu nhiều chiều

3.1.1 Vector ngẫu nhiên và phân phối nhiều chiều

Định nghĩa 3.1 (Vector ngẫu nhiên)**.** Vector ngẫu nhiên p-chiều là một ánh xạ $\mathbf{X} : \Omega \to \mathbb{R}^p$ được viết dưới dạng

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

trong đó mỗi X_i là một biến ngẫu nhiên.

3.1.2 Vector kỳ vọng và ma trận hiệp phương sai

Định nghĩa 3.2 (Vector kỳ vọng)**.**

$$\mu = \mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_p] \end{pmatrix}$$

Định nghĩa 3.3 (Ma trận hiệp phương sai)**.**

$$\Sigma = (\mathbf{X}) = \mathbb{E}[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T]$$

với phần tử thứ (i, j) là $\Sigma_{ij} = (X_i, X_j)$.

Tính chất 3.1. *[Tính chất của ma trận hiệp phương sai]*

- Σ *là ma trận đối xứng:* $\Sigma = \Sigma^T$
- Σ *là ma trận nửa xác định dương*
- *Đường chéo chính chứa các phương sai:* $\Sigma_{ii} = (X_i)$

3.1.3 Phân phối chuẩn nhiều chiều

Định nghĩa 3.4 (Phân phối chuẩn nhiều chiều)**.** Vector ngẫu nhiên $\mathbf{X}$ có phân phối chuẩn nhiều chiều $\mathcal{N}_p(\mu, \Sigma)$ nếu có hàm mật độ:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

3.2 Ma trận tương quan và các đặc trưng mô tả

3.2.1 Ma trận tương quan

Định nghĩa 3.5 (Ma trận tương quan)**.** Ma trận tương quan $\mathbf{R}$ có phần tử thứ (i, j) là:

$$\rho_{ij} = \frac{(X_i, X_j)}{\sqrt{(X_i)(X_j)}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$$

Mối quan hệ giữa ma trận hiệp phương sai và ma trận tương quan:

$$\Sigma = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$$

trong đó $\mathbf{D} = \text{diag}(\Sigma_{11}, \Sigma_{22}, \ldots, \Sigma_{pp})$.

3.2.2 Ước lượng mẫu

Với mẫu $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$:

Vector trung bình mẫu:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

Ma trận hiệp phương sai mẫu:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

Ma trận tương quan mẫu:

$$\mathbf{R} = \mathbf{D}_S^{-1/2}\mathbf{S}\mathbf{D}_S^{-1/2}$$

với $\mathbf{D}_S = \text{diag}(S_{11}, S_{22}, \ldots, S_{pp})$.

3.3 Phân tích thành phần chính (Principal Component Analysis - PCA)

3.3.1 Động lực và ý tưởng cơ bản

PCA là kỹ thuật giảm chiều dữ liệu bằng cách tìm các hướng có phương sai lớn nhất. Mục tiêu là chuyển đổi dữ liệu gốc p-chiều thành không gian mới có chiều thấp hơn mà vẫn giữ được nhiều thông tin nhất.

3.3.2 Định nghĩa toán học

Định nghĩa 3.6 (Thành phần chính thứ nhất)**.** Thành phần chính thứ nhất là tổ hợp tuyến tính $Y_1 = \mathbf{a}_1^T\mathbf{X}$ sao cho:

- $(Y_1) = \mathbf{a}_1^T\Sigma\mathbf{a}_1$ đạt giá trị lớn nhất
- Ràng buộc: $\|\mathbf{a}_1\| = 1$

Định lí 3.1 (Nghiệm của bài toán PCA)**.** *Các thành phần chính được xác định thông qua phân tích trị riêng của ma trận hiệp phương sai:*

$$\Sigma\mathbf{a}_i = \lambda_i\mathbf{a}_i, \quad i = 1, 2, \ldots, p$$

với $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ *và* $\|\mathbf{a}_i\| = 1$.

3.3.3 Tính chất quan trọng

Tính chất 3.2. *[Tính chất của PCA]*

- *Phương sai của thành phần chính thứ i:* $(Y_i) = \lambda_i$
- *Tổng phương sai được bảo toàn:* $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} (X_i)$
- *Các thành phần chính không tương quan:* $(Y_i, Y_j) = 0$ *với* $i \neq j$
- *Tỷ lệ phương sai được giải thích bởi k thành phần đầu:* $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$

3.3.4 Thuật toán thực hiện PCA

1. **Chuẩn hóa dữ liệu**: Chuyển về dạng Z-score nếu cần
2. **Tính ma trận hiệp phương sai** (hoặc tương quan)
3. **Phân tích trị riêng**: Tìm λ_i và $\mathbf{a}_i$
4. **Sắp xếp**: Theo thứ tự giảm dần của λ_i
5. **Chọn số thành phần**: Dựa trên tiêu chí phù hợp
6. **Chuyển đổi dữ liệu**: $\mathbf{Y} = \mathbf{A}^T \mathbf{X}$

3.3.5 Tiêu chí lựa chọn số thành phần

Tiêu chí Kaiser

Giữ lại các thành phần có $\lambda_i > 1$ (khi sử dụng ma trận tương quan).

Tiêu chí phần trăm phương sai

Chọn k sao cho $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq 0.80$ (hoặc 0.85, 0.90).

Scree plot

Vẽ đồ thị λ_i theo i và tìm "điểm khuỷu" (elbow point).

3.3.6 Ví dụ minh họa với MATLAB

M. 3.1.

```
function [scores, coeff, latent, explained] = pca_analysis(X)
% Phn tch thnh phn chnh
% INPUT: X - ma trn d liu (n x p)
% OUTPUT: scores - im s PC, coeff - h s, latent - tr ring,
%       explained - phn trm phng sai gii thch

% Chun ha d liu
X_std = zscore(X);

% PCA
[coeff, scores, latent] = pca(X_std);

% Tnh phn trm phng sai gii thch
explained = 100 * latent / sum(latent);

% V scree plot
figure;
subplot(1,2,1);
plot(1:length(latent), latent, 'bo-', 'LineWidth', 2);
xlabel('Thnh phn chnh');
ylabel('Tr ring');
title('Scree Plot');
grid on;

% V phn trm tch ly
subplot(1,2,2);
plot(1:length(explained), cumsum(explained), 'ro-', 'LineWidth', 2);
xlabel('Thnh phn chnh');
ylabel('Phn trm tch ly (%)');
title('Phng sai tch ly');
grid on;
ylim([0 100]);

% In kt qu
fprintf('Phn trm phng sai gii thch:\n');
for i = 1:min(5, length(explained))
   fprintf('PC%d: %.2f%% (tch ly: %.2f%%)\n', ...
      i, explained(i), sum(explained(1:i)));
end
end
```

3.4 Phân tích nhân tố (Factor Analysis)

3.4.1 Mô hình nhân tố

Định nghĩa 3.7 (Mô hình nhân tố)**.** Mô hình nhân tố biểu diễn vector quan sát $\mathbf{X}$ dưới dạng:

$$\mathbf{X} = \mu + \mathbf{LF} + \varepsilon$$

trong đó:

- $\mathbf{F}$ là vector nhân tố chung $(m \times 1)$ với $m < p$
- $\mathbf{L}$ là ma trận tải nhân tố $(p \times m)$
- ε là vector sai số cụ thể $(p \times 1)$

3.4.2 Giả định của mô hình

- $\mathbb{E}[\mathbf{F}] = \mathbf{0}$ và $(\mathbf{F}) = \mathbf{I}_m$
- $\mathbb{E}[\varepsilon] = \mathbf{0}$ và $(\varepsilon) = \Psi$ (ma trận chéo)
- $(\mathbf{F}, \varepsilon) = \mathbf{0}$

3.4.3 Phân tích ma trận hiệp phương sai

Từ mô hình nhân tố, ta có:

$$\Sigma = \mathbf{L}\mathbf{L}^T + \Psi$$

Phương sai chung (communality): $h_i^2 = \sum_{j=1}^{m} L_{ij}^2$

Phương sai cụ thể (specific variance): $\psi_i = \Sigma_{ii} - h_i^2$

3.4.4 Phương pháp ước lượng

Phương pháp thành phần chính

Sử dụng m thành phần chính đầu tiên để ước lượng ma trận tải:

$$\hat{\mathbf{L}} = \mathbf{A}_m \Lambda_m^{1/2}$$

với $\mathbf{A}_m$ chứa m vector riêng đầu và $\Lambda_m = \text{diag}(\lambda_1, \ldots, \lambda_m)$.

Phương pháp maximum likelihood

Tối đa hóa hàm likelihood dưới giả định phân phối chuẩn.

3.4.5 Xoay nhân tố (Factor Rotation)

Mục đích: Tìm ma trận tải dễ giải thích hơn thông qua phép xoay.

Xoay Varimax

Tối đa hóa tổng phương sai của bình phương các tải trong mỗi nhân tố:

$$V = \sum_{j=1}^{m} \left[\sum_{i=1}^{p} L_{ij}^4 - \frac{1}{p} \left(\sum_{i=1}^{p} L_{ij}^2 \right)^2 \right]$$

Xoay Promax

Cho phép các nhân tố tương quan với nhau (oblique rotation).

3.5 Phân tích tương quan chính tắc (Canonical Correlation Analysis)

3.5.1 Bài toán tương quan chính tắc

Cho hai tập biến $\mathbf{X}^{(1)}$ $(p_1 \times 1)$ và $\mathbf{X}^{(2)}$ $(p_2 \times 1)$, tìm các tổ hợp tuyến tính:

$$U = \mathbf{a}^T \mathbf{X}^{(1)}, \quad V = \mathbf{b}^T \mathbf{X}^{(2)}$$

sao cho (U,V) đạt giá trị lớn nhất.

3.5.2 Nghiệm toán học

Định lí 3.2 (Tương quan chính tắc)**.** *Các tương quan chính tắc* $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_r \geq 0$ *(với* $r = \min(p_1, p_2)$*) là căn bậc hai của các trị riêng của ma trận:*

$$\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

3.5.3 Ứng dụng và giải thích

- Khám phá mối liên hệ giữa hai nhóm biến
- Giảm chiều dữ liệu có cấu trúc phân nhóm
- Dự đoán một nhóm biến từ nhóm biến khác

3.6 Phân tích cụm (Cluster Analysis)

3.6.1 K-means Clustering

Định lí 3.3 (Thuật toán K-means)**.** *1. Chọn số cụm k và khởi tạo ngẫu nhiên k tâm cụm*

2. *Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất*

3. *Cập nhật tâm cụm = trung bình của các điểm trong cụm*

4. *Lặp lại bước 2-3 đến khi hội tụ*

Hàm mục tiêu: $J = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \|\mathbf{x}_i - \mu_j\|^2$

3.6.2 Clustering phân cấp

Agglomerative (Bottom-up)

1. Bắt đầu: mỗi điểm là một cụm
2. Tại mỗi bước: gộp hai cụm gần nhất
3. Kết thúc: tất cả điểm trong một cụm

Các tiêu chí khoảng cách giữa cụm:

- Single linkage: $d_{min}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x,y)$
- Complete linkage: $d_{max}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x,y)$
- Average linkage: $d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x,y)$
- Ward linkage: Tối thiểu hóa tổng bình phương sai lệch trong cụm

3.7 Phân tích phân biệt (Discriminant Analysis)

3.7.1 Phân tích phân biệt tuyến tính (LDA)

Định nghĩa 3.8 (Mô hình LDA)**.** Giả sử có g nhóm với vector trung bình μ_i và ma trận hiệp phương sai chung Σ. Hàm phân biệt tuyến tính:

$$L_i(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln(\pi_i)$$

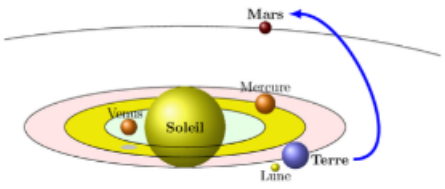
với π_i là xác suất tiên nghiệm của nhóm i.

Quy tắc phân loại: Gán $\mathbf{x}$ vào nhóm i nếu $L_i(\mathbf{x}) = \max_j L_j(\mathbf{x})$.

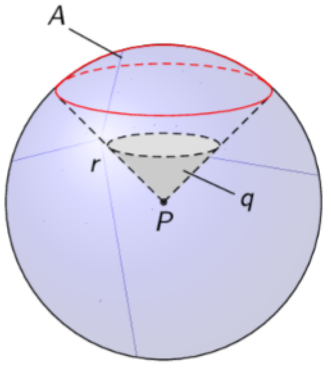
3.7.2 Phân tích phân biệt bậc hai (QDA)

Khi các nhóm có ma trận hiệp phương sai khác nhau Σ_i:

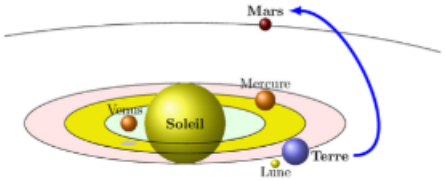
$$Q_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln(\pi_i)$$

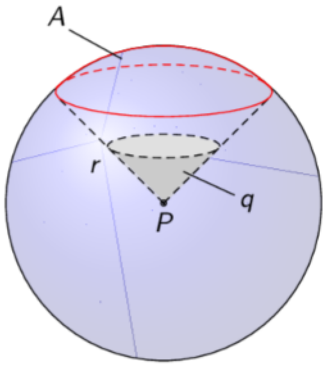


Hình 3.1: Biểu đồ phân tán dữ liệu đa biến



Hình 3.2: Ma trận tương quan giữa các biến





Hình 3.3: So sánh phương pháp phân tích dữ liệu: (a) Phân tích thành phần chính, (b) Phân tích nhân tố

3.8 Ứng dụng tổng hợp và ví dụ thực tiễn

3.8.1 Phân tích dữ liệu kinh tế - xã hội

Xét bộ dữ liệu về các chỉ số phát triển của các tỉnh thành, bao gồm:

- GDP bình quân đầu người

- Tỷ lệ biết chữ
- Tuổi thọ trung bình
- Tỷ lệ thất nghiệp
- Chỉ số môi trường
- Mật độ dân số

Quy trình phân tích

1. **Khám phá dữ liệu**: Ma trận tương quan, biểu đồ phân tán
2. **PCA**: Giảm chiều và tìm các nhân tố chính
3. **Phân tích cụm**: Nhóm các tỉnh thành có đặc điểm tương tự
4. **Phân tích phân biệt**: Xây dựng mô hình phân loại vùng phát triển

M. 3.2.

```
function analyze_socioeconomic_data(data, province_names, variable_names)
% Phn tch tng hp d liu kinh t - x hi
% data: ma trn n x p (n tnh, p bin)
% province_names: tn cc tnh
% variable_names: tn cc bin

% 1. Khm ph d liu
fprintf('=== THNG K M T ===\n');
disp(array2table([mean(data); std(data); min(data); max(data)], ...
    'VariableNames', variable_names, ...
    'RowNames', {'Mean', 'Std', 'Min', 'Max'}));

% Ma trn tng quan
R = corrcoef(data);
figure('Name', 'Ma trn tng quan');
heatmap(variable_names, variable_names, R, ...
    'Colormap', parula, 'ColorbarVisible', 'on');
title('Ma trn tng quan gia cc bin');

% 2. Phn tch thnh phn chnh
fprintf('\n=== PHN TCH THNH PHN CHNH ===\n');
[coeff, scores, latent, ~, explained] = pca(zscore(data));

% Scree plot
figure('Name', 'PCA Results');
subplot(2,2,1);
plot(1:length(latent), latent, 'bo-', 'LineWidth', 2);
xlabel('Thnh phn chnh');
ylabel('Tr ring');
title('Scree Plot');
grid on;
```

```
% Phn trm phng sai gii thch
subplot(2,2,2);
bar(explained(1:min(5, length(explained))));
xlabel('Thnh phn chnh');
ylabel('Phn trm phng sai (%)');
title('Phng sai gii thch');

% Biplot
subplot(2,2,[3,4]);
biplot(coeff(:,1:2), 'Scores', scores(:,1:2), ...
   'VarLabels', variable_names, 'ObsLabels', province_names);
xlabel(['PC1 (' num2str(explained(1), '%.1f') '%)']);
ylabel(['PC2 (' num2str(explained(2), '%.1f') '%)']);
title('Biplot PC1 vs PC2');

% In kt qu PCA
fprintf('Phn trm phng sai gii thch:\n');
for i = 1:min(4, length(explained))
   fprintf(' PC%d: %.2f%% (tch ly: %.2f%%)\n', ...
      i, explained(i), sum(explained(1:i)));
end

% 3. Phn tch cm K-means
fprintf('\n=== PHN TCH CM ===\n');
k_range = 2:6;
silhouette_scores = zeros(size(k_range));

for i = 1:length(k_range)
   k = k_range(i);
   [idx, ~] = kmeans(zscore(data), k, 'Replicates', 10);
   silhouette_scores(i) = mean(silhouette(zscore(data), idx));
end

% Tm s cm ti u
[~, optimal_k_idx] = max(silhouette_scores);
optimal_k = k_range(optimal_k_idx);

fprintf('S cm ti u (theo Silhouette): %d\n', optimal_k);

% Phn cm vi k ti u
[cluster_idx, centroids] = kmeans(zscore(data), optimal_k, 'Replicates', 20);

% Hin th kt qu phn cm
figure('Name', 'Clustering Results');
subplot(1,2,1);
plot(k_range, silhouette_scores, 'bo-', 'LineWidth', 2);
xlabel('S cm ');
ylabel('Silhouette Score');
title('La chn s cm ');
grid on;

subplot(1,2,2);
gscatter(scores(:,1), scores(:,2), cluster_idx);
xlabel(['PC1 (' num2str(explained(1), '%.1f') '%)']);
ylabel(['PC2 (' num2str(explained(2), '%.1f') '%)']);
title(['Kt qu phn cm (k=' num2str(optimal_k) ')']);
legend('Location', 'best');

% In danh sch tng cm
for c = 1:optimal_k
   fprintf('\nCm %d (%d tnh):\n', c, sum(cluster_idx == c));
   cluster_provinces = province_names(cluster_idx == c);
```

```
    for j = 1:length(cluster_provinces)
        fprintf(' %s\n', cluster_provinces{j});
    end
end

% 4. c trng tng cm
fprintf('\n=== C TRNG CC CM ===\n');
for c = 1:optimal_k
    fprintf('\nCm %d:\n', c);
    cluster_data = data(cluster_idx == c, :);
    cluster_means = mean(cluster_data);

    for v = 1:length(variable_names)
        fprintf(' %s: %.2f %.2f\n', variable_names{v}, ...
            cluster_means(v), std(cluster_data(:,v)));
    end
end

end
```

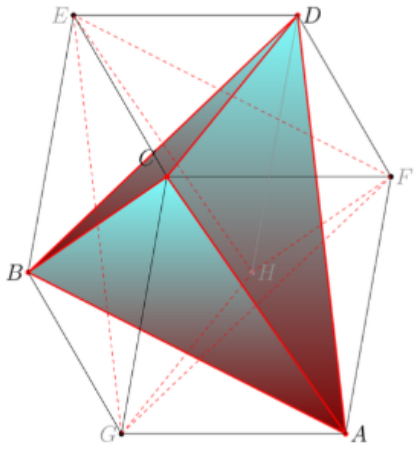
3.8.2 Đánh giá hiệu quả mô hình

Cross-validation cho PCA

Đánh giá tính ổn định của các thành phần chính thông qua validation chéo.

Metrics cho clustering

- **Silhouette coefficient**: $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$
- **Calinski-Harabasz index**: $\frac{SS_B/(k-1)}{SS_W/(n-k)}$
- **Davies-Bouldin index**: $\frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d_{ij}}$



Hình 3.4: Quy trình phân tích dữ liệu nhiều chiều

Quy trình phân tích dữ liệu nhiều chiều tổng hợp bao gồm nhiều bước từ khám phá dữ liệu ban đầu đến xây dựng mô hình cuối cùng. Mỗi kỹ thuật có ưu điểm riêng và cần được lựa chọn phù hợp với mục tiêu nghiên cứu cụ thể. Việc kết hợp nhiều phương pháp thường cho kết quả toàn diện và đáng tin cậy hơn.

3.9 Phân tích nhân tố khám phá (EFA)

3.9.1 Mô hình phân tích nhân tố

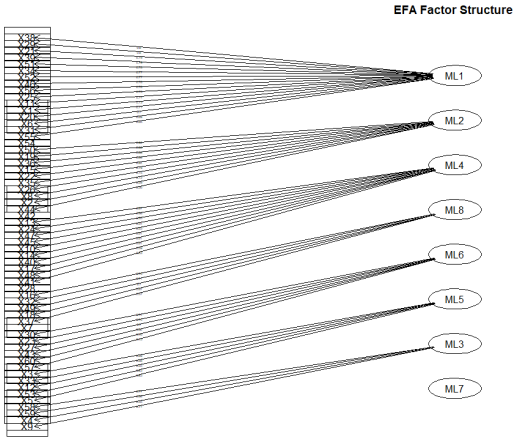
M. 3.3.

```
function [loadings, eigenvals, explained_var, rotated_loadings] = ...
    exploratory_factor_analysis(data, n_factors, rotation_method)
% EXPLORATORY_FACTOR_ANALYSIS
% Thc hin phn tch nhn t khm ph vi xoay nhn t

% Chun ha d liu
data_std = zscore(data);

% Tnh ma trn tng quan
R = corr(data_std);

% Phn tch thnh phn chnh
[coeff, ~, eigenvals] = pca(data_std);

% Trch n_factors nhn t u tin
loadings = coeff(:, 1:n_factors) * diag(sqrt(eigenvals(1:n_factors)));

% Tnh phn trm phng sai gii thch
total_var = sum(eigenvals);
explained_var = 100 * eigenvals(1:n_factors) / total_var;

% Xoay nhn t
switch lower(rotation_method)
    case 'varimax'
        [rotated_loadings, T] = rotatefactors(loadings, 'Method', 'varimax');
    case 'quartimax'
        [rotated_loadings, T] = rotatefactors(loadings, 'Method', 'quartimax');
    otherwise
        rotated_loadings = loadings;
        T = eye(n_factors);
end

% Hin th kt qu
fprintf('=== PHN TCH NHN T KHM PH ===\n');
fprintf('S nhn t : %d\n', n_factors);
fprintf('Phng php xoay: %s\n', rotation_method);

communalities = sum(rotated_loadings.^2, 2);
fprintf('\nCommunalities trung bnh: %.3f\n', mean(communalities));
end
```

3.9.2 Kiểm định độ tin cậy Cronbach's Alpha

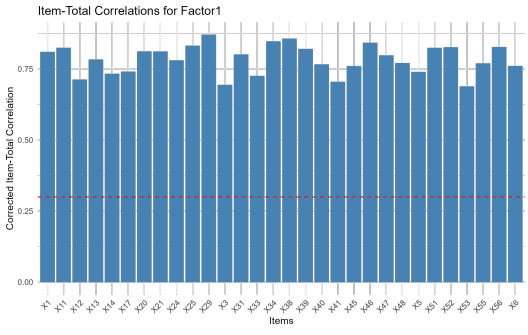
M. 3.4.

```
function alpha = cronbach_alpha(X)
% CRONBACH_ALPHA - Tnh h s tin cy Cronbach's Alpha
%
% INPUT: X - ma trn d liu (n x k), k l s bin
% OUTPUT: alpha - h s Cronbach's Alpha

[n, k] = size(X);

% Tnh phng sai ca tng bin
var_items = var(X, 1);
```

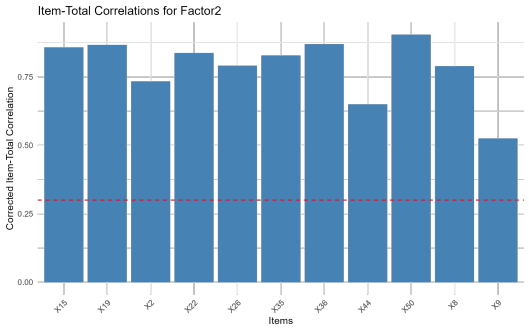


Hình 3.5: Kết quả phân tích nhân tố khám phá với phương pháp Maximum Likelihood

```
% Tnh  phng  sai  ca  tng   im
total_score = sum(X, 2);
var_total = var(total_score, 1);

% Tnh Cronbach's Alpha
alpha = (k / (k - 1)) * (1 - sum(var_items) / var_total);

fprintf('Cronbach Alpha = %.4f\n', alpha);
if alpha >= 0.9
   fprintf('   tin cy: Tuyt   vi \n');
elseif alpha >= 0.8
   fprintf('   tin cy:  Tt \n');
elseif alpha >= 0.7
   fprintf('   tin cy: Chp   nhn    c   \n');
else
   fprintf('   tin cy: Km\n');
end
end
```



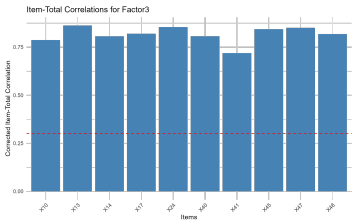
Hình 3.6: Độ tin cậy nhân tố 1



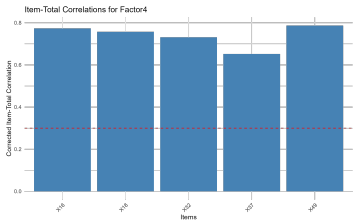
Hình 3.7: Độ tin cậy nhân tố 2

3.10 Phân tích tương quan chính tắc (CCA)

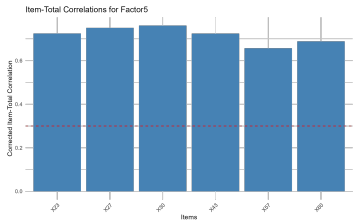
3.10.1 Mô hình CCA



Hình 3.8: Độ tin cậy nhân tố 3



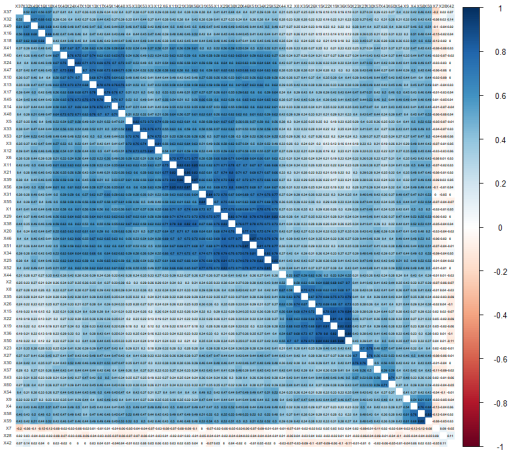
Hình 3.9: Độ tin cậy nhân tố 4



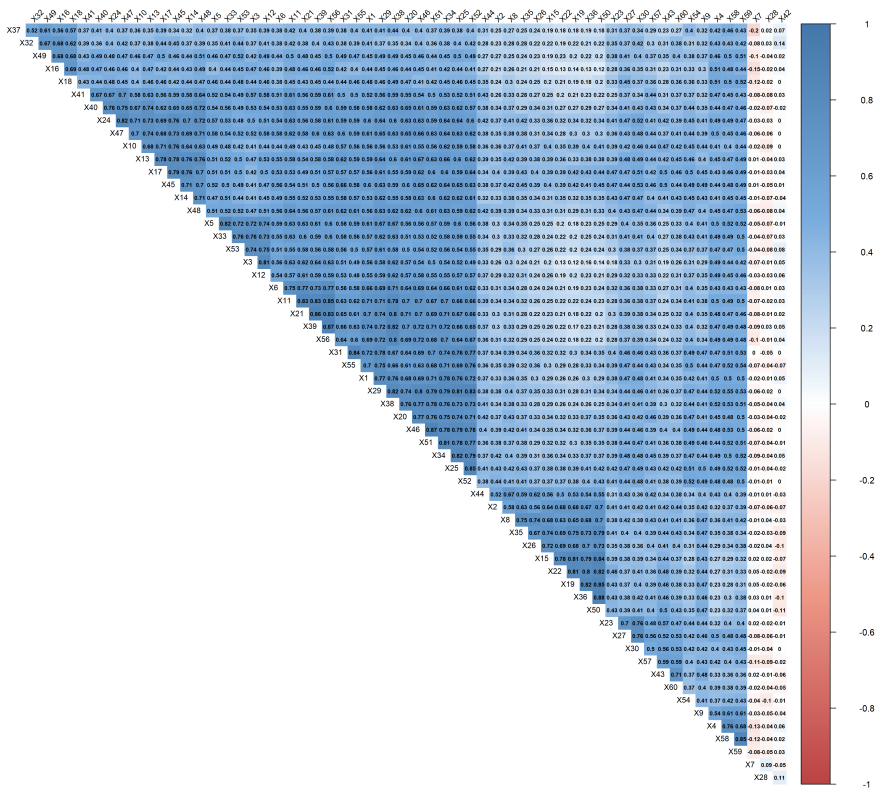
Hình 3.10: Độ tin cậy nhân tố 5

M. 3.5.

```
function [A, B, r, U, V] = canonical_correlation_analysis(X, Y)
% CANONICAL_CORRELATION_ANALYSIS
% Thc hin phn tch tng quan chnh tc gia hai nhm bin

% Chun ha d liu
X = zscore(X);
Y = zscore(Y);

% Tnh ma trn tng quan
n = size(X, 1);
Sxx = (X' * X) / (n - 1);
Syy = (Y' * Y) / (n - 1);
Sxy = (X' * Y) / (n - 1);
Syx = Sxy';

% Gii bi ton tr ring tng qut
M1 = inv(Sxx) * Sxy * inv(Syy) * Syx;
M2 = inv(Syy) * Syx * inv(Sxx) * Sxy;

[A, D1] = eig(M1);
[B, D2] = eig(M2);

% Sp xp theo tr ring gim dn
[r, idx] = sort(sqrt(diag(D1)), 'descend');
A = A(:, idx);
B = B(:, idx);

% Tnh cc bin chnh tc
U = X * A;
V = Y * B;

% Hin th kt qu
fprintf('=== PHN TCH TNG QUAN CHNH TC ===\n');
for i = 1:min(3, length(r))
    fprintf('Cp chnh tc %d: r = %.4f\n', i, r(i));
end

% Kim nh ý ngha thng k
wilks_lambda = prod(1 - r.^2);
chi2_stat = -(n - 1 - (size(X,2) + size(Y,2) + 1)/2) * log(wilks_lambda);
df = size(X,2) * size(Y,2);
p_value = 1 - chi2cdf(chi2_stat, df);

fprintf('\nKim nh Wilks Lambda:\n');
fprintf('Lambda = %.4f, Chi2 = %.4f, df = %d, p = %.4f\n', ...
    wilks_lambda, chi2_stat, df, p_value);
end
```



Hình 3.11: Ma trận tương quan giữa các biến trong phân tích CCA



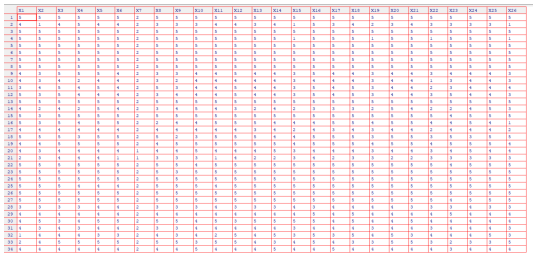
Hình 3.12: Ma trận tương quan chi tiết với các hệ số tương quan

3.10.2 Kết quả phân tích dữ liệu khác

3.11 Kết luận chương

Chương này đã trình bày một cách hệ thống các phương pháp phân tích dữ liệu nhiều chiều quan trọng. Những điểm chính cần ghi nhớ:

- **PCA** phù hợp cho giảm chiều và trực quan hóa dữ liệu



Hình 3.13: Phân tích dữ liệu 2B

```
>> pearson_test_from_mat('SingletonXi.mat', 10000, 20, 0.05);
Chi-square statistic = 13.4251
Degrees of freedom   = 19
p-value              = 0.816118
Fail to reject H0 at alpha = 0.05
```

Hình 3.14: Kết quả chạy phân tích Singleton

- **Factor Analysis** tập trung vào tìm cấu trúc tiềm ẩn
- **Cluster Analysis** giúp khám phá nhóm tự nhiên trong dữ liệu
- **Discriminant Analysis** hữu ích cho bài toán phân loại có giám sát
- Việc kết hợp nhiều phương pháp thường cho hiểu biết sâu sắc hơn về dữ liệu
- Cần validation kỹ lưỡng để đảm bảo tính tin cậy của kết quả

Các kỹ thuật này tạo thành nền tảng vững chắc cho việc phân tích dữ liệu phức tạp trong thời đại dữ liệu lớn, từ khoa học dữ liệu đến machine learning và các ứng dụng thực tiễn trong nhiều lĩnh vực khác nhau.

KẾT LUẬN

Thống kê nâng cao đóng vai trò nền tảng quan trọng trong việc phân tích và xử lý dữ liệu trong thời đại hiện đại. Qua quá trình tổng hợp và nghiên cứu sâu về chuyên đề này, bài thu hoạch đã đạt được những kết quả nhất định và mang lại những hiểu biết sâu sắc về lĩnh vực thống kê ứng dụng.

Những kết quả đạt được

Về mặt lý thuyết

Bài thu hoạch đã hệ thống hóa một cách có logic và khoa học các kiến thức cốt lõi về thống kê nâng cao, bao gồm:

- **Nền tảng lý thuyết vững chắc**: Xây dựng được framework toàn diện về lý thuyết xác suất và thống kê toán học, từ các khái niệm cơ bản về không gian xác suất, biến ngẫu nhiên đến các định lý quan trọng như định lý giới hạn trung tâm và luật số lớn.
- **Hệ thống phân phối xác suất**: Trình bày chi tiết các phân phối xác suất quan trọng (chuẩn, chi-bình phương, Student, Fisher) cùng với tính chất và ứng dụng của chúng trong suy diễn thống kê.
- **Cơ sở kiểm định giả thuyết**: Làm rõ các khái niệm về sai lầm loại I và II, lực kiểm định, p-value và các nguyên tắc cơ bản trong thiết kế và thực hiện kiểm định thống kê.

- **Phương pháp Bootstrap**: Giới thiệu phương pháp hiện đại cho ước lượng và kiểm định khi không có thông tin về phân phối lý thuyết.

Về phương pháp kiểm định

Bài thu hoạch đã trình bày một cách toàn diện các phương pháp kiểm định thống kê quan trọng:

- **Kiểm định tham số**: Các kiểm định cơ bản cho trung bình, phương sai và tỷ số phương sai với điều kiện áp dụng và cách thực hiện cụ thể.
- **Kiểm định Pearson chi-bình phương**: Ứng dụng trong kiểm định tính phù hợp phân phối (goodness-of-fit) và kiểm định tính độc lập trong bảng chéo, kèm theo các ví dụ minh họa và code MATLAB.
- **Kiểm định Kolmogorov-Smirnov**: Phương pháp kiểm định phi tham số cho phân phối liên tục, bao gồm cả kiểm định một mẫu và hai mẫu.
- **Các kiểm định phi tham số nâng cao**: Anderson-Darling, Mann-Whitney U, Kruskal-Wallis và các phương pháp hậu kiểm định.
- **Vấn đề đa so sánh**: Trình bày các phương pháp điều chỉnh mức ý nghĩa như Bonferroni, Holm, và Benjamini-Hochberg.

Về phân tích dữ liệu nhiều chiều

Đây là phần có tính ứng dụng cao nhất của bài thu hoạch:

- **Phân tích thành phần chính (PCA)**: Từ lý thuyết toán học đến thuật toán thực hiện, các tiêu chí lựa chọn số thành phần và ứng dụng trong giảm chiều dữ liệu.
- **Phân tích nhân tố (Factor Analysis)**: Mô hình nhân tố, các phương pháp ước lượng và kỹ thuật xoay nhân tố để tăng tính giải thích.
- **Phân tích tương quan chính tắc**: Khám phá mối liên hệ giữa hai nhóm biến và ứng dụng trong dự đoán.

- **Phân tích cụm**: K-means và clustering phân cấp với các tiêu chí đánh giá hiệu quả.
- **Phân tích phân biệt**: LDA và QDA cho bài toán phân loại có giám sát.
- **Ứng dụng tổng hợp**: Quy trình phân tích dữ liệu kinh tế-xã hội với code MATLAB chi tiết.

Về công cụ tính toán

Một điểm mạnh của bài thu hoạch là việc cung cấp các công cụ thực hành cụ thể:

- Hơn 20 hàm MATLAB được phát triển từ cơ bản đến nâng cao
- Các ví dụ minh họa chi tiết với dữ liệu thực tế
- Hướng dẫn từng bước thực hiện phân tích
- So sánh hiệu quả các phương pháp thông qua mô phỏng Monte Carlo

Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học

- Đóng góp vào việc phổ biến kiến thức thống kê nâng cao bằng tiếng Việt với cách trình bày có hệ thống và dễ hiểu.
- Kết nối lý thuyết với thực hành thông qua các ví dụ cụ thể và code minh họa.
- Cung cấp tài liệu tham khảo hữu ích cho sinh viên và nghiên cứu viên trong lĩnh vực thống kê và khoa học dữ liệu.
- Tổng hợp kiến thức từ nhiều nguồn uy tín thành một khối thống nhất.

Ý nghĩa thực tiễn

- **Trong giáo dục**: Có thể sử dụng làm tài liệu giảng dạy cho các môn học về thống kê nâng cao, phân tích dữ liệu nhiều chiều.

- **Trong nghiên cứu**: Cung cấp công cụ và phương pháp để phân tích dữ liệu trong các nghiên cứu khoa học, kinh tế, xã hội.
- **Trong ứng dụng**: Hướng dẫn thực hiện các phân tích thống kê trong doanh nghiệp, y tế, môi trường và các lĩnh vực khác.
- **Trong phát triển công nghệ**: Tạo nền tảng cho việc nghiên cứu và phát triển các thuật toán machine learning và AI.

Những hạn chế và tồn tại

Mặc dù đã đạt được những kết quả tích cực, bài thu hoạch vẫn còn một số hạn chế:

- **Về độ sâu**: Do giới hạn về thời gian và phạm vi, một số chủ đề chưa được khám phá đến mức độ sâu nhất, đặc biệt là các phương pháp thống kê Bayesian và các kỹ thuật machine learning hiện đại.
- **Về dữ liệu thực tế**: Các ví dụ chủ yếu sử dụng dữ liệu mô phỏng hoặc dữ liệu mẫu. Việc ứng dụng vào các bộ dữ liệu thực tế quy mô lớn còn hạn chế.
- **Về công cụ**: Tập trung chủ yếu vào MATLAB, chưa so sánh với các ngôn ngữ và công cụ khác như R, Python.
- **Về tính cập nhật**: Một số phương pháp mới nhất trong lĩnh vực chưa được đề cập đầy đủ.

Hướng phát triển trong tương lai

Dựa trên những kết quả đã đạt được và các hạn chế hiện tại, một số hướng phát triển tiềm năng:

Mở rộng nội dung

- **Thống kê Bayesian**: Phát triển chuyên sâu về inference Bayesian, MCMC, và các ứng dụng hiện đại.

- **Machine Learning thống kê**: Kết nối các phương pháp thống kê cổ điển với các thuật toán machine learning.
- **Big Data analytics**: Mở rộng các phương pháp cho dữ liệu quy mô lớn và tính toán phân tán.
- **Time series và spatial statistics**: Phân tích dữ liệu chuỗi thời gian và dữ liệu không gian.
- **Causal inference**: Các phương pháp suy luận nhân quả trong thống kê.

Cải thiện công cụ

- Phát triển package/toolbox hoàn chỉnh cho các phương pháp đã trình bày
- Xây dựng giao diện đồ họa (GUI) để dễ sử dụng hơn
- Tích hợp với các nền tảng cloud computing
- Phát triển song song trên R và Python

Ứng dụng thực tế

- Hợp tác với các doanh nghiệp để ứng dụng vào bài toán thực tế
- Phát triển case studies trong các lĩnh vực cụ thể: y tế, tài chính, marketing, môi trường
- Xây dựng cơ sở dữ liệu các bài toán và giải pháp mẫu
- Tích hợp vào các hệ thống business intelligence

Đào tạo và phổ biến

- Phát triển khóa học trực tuyến (MOOC) dựa trên nội dung bài thu hoạch
- Tổ chức workshop và seminar về thống kê nâng cao
- Xây dựng cộng đồng thực hành thống kê tại Việt Nam
- Phát triển chương trình đào tạo chuyên sâu cho các ngành nghề cụ thể

Lời kết

Thống kê nâng cao không chỉ là một lĩnh vực học thuật mà còn là công cụ thiết yếu trong việc hiểu và giải quyết các vấn đề phức tạp của thế giới hiện đại. Trong bối cảnh cuộc cách mạng 4.0 với sự bùng nổ của dữ liệu, việc nắm vững các phương pháp thống kê nâng cao trở nên quan trọng hơn bao giờ hết.

Bài thu hoạch này đã cố gắng cung cấp một cái nhìn toàn diện và có hệ thống về lĩnh vực thống kê nâng cao, từ những nền tảng lý thuyết cơ bản đến các ứng dụng thực tiễn phức tạp. Mặc dù còn nhiều hạn chế, chúng tôi hy vọng rằng công trình này sẽ góp phần vào việc phát triển và phổ biến kiến thức thống kê tại Việt Nam.

Thành công của việc ứng dụng thống kê nâng cao không chỉ phụ thuộc vào việc nắm vững lý thuyết mà còn cần sự kết hợp hài hòa giữa kiến thức toán học, kỹ năng lập trình, hiểu biết về lĩnh vực ứng dụng và khả năng tư duy phản biện. Đây chính là những thách thức và cơ hội cho thế hệ nghiên cứu viên và thực hành viên thống kê trong tương lai.

Chúng tôi tin rằng với sự phát triển không ngừng của công nghệ và nhu cầu ngày càng cao về phân tích dữ liệu, thống kê nâng cao sẽ tiếp tục đóng vai trò trung tâm trong việc thúc đẩy tiến bộ khoa học và công nghệ, góp phần xây dựng một xã hội dựa trên bằng chứng và ra quyết định thông minh.

Cuối cùng, chúng tôi mong muốn rằng bài thu hoạch này sẽ truyền cảm hứng cho những người yêu thích thống kê, khuyến khích họ tiếp tục khám phá và phát triển lĩnh vực đầy tiềm năng này. Thống kê nâng cao không chỉ là công cụ mà còn là nghệ thuật của việc khám phá tri thức từ dữ liệu - một kỹ năng vô cùng quý giá trong thế kỷ 21.

TÀI LIỆU THAM KHẢO

Tài liệu tham khảo

Tài liệu tiếng Việt

[1] Nguyễn Văn A (1999). *Phương pháp thống kê trong nghiên cứu khoa học*, NXB Đại học Quốc gia, Hà Nội.

[2] Nguyễn Văn B và Trần Thị C (1982). Ứng dụng kiểm định giả thuyết trong phân tích dữ liệu, *Tạp chí Thống kê Việt Nam*, Số 15, trang 25–35.

[3] Lê Văn C (2022). *Phân tích dữ liệu đa chiều trong thống kê*, Luận văn Thạc sĩ, Đại học Cần Thơ.

[4] Hoàng Thị D (2023). Phương pháp Monte Carlo trong mô phỏng thống kê, *Tạp chí Khoa học CTU*, tập 59, số 3, trang 45–58.

Tài liệu tiếng Anh

[5] Johnson, R.A. and Wichern, D.W. (2020). *Applied Multivariate Statistical Analysis*, 6th Edition, Pearson Education.

[6] Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd Edition, Duxbury Press.

[7] Montgomery, D.C., Peck, E.A., and Vining, G.G. (2017). *Introduction to Linear Regression Analysis*, 5th Edition, John Wiley & Sons.

[8] Agresti, A. and Finlay, B. (2018). *Statistical Methods for the Social Sciences*, 5th Edition, Pearson.

[9] Mood, A.M., Graybill, F.A., and Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd Edition, McGraw-Hill.

Phụ lục A

Hướng dẫn trích dẫn tài liệu

	Ba tác giả
	Tác giả là một cơ quan, tổ chức Ghi tên cơ quan và năm
	Nhiều tài liệu Sắp xếp các tài liệu theo năm xuất bản tăng dần. Nếu c
	Nhiều tài liệu cù
	Nhiều tài liệu cùng cách trích dẫn tác giả v
	Trích dẫn từ nguồn thứ cấp Ghi tác giả và năm (nếu có) của tài liệu gốc kè

	Trích dẫn nguyên văn Ghi tác giả, năm và trang viết. Đoạn trích dưới 40 từ: