

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC ĐỒNG THÁP**



BÀI GIẢNG
XÁC SUẤT THỐNG KÊ CHO TIN HỌC
Số tín chỉ: 2

ĐỒNG THÁP, THÁNG 11 NĂM 2024
TÀI LIỆU LƯU HÀNH NỘI BỘ

LỜI NÓI ĐẦU

1. Giới thiệu vắn tắt về học phần Xác suất thống kê cho Tin học:

Học phần Xác suất thống kê cho Tin học được giảng dạy cho sinh viên các ngành Sư phạm Tin học, Khoa học máy tính, Công nghệ thông tin, với thời lượng 2 tín chỉ, tương ứng với 30 tiết lý thuyết trên lớp. Nội dung gồm hai phần: Phần Xác suất và phần Thống kê, các kiến thức được tiếp nối từ xác suất, thống kê phổ thông.

Phần xác suất bao gồm các nội dung cơ bản về xác suất của biến cố, các công thức tính xác suất, biến ngẫu nhiên, hàm mật độ xác suất và phân phối xác suất, một số phân phối xác suất thông dụng, các số đặc trưng cơ bản của biến ngẫu nhiên. Phần thống kê bao gồm một số nội dung cơ bản của thống kê mô tả và thống kê suy diễn như lý thuyết mẫu, bài toán ước lượng tham số, kiểm định giả thuyết về giá trị trung bình, tỷ lệ, kiểm định về sự phù hợp với luật phân phối, kiểm định về sự độc lập, hồi quy và tương quan tuyến tính.

Vì thời lượng không nhiều nên bài giảng chủ yếu giới thiệu những vấn đề cốt lõi của lý thuyết xác suất và thống kê qua một số mô hình chung chung và một số mô hình về tin học. Các bài tập chủ yếu là bài tập tính toán, không có bài tập lý thuyết.

Để học tốt học phần này, người học cần tự ôn lại một số kiến thức cơ bản đã được học ở trung học phổ thông như là giải tích tổ hợp; cách tính nguyên hàm, tích phân, xác suất và lý thuyết mẫu căn bản. Để học tốt trên lớp, người học cần đọc trước bài mới trước khi đến lớp để tiếp thu tốt bài giảng của giảng viên. Về nhà, người học cần ôn lại bài cũ để giải lại các ví dụ trong bài giảng và các bài tập liên quan ở cuối chương. Việc bắt tay tính toán trực tiếp sẽ rèn thêm kỹ năng tính toán chính xác. Trong suốt quá trình học tập học phần, người học nên tham khảo thêm một số tài liệu tham khảo để hiểu rõ hơn về vấn đề đang học cũng như tham khảo cách giải một số dạng bài tập tương tự.

Bài giảng này vẫn tiếp tục được cập nhật, chúng tôi mong được góp ý của đồng nghiệp và người học để quyển bài giảng tóm tắt này được hoàn thiện hơn cho các khóa dạy sau. Mọi vấn đề trao đổi thêm về bài giảng, vui lòng liên lạc tác giả/nhóm tác giả hoặc với giảng viên trực tiếp giảng dạy lớp.

2. Vai trò của Xác suất thống kê trong Tin học:

Lý thuyết xác suất thống kê đóng vai trò quan trọng trong các ngành về Tin học,

Công nghệ thông tin, và Khoa học máy tính, chẳng hạn như:

a) Học máy (Machine Learning) và Trí tuệ nhân tạo (Artificial Intelligence):

Phân loại (Classification): Sử dụng xác suất thống kê để dự đoán lớp của một đối tượng dựa trên dữ liệu đã học. Ví dụ: phân loại email thành thư rác hoặc không thư rác sử dụng mô hình Naive Bayes, một thuật toán dựa trên xác suất.

Hồi quy (Regression): Để dự đoán giá trị liên tục như giá nhà, thuật toán hồi quy tuyến tính sử dụng các kỹ thuật thống kê để ước tính mối quan hệ giữa các biến đầu vào và đầu ra.

Mạng Bayes (Bayesian networks): Sử dụng lý thuyết xác suất để mô hình hóa mối quan hệ giữa các biến ngẫu nhiên. Mạng này được áp dụng trong chẩn đoán y tế, nhận diện giọng nói, và nhiều ứng dụng khác.

b) Khai phá dữ liệu (Data mining):

Phân cụm (Clustering): Sử dụng các phương pháp như K-means, dựa trên xác suất để nhóm các điểm dữ liệu thành các cụm dựa trên sự tương đồng. Điều này giúp trong việc phân đoạn khách hàng, phát hiện xu hướng, và tối ưu hóa tiếp thị.

Phát hiện bất thường (Anomaly detection): Các mô hình thống kê được sử dụng để xác định các điểm dữ liệu không tuân theo mô hình thông thường, như phát hiện gian lận trong giao dịch tài chính hoặc sự cố trong hệ thống.

c) Xử lý ngôn ngữ tự nhiên (Natural language processing):

Mô hình ngôn ngữ (Language models): Sử dụng xác suất để dự đoán từ tiếp theo trong một câu, như trong các ứng dụng tự động hoàn thành văn bản hoặc tạo văn bản. Các mô hình như N-gram và mô hình dựa trên xác suất Bayes được sử dụng rộng rãi.

Dịch máy (Machine translation): Dịch một câu từ ngôn ngữ này sang ngôn ngữ khác sử dụng mô hình thống kê để tính xác suất của một câu dịch.

d) Hệ thống khuyến nghị (Recommender systems):

Sử dụng lý thuyết xác suất để dự đoán sở thích của người dùng dựa trên lịch sử tương tác của họ. Các thuật toán như Collaborative Filtering và Matrix Factorization sử dụng xác suất để dự đoán và đề xuất các sản phẩm hoặc nội dung mà người dùng có thể quan tâm.

e) Mạng máy tính và an ninh mạng:

Phân tích lưu lượng mạng (Network traffic analysis): Sử dụng các mô hình thống kê để phát hiện các hành vi bất thường có thể chỉ ra một cuộc tấn công mạng.

Mã hóa và bảo mật thông tin: Sử dụng xác suất để đoán xác suất của một cuộc tấn công thành công và để tối ưu hóa các phương pháp mã hóa.

f) Thị giác máy tính (Computer vision):

Nhận diện đối tượng (Object recognition): Xác suất thống kê giúp xác định và phân loại các đối tượng trong hình ảnh hoặc video, như nhận diện khuôn mặt hoặc biển số xe.

Phân đoạn ảnh (Image segmentation): Sử dụng các mô hình thống kê để chia hình ảnh thành các vùng có ý nghĩa, điều này rất quan trọng trong nhận diện ảnh y tế hoặc xe tự hành.

g) Mô phỏng và tối ưu hóa

Mô phỏng Monte Carlo: Sử dụng phương pháp thống kê để mô phỏng các hệ thống phức tạp và phân tích xác suất của các kết quả khác nhau. Được sử dụng trong tài chính, dự báo thời tiết, và đánh giá hiệu suất hệ thống.

Lý thuyết xác suất thống kê cung cấp nền tảng toán học quan trọng giúp phân tích, dự đoán và ra quyết định dựa trên dữ liệu, là một phần không thể thiếu của các công nghệ và hệ thống hiện đại.

MỤC LỤC

1	SƠ LƯỢC VỀ XÁC SUẤT VÀ BIẾN NGẪU NHIÊN	8
1.1	Sơ lược về xác suất của biến cố	8
1.1.1	Ôn tập về phép thử ngẫu nhiên và biến cố ngẫu nhiên	8
1.1.2	Phép toán trên các biến cố	9
1.1.3	Định nghĩa xác suất theo dạng cổ điển	10
1.1.4	Định nghĩa xác suất theo phương pháp thống kê	11
1.1.5	Định nghĩa xác suất theo hình học	12
1.2	Một số công thức tính xác suất của biến cố	13
1.2.1	Công thức cộng xác suất	13
1.2.2	Xác suất có điều kiện, công thức nhân xác suất	13
1.2.3	Công thức xác suất đầy đủ và công thức Bayes	15
1.2.4	Công thức xác suất nhị thức	16
1.3	Biến ngẫu nhiên	18
1.3.1	Khái niệm về biến ngẫu nhiên	18
1.3.2	Biến ngẫu nhiên rời rạc	19
1.3.3	Biến ngẫu nhiên liên tục	21
1.3.4	Hàm phân phối xác suất của biến ngẫu nhiên	21
1.4	Các số đặc trưng của biến ngẫu nhiên	24
1.4.1	Kỳ vọng toán học	24
1.4.2	Phương sai, độ lệch chuẩn	26
1.4.3	Trung vị (median)	27
1.4.4	Mốt (Yếu vị)	28

1.5	Các phân phối xác suất thông dụng	28
1.5.1	Đối với biến ngẫu nhiên rời rạc	28
1.5.2	Đối với biến ngẫu nhiên liên tục	30
1.5.3	Tính gần đúng xác suất của phân phối nhị thức	34
	Bài tập Chương 1	37
2	LÝ THUYẾT MẪU VÀ BÀI TOÁN ƯỚC LƯỢNG THAM SỐ	41
2.1	Đám đông và mẫu	41
2.1.1	Đám đông và đặc tính nghiên cứu	41
2.1.2	Khái niệm mẫu và cách chọn mẫu	42
2.1.3	Cách biểu diễn mẫu, hàm phân phối mẫu	44
2.2	Các đặc trưng của mẫu	47
2.2.1	Các đặc trưng của mẫu	47
2.2.2	Phân phối của các đặc trưng mẫu	49
2.3	Ước lượng điểm	51
2.3.1	Tiêu chuẩn ước lượng điểm	51
2.3.2	Ước lượng điểm cho kỳ vọng, xác suất và phương sai	52
2.4	Ước lượng khoảng	53
2.4.1	Khái niệm về khoảng tin cậy	53
2.4.2	Khoảng tin cậy cho giá trị trung bình	54
2.4.3	Khoảng tin cậy cho tỉ lệ	57
2.4.4	Khoảng tin cậy cho phương sai	57
2.4.5	Tìm cỡ mẫu khi cho biết độ chính xác của ước lượng	58
	Bài tập Chương 2	61
3	KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ, TƯƠNG QUAN, HỒI QUY TUYẾN TÍNH	63
3.1	Bài toán kiểm định giả thuyết thống kê	63
3.1.1	Cặp giả thuyết thống kê	63

3.1.2	Tiêu chuẩn kiểm định giả thuyết thống kê	64
3.2	Kiểm định giả thuyết về giá trị trung bình	65
3.2.1	Khi đã biết phương sai σ^2	65
3.2.2	Khi chưa biết phương sai σ^2 , cỡ mẫu lớn $n \geq 30$	66
3.2.3	Khi chưa biết phương sai σ^2 , cỡ mẫu bé $n < 30$, X có phân phối chuẩn	67
3.3	Kiểm định giả thuyết về tỉ lệ	69
3.3.1	Bài toán kiểm định hai phía	69
3.3.2	Bài toán kiểm định một phía	70
3.4	Kiểm định (so sánh) hai tham số	70
3.4.1	Kiểm định (so sánh) hai giá trị trung bình	70
3.4.2	Kiểm định (so sánh) hai tỉ lệ	73
3.5	Kiểm định phi tham số	74
3.5.1	Kiểm định một phân phối (kiểm định về sự phù hợp)	74
3.5.2	Kiểm định về sự độc lập	77
3.6	Tương quan, hồi quy tuyến tính	79
3.6.1	Mở đầu về tương quan tuyến tính	79
3.6.2	Hệ số tương quan tuyến tính thực nghiệm	79
3.7	Phương trình hồi quy tuyến tính thực nghiệm	81
	Bài tập Chương 3	83
	PHỤ LỤC	87
	Phụ lục 1: Sử dụng máy tính cầm tay	87
	Phụ lục 2: Bảng phân phối chuẩn tắc $N(0, 1)$	92
	Phụ lục 3: Bảng phân phối t (Student)	93
	Phụ lục 4: Bảng phân phối χ^2	94
	Phụ lục 5: Đề thi tham khảo	95
	Phụ lục 6: Minh họa sử dụng gói công cụ Data Analysis trong Excel	97

CHƯƠNG 1

SƠ LƯỢC VỀ XÁC SUẤT VÀ BIẾN NGẪU NHIÊN

1.1 Sơ lược về xác suất của biến cố

1.1.1 Ôn tập về phép thử ngẫu nhiên và biến cố ngẫu nhiên

a) Phép thử (trial, experiment) và biến cố (event):

- *Hiện tượng tất yếu:* Là những hiện tượng nếu được thực hiện trong những điều kiện giống nhau thì cho kết quả hoàn toàn giống nhau. Ví dụ, đun nước đến 100°C thì nước sôi, giấy quỳ tím sẽ hóa đỏ khi được tẩm axit. Hiện tượng tất yếu là đối tượng nghiên cứu của Vật lý, Hóa học, ...
- *Hiện tượng ngẫu nhiên:* Là những hiện tượng dù đã được quan sát ở những điều kiện giống nhau nhưng cho kết quả có thể khác nhau và không thể biết trước được. Ví dụ, gieo đồng tiền, gieo con xúc sắc, chơi trò chơi xổ số, kết quả thi cuối kỳ của môn học Xác suất thống kê đang học. Hiện tượng ngẫu nhiên là đối tượng nghiên cứu của Xác suất thống kê.
- *Phép thử:* Là việc tiến hành các hoạt động thực nghiệm với điều kiện đặt ra ban đầu để nghiên cứu một hiện tượng ngẫu nhiên nào đó. Thường được ký hiệu bởi chữ \mathcal{T} .
- *Biến cố:* Mỗi kết quả của phép thử gọi là một *biến cố* hay *sự kiện*, ký hiệu biến cố bởi các chữ cái in hoa A, B, C, \dots . Tập hợp tất cả các kết quả của \mathcal{T} lập thành *không gian mẫu* (sample space), ký hiệu là Ω .

b) Các loại biến cố: Người ta chia thành một số loại biến cố sau đây

- *Biến cố rỗng (không thể, trống, bất khả)* (empty event): Là biến cố luôn không xảy ra khi thực hiện phép thử. Ký hiệu \emptyset .
- *Biến cố chắc chắn* (sure event): Là biến cố luôn xảy ra khi thực hiện phép thử. Biến cố đó chính là Ω , là biến cố lớn nhất của không gian mẫu Ω .

- *Biến cố sơ cấp (biến cố cơ bản, elementary event)*: Là biến cố đơn giản nhất, không thể biểu diễn được thành hợp của nhiều biến cố khác rỗng khác, và có thể xảy ra khi thực hiện phép thử. Ký hiệu biến cố sơ cấp là ω , ω chính là một điểm của không gian mẫu Ω .
- *Biến cố ngẫu nhiên (outcome)*: Là biến cố khác rỗng, có thể xảy ra khi thực hiện phép thử. Một biến cố ngẫu nhiên có thể chứa trong nó một hoặc nhiều biến cố sơ cấp.

Ví dụ 1.1.1. Gieo một con xúc sắc cân đối đồng chất trên mặt phẳng, đó là phép thử. Khi đó,

"Xuất hiện mặt 1", ..., "xuất hiện mặt 6" là các biến cố sơ cấp. Không gian các biến cố sơ cấp (không gian mẫu) là $\Omega = \{1, 2, 3, 4, 5, 6\}$.

"Xuất hiện mặt 7", "xuất hiện mặt 8" là các biến cố rỗng.

"Xuất hiện mặt có số chấm từ 1 đến 6" là biến cố chắc chắn.

B : "Xuất hiện mặt chẵn" là biến cố ngẫu nhiên. Biến cố B chứa trong nó 3 biến cố sơ cấp là 2, 4, 6, như vậy $B = \{2, 4, 6\}$.

C : "Xuất hiện số chấm lớn hơn 4" là biến cố ngẫu nhiên, C chứa 2 biến cố sơ cấp là 5, 6, như vậy $C = \{5, 6\}$.

1.1.2 Phép toán trên các biến cố

Cho phép thử \mathcal{T} và các biến cố A, B, C , ta có các khái niệm sau đây:

- *Quan hệ kéo theo*: Nếu A xảy ra kéo theo B cũng xảy ra thì ta ký hiệu $A \subset B$. Khi đó, tất cả các biến cố sơ cấp chứa trong A đều thuộc B .
- *Tổng (hợp, union) của hai biến cố*: Tổng của hai biến cố A và B (ký hiệu $A \cup B$ hay $A + B$) là một biến cố sao cho nó xảy ra khi và chỉ khi A xảy ra hoặc B xảy ra (nói cách khác có ít nhất một trong hai biến cố A và B xảy ra).
- *Tích (giao, intersection) của hai biến cố*: Tích của hai biến cố A và B (ký hiệu là $A \cap B$ hay AB) là một biến cố sao cho nó xảy ra khi và chỉ khi A xảy ra và B xảy ra.
- *Hai biến cố độc lập (independent events)* Hai biến cố A và B được gọi là độc lập với nhau nếu sự xảy ra hay không xảy ra của A đều không ảnh hưởng đến sự xảy ra hay không xảy ra của B và ngược lại.

Ví dụ về hai biến cố độc lập: Hai bóng đèn mắc song song. Gọi A_i : "Bóng đèn thứ i bị hỏng (không sáng)" $i = 1, 2$, khi đó A_1 và A_2 độc lập vì bóng đèn thứ nhất bị hỏng hay không hỏng đều không ảnh hưởng đến việc bóng đèn kia bị hỏng và ngược lại.

Tuy nhiên, nếu ta thêm điều kiện hai bóng đèn mắc nối tiếp thì khi đó A_1, A_2 không độc lập, vì bóng thứ nhất bị hỏng sẽ ảnh hưởng đến bóng đèn thứ hai.

- **Hai biến cố xung khắc** (mutually exclusive events): A và B được gọi xung khắc với nhau nếu $A \cap B = \emptyset$, nói cách khác chúng không đồng thời xảy ra trong cùng một phép thử.
- **Biến cố đối** (opposite event, complementary event): Biến cố đối (còn gọi là biến cố bù) của biến cố A ký hiệu là \bar{A} là biến cố thỏa điều kiện sau

$$\begin{cases} A \text{ và } \bar{A} \text{ xung khắc, tức là } A \cap \bar{A} = \emptyset; \\ A \cup \bar{A} = \Omega. \end{cases}$$

Khi tiến hành một phép thử bất kì, các biến cố trong mỗi phép thử đều có một mức độ xuất hiện khác nhau, trong nghiên cứu người ta cần đánh giá và biểu thị khả năng xuất hiện này bằng một con số. Số đo khả năng xuất hiện đó được gọi là xác suất của biến cố.

Có nhiều dạng định nghĩa khác nhau về xác suất của biến cố, sau đây là một số định nghĩa thường gặp, chủ yếu theo quan niệm cổ điển.

1.1.3 Định nghĩa xác suất theo dạng cổ điển

Giả sử \mathcal{T} là một phép thử mà không gian mẫu Ω có n biến cố sơ cấp đồng khả năng. A là một biến cố nào đó của \mathcal{T} . Xác suất của A được định nghĩa như sau:

Định nghĩa 1.1.2. Xác suất (probability) của biến cố A , ký hiệu $\mathbb{P}(A)$, là số không âm, biểu thị khả năng xảy ra của biến cố A và được xác định bởi

$$\mathbb{P}(A) = \frac{n(A)}{n(\Omega)} = \frac{\text{Số trường hợp thuận lợi của } A}{\text{Số trường hợp có thể xảy ra của phép thử}}$$

Trong đó, số *trường hợp thuận lợi* (number of favorable cases; number of favorable choices; number of successes) của A là số các biến cố sơ cấp mà nếu chúng xảy ra thì A xảy ra.

Ví dụ 1.1.3. Một hộp có 5 bi xanh, 7 bi đỏ, 8 bi vàng có kích cỡ và hình dạng như nhau. Chọn ngẫu nhiên 5 bi từ hộp. Tính xác suất chọn được:

- (a) 1 bi xanh, 2 bi đỏ, 2 bi vàng. (b) 2 bi xanh, 1 bi đỏ.
(c) 3 bi đỏ. (d) Ít nhất 4 bi đỏ.

Giải. (a) Số trường hợp có thể xảy ra của phép thử $n(\Omega) = C_{20}^5$. Số trường hợp thuận lợi để chọn được 1 bi xanh, 2 bi đỏ, 2 bi vàng: $C_5^1 C_7^2 C_8^2$.

Vậy xác suất cần tìm là $\frac{C_5^1 C_7^2 C_8^2}{C_{20}^5}$.

(b) $\frac{C_5^2 C_7^1 C_8^2}{C_{20}^5}$; (c) $\frac{C_7^3 C_{13}^2}{C_{20}^5}$; (d) $\frac{C_7^4 C_{13}^1 + C_7^5 C_{13}^0}{C_{20}^5}$.

Ví dụ 1.1.4. Một hệ thống mã hóa có 4 ký tự, mỗi ký tự có thể là một trong các chữ số từ 0 đến 9. Một hacker cố gắng đoán mật khẩu của hệ thống. Tính xác suất để hacker đoán đúng mật khẩu ngay lần thử đầu tiên. (Đáp số: 0,0001).

✂ Chú ý: Hạn chế của định nghĩa xác suất dạng cổ điển nêu trên là nó chỉ phù hợp đối với các phép thử có không gian mẫu gồm hữu hạn biến cố sơ cấp đồng khả năng. Nhưng trong thực tế, có rất nhiều phép thử quen thuộc có vô hạn biến cố sơ cấp, vô hạn kết quả.

1.1.4 Định nghĩa xác suất theo phương pháp thống kê

Làm đi làm lại một phép thử nào đó n lần, nếu có m lần biến cố A xuất hiện thì m được gọi là *tần số* và tỷ số m/n gọi là *tần suất* của biến cố A . Ký hiệu tần suất của A là $f_n(A)$.

Khi n thay đổi, tần suất $f_n(A)$ cũng thay đổi nhưng nó luôn dao động quanh một số cố định nào đó, n càng lớn thì $f_n(A)$ càng gần số cố định đó. Số cố định này gọi là xác suất của biến cố A theo nghĩa thống kê. Như vậy

$$\mathbb{P}(A) = \lim_{n \rightarrow \infty} f_n(A).$$

Trong thực hành, khi n đủ lớn ta xấp xỉ $\mathbb{P}(A)$ bởi m/n .

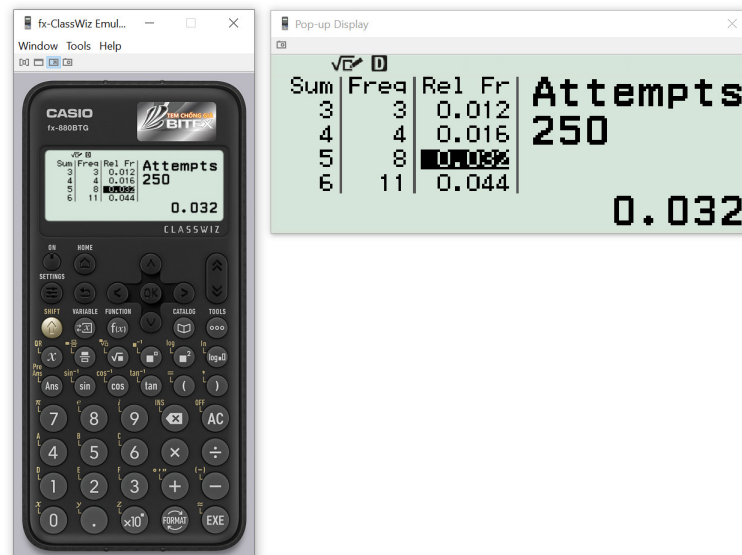
Ví dụ 1.1.5. Buffon đã gieo một đồng tiền cân đối, đồng chất 4040 lần thấy có 2048 lần xuất hiện mặt sấp. Khi đó, $\frac{m}{n} = 0,5080$.

Pearson đã gieo 12000 lần thấy có 6019 lần xuất hiện mặt sấp. Khi đó, $\frac{m}{n} = 0,5016$.

Pearson đã gieo 24000 lần thấy có 12012 lần xuất hiện mặt sấp. Khi đó, $\frac{m}{n} = 0,5005$.

Số cố định cần tìm trong trường hợp này là 0,5. Tức là xác suất xuất hiện mặt sấp khi ta gieo đồng tiền cân đối và đồng chất bằng 0,5.

Chú ý: Ngày nay người ta có nhiều phần mềm máy tính mô phỏng phép thử kiểu này trên máy vi tính, máy tính cầm tay. Ví dụ, trên máy tính Casio fx880BTG, có chức năng mô phỏng (Math Box) có thể thực hiện mô phỏng gieo 1, hoặc 2, hoặc 3 con xúc sắc tối đa $n = 250$ lần.



1.1.5 Định nghĩa xác suất theo hình học

Cho miền Ω đo được (trong đường thẳng, mặt phẳng, không gian ba chiều, ...) và miền con đo được S của Ω . Chọn ngẫu nhiên một điểm M trong miền Ω . Đặt A là biến cố " M thuộc miền S ". Khi đó, xác suất của biến cố A được xác định như sau:

$$\mathbb{P}(A) = \frac{\text{Độ đo}(S)}{\text{Độ đo}(\Omega)}.$$

Miền Ω chính là không gian biến cố sơ cấp.

- Nếu miền Ω là đường cong hay đoạn thẳng thì "độ đo" của Ω chính là độ dài của nó.
- Nếu miền Ω là hình phẳng hay mặt cong thì "độ đo" của Ω chính là diện tích của nó.
- Nếu miền Ω là hình khối ba chiều thì "độ đo" của Ω chính là thể tích của nó.

Ví dụ 1.1.6. Chọn ngẫu nhiên một điểm M trong hình vuông cạnh $2m$. Tìm xác suất để M không rơi vào hình tròn nội tiếp hình vuông này. (Kết quả: $1 - \pi/4$).

✂ Chú ý: Mỗi định nghĩa xác suất dạng nêu trên đều có hạn chế nhất định, không bao quát được hết các dạng khác nhau của phép thử. Do đó người ta xây dựng lý thuyết xác suất một cách chặt chẽ hơn bằng cách dùng công cụ của giải tích toán học hiện đại để định nghĩa độ đo xác suất. Một trong những cách như thế là định nghĩa độ đo xác suất theo hệ tiên đề Kolmogorov (xem [7]).

1.2 Một số công thức tính xác suất của biến cố

Mục này trình bày 5 công thức cơ bản để tính xác suất: (1) Công thức cộng xác suất; (2) Công thức nhân xác suất; (3) Công thức xác suất đầy đủ (toàn phần); (4) Công thức Bayes; (5) Công thức xác suất nhị thức.

1.2.1 Công thức cộng xác suất

a) Công thức cộng tổng quát: Cho A, B, C là các biến cố tùy ý, ta có công thức cộng tổng quát trường hợp 2 và 3 biến cố như sau

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB) \quad (1.1)$$

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(AB) - \mathbb{P}(BC) - \mathbb{P}(CA) + \mathbb{P}(ABC) \quad (1.2)$$

Tổng quát: Cho n biến cố tùy ý A_1, A_2, \dots, A_n , khi đó công thức cộng tổng quát là

$$\begin{aligned} \mathbb{P}(A_1 \cup A_2 \cup \dots \cup A_n) = & \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbb{P}(A_i A_j) + \sum_{1 \leq i < j < k \leq n} \mathbb{P}(A_i A_j A_k) \\ & - \dots + (-1)^{n+1} \mathbb{P}(A_1 \dots A_n). \end{aligned}$$

b) Công thức cộng đơn giản (cho các biến cố xung khắc): Nếu A, B, C là các biến cố xung khắc đôi một thì ta có công thức cộng đơn giản sau

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \quad \text{và} \quad \mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C). \quad (1.3)$$

✂ Chú ý: - Công thức (1.3) vẫn đúng cho n biến cố xung khắc đôi một ($n \geq 4$).

- Vì A, \bar{A} là hai biến cố đối nhau nên xung khắc và $A \cup \bar{A} = \Omega$, do đó $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup \bar{A}) = \mathbb{P}(A) + \mathbb{P}(\bar{A})$. Suy ra $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.

1.2.2 Xác suất có điều kiện, công thức nhân xác suất

a) Xác suất có điều kiện: Cho một phép thử \mathcal{S} và hai biến cố A và B . Xác suất của biến cố A với điều kiện B (kí hiệu $\mathbb{P}(A|B)$) là số không âm, biểu thị khả năng xảy ra của biến cố A khi biết biến cố B đã xảy ra và được xác định như sau

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} \quad (1.4)$$

trong đó, điều kiện $\mathbb{P}(B) > 0$.

Chú ý rằng, khi biến cố B xảy ra, không gian mẫu cũng sẽ thay đổi. Khi đó, chúng ta tính xác suất $\mathbb{P}(A|B)$ có nghĩa là tính xác suất của A trong điều kiện không gian mẫu mới, khi B đã xảy ra.

✂ Chú ý: Trường hợp A và B là hai biến cố độc lập thì sự xuất hiện (hay không xuất hiện) của B đều không ảnh hưởng đến A và ngược lại. Khi đó, xác suất của A với điều kiện B chính bằng xác suất của A , tức là $\mathbb{P}(A|B) = \mathbb{P}(A)$. Tương tự, $\mathbb{P}(B|A) = \mathbb{P}(B)$.

b) Công thức nhân xác suất đơn giản và tổng quát:

Cho phép thử \mathcal{T} và các biến cố A, B, C , ta có các dạng công thức nhân sau

• **Công thức nhân xác suất tổng quát:**

Từ công thức xác suất điều kiện (1.4) ta có công thức sau gọi là *công thức nhân xác suất* cho hai biến cố tùy ý

$$\mathbb{P}(AB) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) = \mathbb{P}(B) \cdot \mathbb{P}(A|B). \quad (1.5)$$

Công thức nhân cho n biến cố tùy ý A_1, A_2, \dots, A_n là

$$\mathbb{P}(A_1 A_2 \dots A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \mathbb{P}(A_3|A_1 A_2) \dots \mathbb{P}(A_n|A_1 A_2 \dots A_{n-1}).$$

• **Công thức nhân xác suất đơn giản:**

Khi A, B độc lập, ta có $\mathbb{P}(A|B) = \mathbb{P}(A)$ và $\mathbb{P}(B|A) = \mathbb{P}(B)$. Khi đó,

$$\mathbb{P}(AB) = \mathbb{P}(A) \mathbb{P}(B). \quad (1.6)$$

Ta có thể quy nạp tương tự cho công thức (1.6) đối với n biến cố độc lập A_1, A_2, \dots, A_n .

c) Một vài ví dụ cơ bản:

Ví dụ 1.2.1. Hai sinh viên dự thi kết thúc học phần môn Xác suất thống kê. Khả năng làm bài đạt yêu cầu giảng viên của hai người lần lượt là 70% và 80%. Tìm xác suất để việc làm bài của sinh viên như sau:

- a) Cả hai cùng đạt yêu cầu. b) Chỉ có người thứ nhất đạt yêu cầu.
c) Chỉ có một người đạt yêu cầu. d) Có ít nhất một người đạt yêu cầu.

Giải. Gọi A_i : "Người thứ i làm bài đạt yêu cầu", ($i = 1, 2$).

A_1, A_2 độc lập và $\mathbb{P}(A_1) = 0,7; \mathbb{P}(A_2) = 0,8$.

- a) $\mathbb{P}(A_1 A_2) = \mathbb{P}(A_1) \mathbb{P}(A_2) = 0,7 \cdot 0,8 = 0,56$.
b) $\mathbb{P}(A_1 \bar{A}_2) = \mathbb{P}(A_1) \mathbb{P}(\bar{A}_2) = 0,7 \cdot 0,2 = 0,14$.
c) $\mathbb{P}(A_1 \bar{A}_2 \cup \bar{A}_1 A_2) = \mathbb{P}(A_1) \mathbb{P}(\bar{A}_2) + \mathbb{P}(\bar{A}_1) \mathbb{P}(A_2) = 0,7 \cdot 0,2 + 0,3 \cdot 0,8 = 0,38$.
d) $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 A_2) = 0,7 + 0,8 - 0,56 = 0,94$.

Câu d) có thể giải bằng cách chia nhiều trường hợp rồi cộng xác suất của các trường hợp lại với nhau.

Ví dụ 1.2.2. Một máy chủ có 3 phần cứng A, B và C. Xác suất phần cứng A gặp sự cố trong 1 năm là 0,02; phần cứng B là 0,03 và phần cứng C là 0,05. Giả sử các phần cứng này hoạt động độc lập, tính xác suất trong 1 năm có ít nhất một phần cứng gặp sự cố. (Đáp số: 0,0963).

1.2.3 Công thức xác suất đầy đủ và công thức Bayes

a) Khái niệm về hệ biến cố đầy đủ (hệ toàn phần):

Hệ n biến cố $\{A_1, A_2, \dots, A_n\}$ được gọi là *hệ biến cố đầy đủ* (hệ toàn phần; complete system of events) nếu thỏa mãn đồng thời hai điều kiện sau

i) A_1, A_2, \dots, A_n từng đôi một xung khắc (tức là $A_i \cap A_j = \emptyset, i \neq j, i, j = \overline{1, n}$);

ii) $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$.

Chú ý: Nếu A là một biến cố thì hệ $\{A, \bar{A}\}$ là hệ đầy đủ vì hệ này thỏa mãn đồng thời (i), (ii).

b) Công thức xác suất đầy đủ (toàn phần):

Cho một hệ biến cố đầy đủ $\{A_1, A_2, \dots, A_n\}$ và B là một biến cố bất kỳ của phép thử, ta có công thức tính xác suất của biến cố B như sau

$$\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B|A_1) + \mathbb{P}(A_2)\mathbb{P}(B|A_2) + \dots + \mathbb{P}(A_n)\mathbb{P}(B|A_n). \quad (1.7)$$

(1.7) được gọi là *công thức xác suất đầy đủ* (hay công thức xác suất toàn phần).

c) Công thức Bayes: Giả sử $\{A_1, A_2, \dots, A_n\}$ là hệ biến cố đầy đủ và B là một biến cố bất kỳ của phép thử. $\mathbb{P}(B)$ được tính theo công thức xác suất đầy đủ (1.7).

Viết lại công thức nhân xác suất

$$\mathbb{P}(A_i B) = \mathbb{P}(A_i)\mathbb{P}(B|A_i) = \mathbb{P}(B)\mathbb{P}(A_i|B), \quad i = 1, 2, \dots, n.$$

Suy ra

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(B)}, \quad i = 1, 2, \dots, n. \quad (1.8)$$

(1.8) được gọi là *công thức Bayes*¹, trong đó $\mathbb{P}(B)$ được tính theo công thức xác suất đầy đủ (1.7).

¹Thomas Bayes (1701-1761) là một nhà thống kê, nhà triết học người Anh. Nguồn: en.wikipedia.org.

d) Một vài ví dụ cơ bản:

Ví dụ 1.2.3. Một kho hàng có 20 thùng cà phê loại 1 (xuất khẩu) và 70 thùng cà phê loại 2 (tiêu thụ nội địa). Mỗi thùng cà phê loại 1 có 50 hộp cà phê trong đó có 4 hộp có quà trúng thưởng. Mỗi thùng cà phê loại 2 có 40 hộp cà phê trong đó có 3 hộp có quà trúng thưởng.

a) Chọn ngẫu nhiên một thùng cà phê trong kho rồi lấy ngẫu nhiên từ thùng này ra một hộp. Tính xác suất để hộp cà phê này có quà trúng thưởng.

b) Giả sử rằng chọn được hộp cà phê có trúng thưởng. Tính xác suất để hộp cà phê đó là cà phê thuộc thùng loại 1 dành cho xuất khẩu.

Giải. a) Gọi B : "Hộp cà phê có quà trúng thưởng"; A_i : "Hộp cà phê loại i ", $i = 1, 2$. Khi đó, $\{A_1, A_2\}$ là hệ đầy đủ các biến cố và

$$\mathbb{P}(A_1) = \frac{20}{90} = \frac{2}{9}; \quad \mathbb{P}(A_2) = \frac{70}{90} = \frac{7}{9}; \quad \mathbb{P}(B|A_1) = \frac{4}{50} = \frac{2}{25}; \quad \mathbb{P}(B|A_2) = \frac{3}{40}.$$

Theo công thức xác suất đầy đủ, xác suất để gặp hộp cà phê có trúng thưởng là

$$\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B|A_1) + \mathbb{P}(A_2)\mathbb{P}(B|A_2) = \frac{137}{1800} = 0,0761.$$

b) Áp dụng công thức Bayes ta có

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(A_1)\mathbb{P}(B|A_1)}{\mathbb{P}(B)} = \frac{(2/9) \times (2/25)}{0,0761} = 0,2336.$$

1.2.4 Công thức xác suất nhị thức

a) Phép thử Bernoulli: Một phép thử được gọi là *phép thử Bernoulli* nếu thỏa mãn 2 điều kiện

- i) Chỉ xét hai kết quả là thành công và thất bại A và \bar{A} . Phép thử thành công nếu A xuất hiện, ngược lại phép thử thất bại nếu \bar{A} xuất hiện, trong đó A là một biến cố nào đó của phép thử mà ta đã quan tâm từ trước.
- ii) Xác suất $\mathbb{P}(A) = p$, $\mathbb{P}(\bar{A}) = 1 - p = q$ là như nhau đối với mọi lần thực hiện phép thử.

Tiến hành n phép thử Bernoulli một cách độc lập (tức kết quả của phép thử này không làm ảnh hưởng đến kết quả của phép thử kia và ngược lại), n kết quả ngẫu nhiên của n phép thử này lập thành dãy phép thử Bernoulli.

Ví dụ 1.2.4. Gieo một con xúc sắc, gọi A là biến cố "xuất hiện mặt hai chấm", nếu trong phép thử này ta chỉ quan tâm biến cố A có xảy ra hay không thì đây chính là một phép thử Bernoulli. Phép thử này chỉ có hai kết quả cần nghiên cứu là A và \bar{A} . Hơn nữa, sau khi thực hiện phép thử này:

- Mặt hai chấm xuất hiện (A xuất hiện) thì phép thử được gọi là thành công với xác suất như nhau đối với mỗi lần gieo là $\mathbb{P}(A) = p = \frac{1}{6}$.

- Không phải mặt hai chấm xuất hiện (\bar{A} xuất hiện) thì phép thử được gọi là thất bại với xác suất $\mathbb{P}(\bar{A}) = 1 - p = \frac{5}{6}$.

b) Công thức xác suất nhị thức:

Bài toán: Thực hiện n phép thử Bernoulli, xác suất thành công trong mỗi phép thử là p . Tìm xác suất để cho trong n lần thử trên có k lần thành công ($0 \leq k \leq n$).

Ký hiệu xác suất này là $\mathbb{P}_n(k; p)$ hoặc $\mathbb{P}_n(k)$, đôi khi ta viết tắt là $\mathbb{P}(k)$.

Ta có công thức sau gọi là *công thức xác suất nhị thức* (hay công thức Bernoulli)²

$$\mathbb{P}(k) = C_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n. \quad (1.9)$$

Ví dụ 1.2.5. Một game thủ bắn liên tiếp 15 viên đạn vào bia, trong một phần mềm game máy tính. Xác suất trúng bia của game thủ này là 85%. Tìm xác suất để trong 15 viên vừa bắn có:

- (1) 5 viên trúng bia.
- (2) Từ 5 đến 7 viên trúng bia.
- (3) Ít nhất 1 viên trúng bia.

Giải. Đây là $n = 15$ phép thử Bernoulli với xác suất thành công $p = 85\%$.

$$(1) \mathbb{P}(k = 5) = C_n^k p^k q^{n-k} = C_{15}^5 0,85^5 \cdot 0,15^{10} = 7,6836 \cdot 10^{-6}.$$

$$(2) \mathbb{P}(5 \leq k \leq 7) = \mathbb{P}(k = 5) + \mathbb{P}(k = 6) + \mathbb{P}(k = 7) = C_{15}^5 0,85^5 \cdot 0,15^{10} + C_{15}^6 0,85^6 \cdot 0,15^9 + C_{15}^7 0,85^7 \cdot 0,15^8.$$

$$(3) \mathbb{P}(k \geq 1) = 1 - \mathbb{P}(k < 1) = 1 - \mathbb{P}(k = 0) = 1 - C_{15}^0 0,85^0 \cdot 0,15^{15}.$$

c) Số có khả năng nhất

Số lần thành công m_0 có xác suất $\mathbb{P}(m_0)$ lớn nhất được gọi là *số có khả năng nhất*. Bằng suy luận toán học, người ta chứng minh được $np - q \leq m_0 \leq np - q + 1$. Suy ra cách tìm m_0 như sau

- Số có khả năng nhất bằng $np - q$ hoặc bằng $np - q + 1$ nếu $np - q$ là số nguyên.

²probability mass function: Hàm khối xác suất.

- Số có khả năng nhất bằng $[np - q] + 1$ nếu $np - q$ không là số nguyên.

Trong đó, $[x]$ là phần nguyên của số thực x (là số nguyên $\leq x$ và gần x nhất), ví dụ $[2,95] = 2$; $[0,15] = 0$; $[-2,95] = -3$; $[-0,15] = -1$.

Ví dụ 1.2.6. Một game thủ bắn liên tiếp 15 viên đạn vào bia trong một phần mềm game máy tính. Xác suất trúng bia của xạ thủ này là 85%. Tìm số đạn bắn trúng bia có khả năng nhất (trong số 15 viên vừa bắn).

Giải. Ta có $np - q = 15 \cdot 0,85 - 0,15 = 12,6$ là số không nguyên. Suy ra số đạn bắn trúng có khả năng nhất là $[np - q] + 1 = [12,6] + 1 = 12 + 1 = 13$ viên.

Ví dụ 1.2.7. Một đề thi trắc nghiệm gồm 40 câu, mỗi câu có 4 đáp án trong đó chỉ có 1 đáp án đúng. Điểm mỗi câu đúng là 0,25 điểm. Sinh viên chọn ngẫu nhiên các kết quả trong đề thi.

a) Tính xác suất bài thi được 6 điểm. b) Tìm số điểm có khả năng nhất của sinh viên. Kết quả: a) $2,238 \cdot 10^{-6}$; b) 2,5đ.

1.3 Biến ngẫu nhiên

1.3.1 Khái niệm về biến ngẫu nhiên

a) **Khái niệm:** Một đại lượng X nhận các giá trị của nó với xác suất tương ứng nào đó được gọi là đại lượng ngẫu nhiên hay biến ngẫu nhiên (Random variable; Stochastic variable) ([5]).

Nói cách khác, biến ngẫu nhiên X là một hàm xác định trên không gian các biến cố sơ cấp Ω và nhận mỗi giá trị thực tương ứng với một xác suất nào đó ([1]). Tức là, $X : \Omega \rightarrow \mathbb{R}$, $\omega \mapsto X(\omega) \in \mathbb{R}$.

Ta thường ký hiệu biến ngẫu nhiên bởi các chữ cái in hoa X, Y, Z, \dots , hoặc ξ, η, ζ, \dots

Hai biến ngẫu nhiên X, Y được gọi là độc lập với nhau nếu mọi biến cố liên quan đến X độc lập với biến cố bất kỳ liên quan đến Y .

b) **Các ví dụ:**

Ví dụ 1.3.1. Gieo một đồng tiền. Gọi X là biến ngẫu nhiên với quy ước nếu ra mặt ngửa thì $X = 0$, ra mặt sấp thì $X = 1$. Ta thấy xác suất xuất hiện mặt sấp là $1/2$, xác suất ra mặt ngửa là $1/2$. Ghi lại kết quả trên dưới dạng bảng

X	0	1
\mathbb{P}	$\frac{1}{2}$	$\frac{1}{2}$

Ví dụ 1.3.2. Cũng phép thử gieo đồng tiền nhưng quy ước nếu ra mặt ngửa thì coi như thua và phải nộp phạt 10đ, sấp coi như thắng và nhận được 10đ. Gọi Y là biến ngẫu nhiên chỉ số tiền nhận được, khi đó Y sẽ là -10 hay +10 và đều có xác suất như nhau bằng $1/2$. Khi đó ta có bảng

Y	-10	10
\mathbb{P}	$\frac{1}{2}$	$\frac{1}{2}$

Ví dụ 1.3.3. Gọi X là chiều cao (đơn vị: mét) của sinh viên trường ĐH Đồng Tháp thì X là biến ngẫu nhiên nhận giá trị tùy ý trong khoảng $[1, 0m; 2, 0m]$. Gọi Y (đơn vị: giờ) là tuổi thọ của một loại bóng đèn điện thì Y là biến ngẫu nhiên nhận giá trị tùy ý trong khoảng $[0; +\infty)$.

Ví dụ 1.3.4. Trồng 10 cây con, xác suất sống của mỗi cây là 0,8. Coi việc trồng các cây là các phép thử lặp (thử trong cùng điều kiện như nhau và các kết quả mỗi lần thử độc lập với nhau). Gọi X là số cây sống, ta có $X = \{0, 1, 2, \dots, 10\}$.

Ví dụ 1.3.5. Gieo một đồng tiền cho đến khi nào xuất hiện mặt sấp thì dừng lại. Gọi X là số mặt ngửa của mỗi lần thử. Ta có X là biến ngẫu nhiên, tập giá trị của X là vô hạn đếm được $X = \{0, 1, 2, \dots, k, \dots\}$.

Từ các ví dụ trên ta thấy tập giá trị có thể nhận của biến ngẫu nhiên có thể hữu hạn, vô hạn đếm được hoặc vô hạn không đếm được.

- c) **Chú ý:** Tổng, tích, thương (điều kiện biến ngẫu nhiên ở mẫu khác không) của hai hay nhiều biến ngẫu nhiên cũng là biến ngẫu nhiên. Tổng quát ta có các hàm sơ cấp của một biến ngẫu nhiên nếu tồn tại cũng là biến ngẫu nhiên.
- d) **Phân loại biến ngẫu nhiên:** Căn cứ theo giá trị của biến ngẫu nhiên người ta phân chia biến ngẫu nhiên thành hai loại gồm *biến ngẫu nhiên rời rạc* (discrete random variable) và *biến ngẫu nhiên liên tục* (continuous random variable).

1.3.2 Biến ngẫu nhiên rời rạc

- a) **Định nghĩa:** Biến ngẫu nhiên X được gọi là *rời rạc* nếu tập các giá trị của nó hữu hạn hoặc vô hạn đếm được.

Giả sử biến ngẫu nhiên $X = \{x_1, x_2, \dots, x_n, \dots\}$ và $\mathbb{P}(X = x_i) = p_i, i = 1, 2, \dots$. Để mô tả biến ngẫu nhiên rời rạc X ta có bảng sau gọi là *bảng phân bố xác suất*

X	x_1	x_2	\dots	x_n	\dots
\mathbb{P}	p_1	p_2	\dots	p_n	\dots

trong đó, $\sum_i p_i = p_1 + p_2 + \dots = 1$.

Lập bảng phân bố xác suất của biến ngẫu nhiên rời rạc X gồm 2 bước:

- + Bước 1: Liệt kê tất cả các giá trị có thể có của X : $X = \{x_1, x_2, \dots, x_n, \dots\}$.
- + Bước 2: Tính $\mathbb{P}(X = x_i) = p_i$, $i = 1, 2, \dots$ và viết bảng phân bố xác suất.

b) Các ví dụ:

Ví dụ 1.3.6. Trong số 10 hạt giống đem trồng có 7 hạt ra hoa vàng, 3 hạt ra hoa trắng. Chọn ngẫu nhiên 2 hạt. Gọi X là số hạt ra hoa vàng trong số 2 hạt được chọn.

(a) Tìm bảng phân bố xác suất của X ?

(b) Từ kết quả câu a), tính $\mathbb{P}(X \leq 1)$, $\mathbb{P}(X < 1)$?

Giải. a) X là số hạt ra hoa vàng trong số 2 hạt được chọn ra nên $X = \{0, 1, 2\}$.

$$\mathbb{P}(X = 0) = \frac{C_7^0 C_3^2}{C_{10}^2} = \frac{1}{15}; \quad \mathbb{P}(X = 1) = \frac{C_7^1 C_3^1}{C_{10}^2} = \frac{7}{15}; \quad \mathbb{P}(X = 2) = \frac{C_7^2 C_3^0}{C_{10}^2} = \frac{7}{15}$$

Bảng phân bố xác suất của X :

X	0	1	2
\mathbb{P}	$\frac{1}{15}$	$\frac{7}{15}$	$\frac{7}{15}$

b) Dựa vào bảng phân bố xác suất của X ta có

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \frac{8}{15}.$$

$$\mathbb{P}(X < 1) = \mathbb{P}(X = 0) = \frac{1}{15}.$$

Ví dụ 1.3.7. Một người mở 2 cửa hàng kinh doanh một loại mặt hàng ở hai khu vực 1 và 2. Khả năng kinh doanh có lãi sau 1 tháng ở khu vực 1 và 2 lần lượt là 60% và 70%. Gọi Y là số cửa hàng kinh doanh có lãi. Tìm bảng phân bố xác suất của Y . Từ đó tính $\mathbb{P}(Y < 2)$.

Giải. Gọi A_i : "Cửa hàng ở khu vực i kinh doanh có lãi", $i = 1, 2$. Ta có A_1, A_2 độc lập.

Y là số cửa hàng kinh doanh có lãi nên $Y = \{0, 1, 2\}$.

$$\mathbb{P}(Y = 0) = \mathbb{P}(\bar{A}_1 \bar{A}_2) = \mathbb{P}(\bar{A}_1) \mathbb{P}(\bar{A}_2) = 0,4 \cdot 0,3 = 0,12.$$

$$\mathbb{P}(Y = 1) = \mathbb{P}(A_1 \bar{A}_2) + \mathbb{P}(\bar{A}_1 A_2) = 0,6 \cdot 0,3 + 0,4 \cdot 0,7 = 0,46.$$

$$\mathbb{P}(Y = 2) = \mathbb{P}(A_1 A_2) = \mathbb{P}(A_1) \mathbb{P}(A_2) = 0,6 \cdot 0,7 = 0,42.$$

Bảng phân phối của Y :

Y	0	1	2
\mathbb{P}	0,12	0,46	0,42

Từ bảng phân bố xác suất của Y , ta có

$$\mathbb{P}(Y < 2) = \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1) = 0,12 + 0,46 = 0,58.$$

1.3.3 Biến ngẫu nhiên liên tục

a) **Định nghĩa:** Biến ngẫu nhiên X được gọi là *liên tục* nếu nó nhận giá trị tùy ý trong một khoảng (hoặc hợp của các khoảng) nào đó trên \mathbb{R} (mở rộng là khoảng $(-\infty; +\infty)$).

Một cách hiểu khác, biến ngẫu nhiên X được gọi là liên tục nếu nó có tập giá trị vô hạn không đếm được.

Biến ngẫu nhiên liên tục ta không mô tả bằng bảng phân bố xác suất mà người ta dùng khái niệm hàm mật độ xác suất. Hàm $p(x)$ được gọi là *hàm mật độ xác suất* (probability density function) (gọi tắt là hàm mật độ) của biến ngẫu nhiên X khi và chỉ khi $p(x)$ thỏa mãn

$$\begin{cases} p(x) \geq 0, \forall x \in \mathbb{R}; \\ \int_{-\infty}^{+\infty} p(x)dx = 1. \end{cases}$$

b) **Các ví dụ:**

Ví dụ 1.3.8. Cho hàm số $p(x) = \begin{cases} kx(4-x) & \text{nếu } x \in [0, 4] \\ 0 & \text{nếu trái lại.} \end{cases}$

Tìm hằng số k để hàm số đã cho là hàm mật độ xác suất của biến ngẫu nhiên X nào đó.

Ví dụ 1.3.9. Hàm số $p(x) = \begin{cases} 0 & (\text{nếu } x < a \text{ hoặc } x > b) \\ \frac{1}{b-a} & (\text{nếu } a \leq x \leq b) \end{cases}$ là hàm mật độ của

biến ngẫu nhiên X nhận mọi giá trị trên $[a; b]$ với khả năng đều như nhau, được gọi là *hàm mật độ đều* (uniform density function) trên $[a; b]$.

Thật vậy ta thấy $p(x) \geq 0, \forall x \in (-\infty; +\infty)$ và

$$\int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^a p(x)dx + \int_a^b p(x)dx + \int_b^{+\infty} p(x)dx = \int_a^b \frac{1}{(b-a)}dx = 1.$$

1.3.4 Hàm phân phối xác suất của biến ngẫu nhiên

a) **Định nghĩa hàm phân phối xác suất:**

Cho biến ngẫu nhiên X , *hàm phân phối xác suất* (probability distribution function) của biến ngẫu nhiên X là hàm $F : \mathbb{R} \rightarrow [0, 1]$, được xác định như sau

$$F_X(x) = \mathbb{P}(X < x). \quad (1.10)$$

- Hàm phân phối xác suất còn được gọi là *hàm phân phối tích lũy* (cumulative distribution function), hàm phân phối.
- Trong định nghĩa trên, x là biến của hàm F , x nhận giá trị thực. Tại một điểm x bất kỳ hàm $F_X(x)$ chính là xác suất để biến ngẫu nhiên X nhận giá trị bé hơn x .³
- Nếu không bị nhầm lẫn ta có thể ký hiệu hàm phân phối của X là $F(x)$ thay cho $F_X(x)$.

b) Tính chất hàm phân phối:

Hàm phân phối F của biến ngẫu nhiên có các tính chất cơ bản sau

- i) $0 \leq F(x) \leq 1$.
- ii) $F(-\infty) = 0, \quad F(\infty) = 1$.
- iii) Hàm F không giảm trên \mathbb{R} , tức là nếu có $x_1 < x_2$ thì $F(x_1) \leq F(x_2)$.
- iv) $\mathbb{P}(a \leq X < b) = F(b) - F(a)$.

c) Cách tính giá trị hàm phân phối: Có hai trường hợp

★ **Trường hợp X là biến ngẫu nhiên rời rạc:** Nếu ta có $\mathbb{P}(X = x_i) = p_i$ thì

$$F(x) = \sum_{x_i < x} p_i \quad (\text{tổng các } p_i \text{ nằm bên trái điểm } x).$$

Ví dụ 1.3.10. Cho biến ngẫu nhiên rời rạc X có bảng phân bố xác suất

X	0	1	2
\mathbb{P}	0,2	0,7	0,1

- a) Tính $F(\sqrt{2}), F(100), \mathbb{P}(X \geq 1), \mathbb{P}(X = 5)$.
- b) Tìm hàm phân phối xác suất $F(x)$ của X và vẽ đồ thị của $F(x)$.

Giải. a) Ta có, $F(\sqrt{2}) = 0,2 + 0,7 = 0,9$; $F(100) = 0,2 + 0,7 + 0,1 = 1$;
 $\mathbb{P}(X \geq 1) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = 0,7 + 0,1 = 0,8$; $\mathbb{P}(X = 5) = 0$.

b) Ta có $F(x) = \sum_{x_i < x} p_i$, nên

$$F(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ 0,2 & \text{nếu } 0 < x \leq 1 \\ 0,2 + 0,7 = 0,9 & \text{nếu } 1 < x \leq 2 \\ 0,2 + 0,7 + 0,1 = 1 & \text{nếu } x > 2. \end{cases}$$

³Một số giáo trình trong nước và quốc tế, người ta định nghĩa $F_X(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$. So với công thức (1.10), hàm phân phối $F_X(x)$ theo cách định nghĩa này "tích lũy" cả trường hợp $\mathbb{P}(X = x)$. Điều này là không quan trọng, bởi vì đối với trường hợp biến ngẫu nhiên liên tục thì xác suất tại 1 điểm bằng 0, đối với trường hợp biến ngẫu nhiên rời rạc, hàm phân phối chỉ khác nhau tại những điểm có xác suất khác 0, tức là tại đó hàm phân phối liên tục phải hay liên tục trái.

Đồ thị của hàm phân phối xác suất của biến ngẫu nhiên rời rạc có dạng bậc thang, bị gián đoạn tại một số điểm x_i , $i = 1, 2, \dots$

★ **Trường hợp X là biến ngẫu nhiên liên tục:** Nếu X có hàm mật độ $p(x)$ thì

$$F(x) = \int_{-\infty}^x p(t)dt.$$

Ví dụ 1.3.11. Cho $p(x) = \begin{cases} \frac{3}{10}(x^2 + 1) & \text{nếu } 1 \leq x \leq 2 \\ 0 & \text{trường hợp còn lại.} \end{cases}$ là hàm mật độ

xác suất của biến ngẫu nhiên liên tục X . Tính $F(1,5)$ và $\mathbb{P}(X < 3)$. Tìm hàm phân phối xác suất $F(x)$.

Giải. Ta có $F(1,5) = \int_{-\infty}^{1,5} p(x)dx = \int_1^{1,5} \frac{3}{10}(x^2 + 1)dx = 0,3875$.

$$\mathbb{P}(X < 3) = F(3) = \int_{-\infty}^3 p(x)dx = \int_1^2 \frac{3}{10}(x^2 + 1)dx = 1.$$

$$F(x) = \int_{-\infty}^x p(t)dt =$$

$$\begin{cases} 0 & \text{nếu } x < 1 \\ \int_1^x \frac{3}{10}(t^2 + 1)dt & \text{nếu } 1 \leq x \leq 2 \\ \int_1^2 \frac{3}{10}(t^2 + 1)dt & \text{nếu } x > 2 \end{cases} = \begin{cases} 0 & \text{nếu } x < 1 \\ \frac{x^3}{10} + \frac{3x}{10} - \frac{3}{5} & \text{nếu } 1 \leq x \leq 2 \\ 1 & \text{nếu } x > 2 \end{cases}.$$

Nhận xét 1.3.12. (i) Nếu hàm mật độ xác suất của biến ngẫu nhiên X liên tục tại x_0 thì $\mathbb{P}(X = x_0) = 0$.

(ii) Nếu X là biến ngẫu nhiên liên tục có là hàm mật độ $p(x)$ xác định trên khoảng $(a; b)$ thì

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) = \int_a^b p(x)dx.$$

Trong công thức trên thì a có thể là $-\infty$, b có thể là $+\infty$.

Tổng hợp lại một số công thức đối với biến ngẫu nhiên liên tục:

$$\begin{aligned}
 1) \quad & \mathbb{P}(a \leq X < b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) \\
 & = \int_a^b p(x)dx = F(b) - F(a); \\
 2) \quad & F(x) = \mathbb{P}(X < x) = \int_{-\infty}^x p(t)dt; \\
 3) \quad & p(x) = F'(x) \text{ (đạo hàm của } F(x) \text{ theo biến } x); \\
 4) \quad & \mathbb{P}(X < a) = F(a) = \int_{-\infty}^a p(x)dx; \quad \mathbb{P}(X \geq a) = 1 - F(a) = \int_a^{+\infty} p(x)dx.
 \end{aligned}$$

Công thức 2) và 3) cho thấy mối quan hệ giữa hàm mật độ $p(x)$ và hàm phân phối xác suất $F(x)$. Nếu biết một hàm nào đó ta sẽ tìm được hàm còn lại.

1.4 Các số đặc trưng của biến ngẫu nhiên

Khi xét một biến ngẫu nhiên nào đó, nếu ta có bảng phân bố xác suất (đối với biến ngẫu nhiên rời rạc) hoặc hàm mật độ xác suất (đối với biến ngẫu nhiên liên tục) hoặc hàm phân phối xác suất thì coi như ta có sự hiểu biết tương đối đầy đủ về biến ngẫu nhiên này.

Trong một số trường hợp không cần phải biết đầy đủ như vậy mà chỉ cần biết một số đặc trưng cho dãy phân phối về một khía cạnh nào đó. Người ta chia các số đặc trưng ra thành hai nhóm: Nhóm đặc trưng cho vị trí (kỳ vọng, trung vị, mốt, tứ phân vị dưới, tứ phân vị trên, ...) và nhóm đặc trưng cho độ phân tán (phương sai, độ lệch chuẩn, mômen bậc k , biên độ, hệ số biến động, ...).

Sau đây ta xét một vài số đặc trưng thường gặp.

1.4.1 Kỳ vọng toán học

a) Định nghĩa: Kỳ vọng (expectation) của biến ngẫu nhiên X (Kí hiệu: $\mathbb{E}(X)$, $\mathbb{E}X$, $\mathbb{M}(X)$) nếu tồn tại là giá trị được xác định như sau

$$\mathbb{E}X = \begin{cases} \sum_i x_i p_i & \text{(nếu } X \text{ rời rạc có } \mathbb{P}(X = x_i) = p_i); \\ \int_{-\infty}^{+\infty} xp(x)dx & \text{(nếu } X \text{ liên tục có hàm mật độ xác suất } p(x)) \end{cases}$$

b) Ý nghĩa: Kỳ vọng là trung bình theo xác suất của tất cả các giá trị có thể nhận của biến ngẫu nhiên đó. Trong nhiều trường hợp kỳ vọng có giá trị khác với giá trị trung bình, chẳng hạn đối với các biến ngẫu nhiên có vô hạn giá trị.

c) Tính chất:

- i) $\mathbb{E}(c) = c$ (c là hằng số).
- ii) $\mathbb{E}(cX) = c\mathbb{E}X$.
- iii) $\mathbb{E}(X \pm Y) = \mathbb{E}X \pm \mathbb{E}Y$ (X, Y là hai biến ngẫu nhiên tùy ý).
- iv) $\mathbb{E}(XY) \neq \mathbb{E}X \cdot \mathbb{E}Y$, nhưng nếu X, Y độc lập thì $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$.
- v) $|\mathbb{E}X| \leq \mathbb{E}|X|$.
- vi) Nếu $X \leq Y$ thì $\mathbb{E}X \leq \mathbb{E}Y$.
- vii) $\mathbb{E}(f(X)) = \sum_i f(x_i)p_i$ (nếu X rời rạc) $= \int_{-\infty}^{+\infty} f(x)p(x)dx$ (nếu X liên tục).

Ví dụ 1.4.1. Tính kỳ vọng của biến ngẫu nhiên rời rạc X có bảng phân bố xác suất như sau

X	0	1	2
\mathbb{P}	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Giải. Ta có $\mathbb{E}X = \sum_i x_i p_i = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$.

Ví dụ 1.4.2. Một trò chơi may rủi như sau: Có 10 viên bi giống nhau trong một cái hộp kín, trong đó có 7 bi xanh và 3 bi đỏ. Người chơi chọn ngẫu nhiên 1 bi trong hộp. Nếu chọn được bi đỏ thì được nhận 20.000 đồng, nếu chọn được bi xanh thì nộp 9000 đồng. Hỏi ta có nên tham gia trò chơi này nhiều lần không?

Giải. Gọi X là số tiền nhận được khi chơi ta có $X = \{-9000, 20000\}$. Nếu muốn biết có nên tham gia trò chơi này nhiều lần hay không ta cần tính số tiền trung bình (theo xác suất) của mỗi lần chơi.

Ta có $\mathbb{P}(X = -9000) = \frac{7}{10} = 0,7$; $\mathbb{P}(X = 20000) = \frac{3}{10} = 0,3$

Bảng phân bố xác suất của X là

X	-9000	20000
\mathbb{P}	0,7	0,3

$\mathbb{E}X = -9000 \cdot 0,7 + 20000 \cdot 0,3 = -300$ đồng.

Vì giá trị trung bình theo xác suất của số tiền nhận được là số âm nên ta không nên tham gia trò chơi này nhiều lần.

Ví dụ 1.4.3. Tính kỳ vọng của biến ngẫu nhiên X liên tục có hàm mật độ xác suất là hàm mật độ đều

$$p(x) = \begin{cases} 0 & \text{nếu } x < a \text{ hoặc } x > b \\ \frac{1}{b-a} & \text{nếu } a \leq x \leq b. \end{cases}$$

Giải. Ta có

$$\begin{aligned}\mathbb{E}X &= \int_{-\infty}^{+\infty} xp(x)dx = \int_{-\infty}^a xp(x)dx + \int_a^b xp(x)dx + \int_b^{+\infty} xp(x)dx \\ &= \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \frac{x^2}{2} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{a+b}{2}.\end{aligned}$$

1.4.2 Phương sai, độ lệch chuẩn

a) Định nghĩa: Phương sai (Deviation, Variance) của biến ngẫu nhiên X (Kí hiệu: $\mathbb{D}(X)$, $\mathbb{D}X$, $\mathbb{V}(X)$, $\text{Var}(X)$) là giá trị được xác định bởi

$$\boxed{\mathbb{D}X = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}(X^2) - (\mathbb{E}X)^2}$$

$$\text{trong đó, } \mathbb{E}(X^2) = \begin{cases} \sum_i x_i^2 p_i & (\text{nếu } X \text{ rời rạc có } \mathbb{P}(X = x_i) = p_i) \\ \int_{-\infty}^{+\infty} x^2 p(x) dx & (\text{nếu } X \text{ liên tục có hàm mật độ xác suất } p(x)). \end{cases}$$

$\sqrt{\mathbb{D}X}$ được gọi là độ lệch chuẩn (Standard deviation) của biến ngẫu nhiên X .

Ví dụ 1.4.4. Xét lại Ví dụ 1.4.1 ở mục trên, ta có $\mathbb{E}X = 1$ và

$$\mathbb{E}(X^2) = \sum_i x_i^2 p_i = 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{4} = \frac{3}{2};$$

$$\mathbb{D}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{3}{2} - 1^2 = \frac{1}{2}.$$

b) Ý nghĩa: Phương sai là trung bình của bình phương sai số giữa X và $\mathbb{E}X$, nó chỉ mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh $\mathbb{E}X$. Phương sai càng nhỏ thì các giá trị của X càng ít phân tán, tức là càng tập trung xung quanh $\mathbb{E}X$.

c) Tính chất:

i) $\mathbb{D}(c) = 0$ (c là hằng số).

ii) $\mathbb{D}(cX) = c^2 \mathbb{D}X$.

iii) Nếu X, Y độc lập thì $\mathbb{D}(X + Y) = \mathbb{D}X + \mathbb{D}Y$.

Chứng minh các tính chất dựa vào công thức xác định phương sai.

d) Chú ý: Có thể dùng công thức sau đây để tính phương sai

$$\mathbb{D}X = \begin{cases} \sum_i (x_i - a)^2 p_i & (\text{nếu } X \text{ rời rạc có } \mathbb{P}(X = x_i) = p_i) \\ \int_{-\infty}^{+\infty} (x - a)^2 p(x) dx & (\text{nếu } X \text{ liên tục có hàm mật độ xác suất } p(x)) \end{cases}$$

trong đó, $a = \mathbb{E}X$.

1.4.3 Trung vị (median)

a) **Định nghĩa:** Trung vị (median) của biến ngẫu nhiên X (K/h: $MedX$) là các số thực a thỏa đồng thời hai điều kiện sau

$$\begin{cases} \mathbb{P}(X < a) \leq 0,5 \\ \mathbb{P}(X > a) \leq 0,5. \end{cases}$$

Từ định nghĩa trên suy ra nếu X là biến ngẫu nhiên có hàm phân phối $F(x)$ thì trung vị là số thực a thỏa mãn phương trình

$$F(a) = \frac{1}{2}. \quad (1.11)$$

Từ (1.11) suy ra nếu X là biến ngẫu nhiên liên tục có hàm mật độ xác suất $p(x)$ thì tìm trung vị a bằng cách giải phương trình sau

$$F(a) = \int_{-\infty}^a p(x)dx = \frac{1}{2}.$$

Ví dụ 1.4.5. Tìm $MedX$ ở Ví dụ 1 1.4.1, Mục 2.1.

X	0	1	2
\mathbb{P}	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Giải. Ta có, $MedX = 1$ vì $\begin{cases} \mathbb{P}(X < 1) = \frac{1}{4} < 0,5 \\ \mathbb{P}(X > 1) = \frac{1}{4} < 0,5. \end{cases}$

b) **Ý nghĩa:** Trung vị là giá trị phân đôi tập giá trị của biến ngẫu nhiên thành hai miền có xác suất đều không vượt quá 50%.

c) **Chú ý:** + Một biến ngẫu nhiên có thể có nhiều trung vị. Khi đó, nếu a_1, a_2 là các trung vị của X thì mọi $a \in [a_1; a_2]$ đều là trung vị của X , hay $MedX \in [a_1; a_2]$.

+ Đối với biến ngẫu nhiên rời rạc, bằng cách dựa vào đồ thị (dạng bậc thang) của hàm phân phối F ta có thể tìm ra trung vị dễ dàng.

Ví dụ 1.4.6. X là biến ngẫu nhiên rời rạc có bảng phân bố xác suất

X	1	2	3	4
\mathbb{P}	0,3	0,2	0,2	0,3

Dễ dàng thấy được 2 và 3 là hai median của X nên suy ra tất cả $a \in [2; 3]$ đều là trung vị.

Thử tìm trung vị của biến ngẫu nhiên X có bảng phân phối xác suất sau đây

X	0	1
\mathbb{P}	$\frac{1}{2}$	$\frac{1}{2}$

1.4.4 Một (Yếu vị)

a) **Định nghĩa:** Một (mode) của biến ngẫu nhiên X (Kí hiệu $ModX$) là giá trị x_0 của X thỏa:

- Tại x_0 , xác suất đạt giá trị lớn nhất trong bảng phân bố xác suất nếu X rời rạc.
- Tại x_0 , hàm mật độ xác suất $p(x)$ đạt giá trị lớn nhất nếu X liên tục.

Ví dụ 1.4.7. Tìm một của biến ngẫu nhiên X có bảng phân phối xác suất

X	-2	0	1	2	5
\mathbb{P}	0,15	0,25	0,2	0,15	0,25

Giải. $ModX = 0$ hoặc $ModX = 5$ do $\mathbb{P}(X = 0) = \mathbb{P}(X = 5) = 0,25$ là giá trị lớn nhất trong bảng phân bố xác suất.

Ví dụ 1.4.8. Tìm kỳ vọng, phương sai, trung vị, một của biến ngẫu nhiên liên tục X có hàm mật độ xác suất như sau

$$p(x) = \begin{cases} 2e^{-2x} & \text{nếu } x \geq 0 \\ 0 & \text{nếu } x < 0. \end{cases}$$

ĐS: $ModX = 0$ ($p(x)$ đạt giá trị lớn nhất là 2 tại $x = 0$); $MedX = \ln 2/2$.

Chú ý, đối với biến ngẫu nhiên rời rạc, một còn gọi là số có khả năng nhất (đã được giới thiệu ở cuối Chương 1).

1.5 Các phân phối xác suất thông dụng

1.5.1 Đối với biến ngẫu nhiên rời rạc

a) **Phân phối Bernoulli:**

Tiến hành một phép thử Bernoulli với xác suất thành công p . Gọi X là biến ngẫu nhiên rời rạc chỉ nhận hai giá trị 0 và 1. Nếu phép thử thành công thì ta viết

$X = 1$, ngược lại ta viết $X = 0$. Khi đó, biến ngẫu nhiên X được gọi là có *phân phối Bernoulli* (Bernoulli distribution) với tham số p , ký hiệu $X \sim A(p)$.

Bảng phân phối xác suất của X là

X	0	1
\mathbb{P}	$1 - p$	p

Tính chất: $\mathbb{E}X = p$, $\mathbb{D}X = pq$ (tính được dễ dàng).

Phân phối Bernoulli gắn liền với phép thử Bernoulli có hai kết quả đối lập, một kết quả quy ước là 1 hay thành công, có xác suất p ; một kết quả quy ước là 0 hay thất bại, có xác suất $q = 1 - p$.

b) Phân phối nhị thức:

Tiến hành n phép thử Bernoulli với xác suất thành công của mỗi phép thử là p . Gọi X là số lần thành công trong n lần thử trên. Khi đó X có phân phối gọi là *phân phối nhị thức* (binomial distribution) với tham số n và p , ký hiệu $X \sim \mathbb{B}(n; p)$.

Bảng phân bố xác suất của X là

X	0	1	...	k	...	n
\mathbb{P}	p_0	p_1	...	p_k	...	p_n

Trong đó $p_k = \mathbb{P}(X = k) = C_n^k p^k q^{n-k}$, $q = 1 - p$, $k = 0, 1, 2, \dots, n$.

Tính chất: Nếu $X \sim \mathbb{B}(n; p)$ thì $\mathbb{E}X = np$ và $\mathbb{D}X = npq$, ($q = 1 - p$).

Ví dụ 1.5.1. Tiến hành ươm 100 hạt giống của một loại hoa quý, xác suất nảy mầm của mỗi hạt giống là 90%. Gọi X là số hạt nảy mầm sau khi ươm. Hãy lập bảng phân bố xác suất của X ?

Giải. Đây là 100 phép thử Bernoulli với xác suất thành công $p = 0,9$.

X là số hạt nảy mầm nên $X \sim \mathbb{B}(100; 0,9)$.

Bảng phân bố xác suất của X là

X	0	1	...	k	...	100
\mathbb{P}	p_0	p_1	...	p_k	...	p_{100}

Trong đó $p_k = \mathbb{P}(X = k) = C_{100}^k (0,9)^k (0,1)^{100-k}$, $k = 0, 1, 2, \dots, 100$.

c) Phân phối Poisson:

Biến ngẫu nhiên X nhận các giá trị $\{0, 1, 2, \dots, n, \dots\}$ được gọi là có *phân phối Poisson* (Poisson distribution) nếu tồn tại số $\lambda > 0$ sao cho

$$\mathbb{P}(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots, n, \dots$$

Khi đó, ta còn gọi X có phân phối Poisson với tham số λ . Ký hiệu $X \sim P(\lambda)$,

Bảng phân bố xác suất của X là

X	0	1	...	k	...	n	...
\mathbb{P}	p_0	p_1	...	p_k	...	p_n	...

trong đó, $p_k = \mathbb{P}(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$, $k = 0, 1, 2, \dots, n, \dots$

Tính chất: Nếu $X \sim P(\lambda)$ thì $\mathbb{E}X = \lambda$ và $\mathbb{D}X = \lambda$.

Bài toán dẫn đến phân phối Poisson:

Gọi X là số lần xuất hiện biến cố A tại các thời điểm ngẫu nhiên trong các thời gian $(t_1; t_2)$ thỏa hai điều kiện

- Số lần xuất hiện biến cố A giữa các khoảng thời gian là độc lập với nhau.
- Số lần xuất hiện biến cố A trong một khoảng thời gian tỉ lệ với độ dài khoảng thời gian đó.

Khi đó, X có phân phối Poisson với $\lambda = c(t_2 - t_1)$, c là cường độ xuất hiện biến cố A .

Ví dụ 1.5.2. Quan sát 5 phút thấy có 15 người ghé vào một đại lý bưu điện. Tính xác suất:

(a) Trong 1 phút có 7 người vào đại lý bưu điện đó?

(b) Trong 3 phút có 16 người vào đại lý bưu điện đó.

Giải. a) Số người trung bình mỗi phút $\lambda = 15/5 = 3$. Gọi X là số người vào đại lý bưu điện đó trong một phút, ta có $X \sim \mathbb{P}(3)$. Vậy $\mathbb{P}(X = 7) = e^{-3} \times 3^7/7!$

b) Số người trung bình trong 3 phút $\lambda = 15/5 \times 3 = 9$ (Kết quả: $e^{-9} \times 9^{16}/16!$).

1.5.2 Đối với biến ngẫu nhiên liên tục

a) Phân phối đều:

Biến ngẫu nhiên X được gọi là có *phân phối đều* (uniform distribution) trên đoạn $[a; b]$ nếu X nhận giá trị trên $[a; b]$ với khả năng đều như nhau. Ký hiệu $X \sim U(a; b)$.

Hàm mật độ xác suất của X là $p(x) = \begin{cases} \frac{1}{b-a} & \text{nếu } a \leq x \leq b \\ 0 & \text{nếu } x < a \text{ hoặc } x > b \end{cases}$ được gọi là *hàm mật độ đều* trên $[a; b]$.

Dễ dàng tìm được hàm phân phối của X là $F(x) = \begin{cases} 0 & \text{nếu } x < a \\ \frac{x-a}{b-a} & \text{nếu } a \leq x \leq b \\ 1 & \text{nếu } x > b \end{cases}$

Tính chất: Nếu $X \sim U(a; b)$ thì $\mathbb{E}X = \frac{a+b}{2}$ và $\mathbb{D}X = \frac{(b-a)^2}{12}$.

b) Phân phối mũ:

Biến ngẫu nhiên X được gọi là có phân phối mũ với tham số $\lambda > 0$ nếu hàm mật độ xác suất của X có dạng

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{nếu } x \geq 0, \\ 0 & \text{nếu } x < 0. \end{cases}$$

Kí hiệu: $X \sim \text{Ex}(\lambda)$, số λ là tham số, thường được gọi là tham số tỉ lệ (rate parameter).

Tính chất: Nếu $X \sim \text{Ex}(\lambda)$ thì $\mathbb{E}X = \frac{1}{\lambda}$ và $\mathbb{D}X = \frac{1}{\lambda^2}$.

c) Phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$:

Biến ngẫu nhiên X được gọi là có *phân phối chuẩn* (phân phối bình thường; normal distribution) với tham số μ và σ^2 , ký hiệu $X \sim \mathcal{N}(\mu, \sigma^2)$, nếu hàm mật độ xác suất của nó là

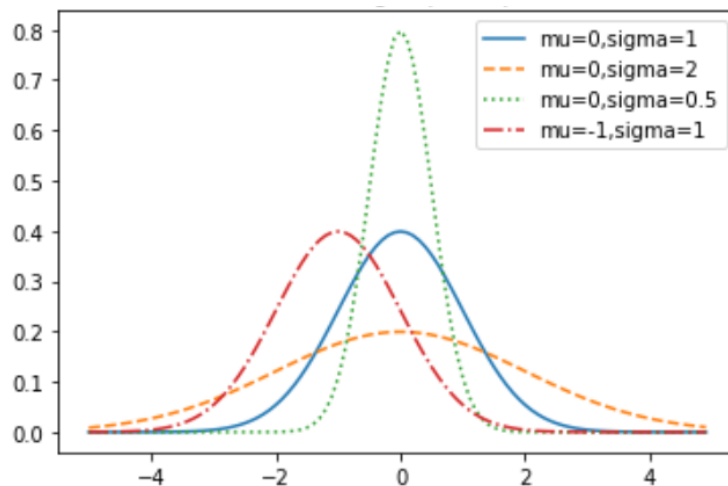
$$\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

• Trường hợp đặc biệt $\mu = 0, \sigma = 1$ ta nói X có *phân phối chuẩn tắc* $\mathcal{N}(0, 1)$.

Nếu $X \sim \mathcal{N}(0, 1)$ thì hàm mật độ xác suất của X lúc này là

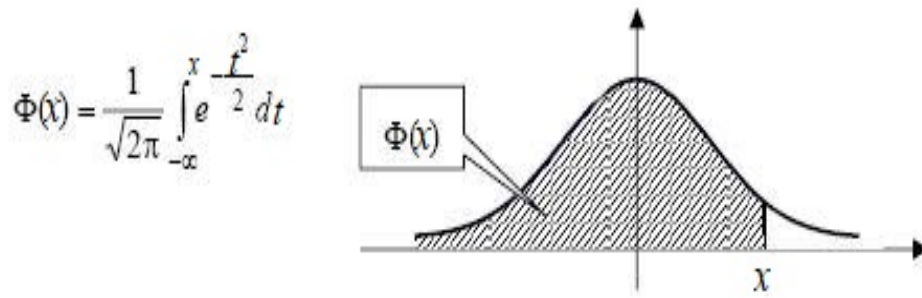
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$$

được gọi là *hàm mật độ Gauss*.



Và hàm phân phối chuẩn tắc $\mathcal{N}(0; 1)$ là

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$



Hàm sau được gọi là *tích phân Laplace*

$$\Phi_1(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt.$$

Mối liên hệ giữa Φ và Φ_1 như sau: $\Phi(x) = \frac{1}{2} + \Phi_1(x)$, $\forall x \in \mathbb{R}$.

• Tính giá trị $\Phi(x)$ (xem phụ lục hướng dẫn bấm máy trang 87, tra bảng Φ trang 92):

Ví dụ, $\Phi(2,16) = 0,9846$; $\Phi(1,96) = 0,9750$; $\Phi(-1,96) = 0,025$; $\Phi(25) = \Phi(180) = 1$.

Dễ dàng chỉ ra mối liên hệ sau: $\Phi(-x) = 1 - \Phi(x)$, $\forall x \in \mathbb{R}$.

Định nghĩa 1.5.3. Cho $Z \sim \mathbb{N}(0; 1)$ và $0 < \alpha < 1$. Ta gọi u_α là *phân vị mức α* của Z nếu

$$\mathbb{P}(Z > u_\alpha) = \alpha.$$

Phân vị được sử dụng thường xuyên trong phân thống kê.

Định lí 1.5.4. Cho $X \sim \mathbb{N}(\mu; \sigma^2)$ ta có

- i) Biến ngẫu nhiên $Y = \frac{X - \mu}{\sigma} \sim \mathbb{N}(0; 1)$;
- ii) $\mathbb{E}X = \mu$; $\mathbb{D}X = \sigma^2$;
- iii) $\mathbb{P}(a \leq X < b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$;
 $\mathbb{P}(X < b) = \Phi\left(\frac{b - \mu}{\sigma}\right)$;
 $\mathbb{P}(X > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right)$.

Vai trò của phân phối chuẩn: Phân phối chuẩn có rất nhiều ứng dụng trong thực tế, nó giữ vai trò vô cùng quan trọng trong lý thuyết xác suất thống kê.

- Các số đo về đặc tính sinh học như chiều cao, cân nặng, huyết áp, nồng độ, sai số trong đo lường vật lý, ... hầu như có phân phối chuẩn.

- Trong xã hội: Lợi tức hàng năm, sản lượng một vụ mùa, đơn vị diện tích, ... tuân theo luật phân phối chuẩn.

d) Phân phối khi (chi) bình phương χ^2 :

Biến ngẫu nhiên X được gọi là có *phân phối khi (chi) bình phương* (Chi-squared distribution) với n bậc tự do, ký hiệu $X \sim \chi^2(n)$, nếu

$$X = X_1^2 + X_2^2 + \dots + X_n^2$$

trong đó X_1, \dots, X_n là các biến ngẫu nhiên độc lập, có phân phối chuẩn tắc $N(0; 1)$.
Hàm mật độ của phân phối khi bình phương là

$$p(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{nếu } x > 0 \\ 0 & \text{nếu trái lại,} \end{cases}$$

với $\Gamma(u) = \int_0^\infty e^{-x} x^{u-1} dx$, $\Gamma(u+1) = u\Gamma(u)$ ($u > 0$), $\Gamma(n+1) = n!$, $\Gamma(1) = 1$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Tính chất: Nếu $X \sim \chi^2(n)$ thì $\mathbb{E}X = n$ và $\mathbb{D}X = 2n$

Phân phối khi bình phương giữ vai trò quan trọng trong kiểm định giả thuyết thống kê.

e) Phân phối Student:

Biến ngẫu nhiên X được gọi là có *phân phối Student* (còn gọi là phân phối t ; Student's t -distribution) với bậc tự do n , ký hiệu $X \sim t(n)$, nếu $X = \frac{U}{\sqrt{\frac{V}{n}}}$, trong

đó $U \sim N(0; 1)$, $V \sim \chi^2(n)$.

Hàm mật độ của phân phối t là hàm

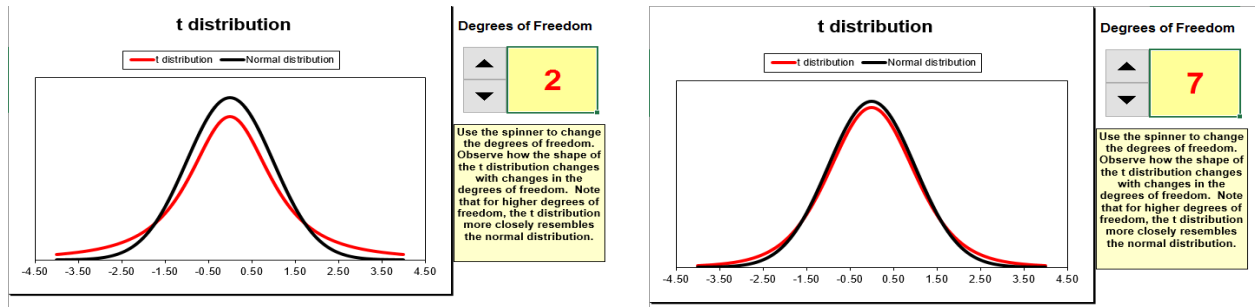
$$p(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Hàm mật độ của phân phối t là hàm đối xứng qua trục tung, dạng đồ thị của nó có dạng hình chuông gần giống như hàm mật độ chuẩn.

Khi bậc tự do $n \geq 30$ thì phân phối t xấp xỉ với phân phối chuẩn tắc.

Tính chất: Nếu $X \sim t(n)$ thì $\mathbb{E}X = 0$ và $\mathbb{D}X = \frac{n}{n-2}$.

Hình sau đây mô phỏng giá trị hàm mật độ xác suất của phân phối t và của phân phối chuẩn gần bằng nhau khi bậc tự do n càng lớn (trong hình, $n = 2, n = 7$):



f) Phân phối F (Phân phối Fisher(n, m)): Biến ngẫu nhiên X được gọi là có phân phối F (phân phối Fisher; Fisher distribution, F -distribution) với n và m bậc tự do, ký hiệu $X \sim F(n, m)$, nếu $X = \frac{U}{n} / \frac{V}{m}$, trong đó U, V là hai biến ngẫu nhiên độc lập và $U \sim \chi^2(n)$, $V \sim \chi^2(m)$.

Hàm mật độ của phân phối F là

$$p(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{x^{\frac{m}{2}-1}}{(1 + \frac{m}{n}x)^{\frac{m+n}{2}}} \quad (\text{với } x > 0).$$

Phân phối t và F được sử dụng nhiều trong thống kê suy đoán. Phân phối t được dùng để giải quyết các bài toán liên quan đến trung bình và tỉ lệ. Phân phối F được dùng để giải quyết các bài toán liên quan đến phương sai. Do đó người ta thiết lập sẵn bảng tính cho những giá trị cần thiết để tiện cho việc tra cứu.

1.5.3 Tính gần đúng xác suất của phân phối nhị thức

Giả sử $X \sim \mathbb{B}(n; p)$, khi đó tại mỗi giá trị k ($0 \leq k \leq n$), ta có thể áp dụng công thức Bernoulli để tính $P_n(X = k)$, hoặc $P_n(X \leq k)$, $P_n(X \geq k)$. Tuy nhiên khi n quá lớn thì việc tính toán trở nên phức tạp, ngay cả việc tính toán trên máy tính cầm tay thì trong nhiều trường hợp kết quả cũng không hiển thị được.

Để khắc phục điều này, trên cơ sở áp dụng các kết quả của các định lý giới hạn người ta đưa ra các công thức về xấp xỉ phân phối nhị thức bởi phân phối chuẩn và phân phối Poisson sau đây.

a) Trường hợp n lớn và p không quá bé (thường xét $n \geq 30$ và $np > 5$):

Có thể tính xấp xỉ phân phối nhị thức bởi phân phối chuẩn $N(\mu, \sigma^2)$ với $\mu \approx np$; $\sigma^2 \approx npq$. Việc tính xác suất thể hiện ở hai dạng sau

+ Xác suất tại một giá trị $\mathbb{P}(X = k) = C_n^k p^k q^{n-k} \approx \frac{1}{\sqrt{npq}} \varphi(x_k)$

trong đó φ là hàm mật độ Gauss $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ và $x_k = \frac{k - np}{\sqrt{npq}}$.

+ Xác suất trong một khoảng giá trị

$$\mathbb{P}(k_1 \leq X < k_2) = \sum_{i=k_1}^{k_2} p_i \approx \Phi\left(\frac{k_2 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{k_1 - np}{\sqrt{npq}}\right)$$

Ví dụ 1.5.5. Một xạ thủ tập bắn trước ngày thi đấu. Xác suất bắn trúng tâm của anh ta ở một khoảng cách cho trước nào đó là 85%. Xạ thủ này bắn liên tiếp 50 viên đạn vào bia. Tính xác suất để anh ta bắn trúng: Đúng 30 viên; Từ 35 đến 45 viên; Ít nhất 20 viên.

Giải. Đây là 50 phép thử Bernoulli với xác suất thành công trong mỗi phép thử là 85%. Do $n > 30, np = 42,5 > 5$, áp dụng công thức xấp xỉ chuẩn ta có

- Xác suất trúng tâm 30 viên đạn là

$$\mathbb{P}(k = 30) = C_{50}^{30} 0,85^{30} 0,15^{20} \approx \frac{1}{\sqrt{npq}} \varphi(x_{30}) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} e^{-\frac{x_{30}^2}{2}} = 1,19 \cdot 10^{-7}.$$

$$\text{Với } x_{30} = \frac{30 - np}{\sqrt{npq}} = -4,95.$$

- Xác suất trúng tâm từ 35 đến 45 viên đạn là

$$\begin{aligned} \mathbb{P}(35 \leq k \leq 45) &= \mathbb{P}(35 \leq k < 46) \approx \Phi\left(\frac{46 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{35 - np}{\sqrt{npq}}\right) \\ &= \Phi(1,39) - \Phi(-2,97) = 0,9177 - 0,0015 = 0,9162. \end{aligned}$$

- Xác suất trúng tâm ít nhất 20 viên đạn là

$$\begin{aligned} \mathbb{P}(k \geq 20) &= 1 - \mathbb{P}(0 \leq k < 20) \approx 1 - \left[\Phi\left(\frac{20 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{0 - np}{\sqrt{npq}}\right) \right] \\ &= 1 - [\Phi(-8,91) - \Phi(-16,83)] = 1 - [0 - 0] = 1. \end{aligned}$$

b) Trường hợp n lớn và p bé (tức $np \approx npq$):

Trường hợp này ta dùng xấp xỉ Poisson $P(\lambda)$, với $\lambda \approx np$. Ta có công thức như sau

$$\mathbb{P}(X = k) = C_n^k p^k q^{n-k} \approx \frac{e^{-\lambda} \lambda^k}{k!}$$

Ví dụ 1.5.6. Xác suất một hạt thóc giống bị lép là 6%. Chọn ngẫu nhiên 100 hạt thóc. Tính xác suất có: Đúng 30 hạt lép, có từ 50 đến 52 hạt lép?

Giải. Đây là 100 phép thử Bernoulli với xác suất thành công trong mỗi phép thử là 6%. Do $n > 30, p = 0,06$ rất bé, áp dụng công thức xấp xỉ Poisson với $\lambda = np = 6$, ta có

- Xác suất có đúng 30 hạt lép là

$$\mathbb{P}(k = 30) \approx \frac{e^{-\lambda} \lambda^{30}}{30!} = 2,07 \cdot 10^{-12}.$$

- Xác suất có từ 50 đến 52 hạt lép là

$$\begin{aligned}\mathbb{P}(50 \leq k \leq 52) &= \mathbb{P}(k = 50) + \mathbb{P}(k = 51) + \mathbb{P}(k = 52) \\ &\approx \frac{e^{-\lambda} \lambda^{50}}{50!} + \frac{e^{-\lambda} \lambda^{51}}{51!} + \frac{e^{-\lambda} \lambda^{52}}{52!} \\ &= 7,4519 \cdot 10^{-29}.\end{aligned}$$

BÀI TẬP CHƯƠNG 1

Công thức định nghĩa xác suất, cộng, nhân xác suất:

1.1. Một người dùng các chữ số từ 0 đến 9 để đặt mật khẩu cho máy tính cá nhân của mình. Mật khẩu gồm 6 chữ số. Tính xác suất để:

- a) Mật khẩu có 6 chữ số giống nhau.
- b) Mật khẩu có 6 chữ số khác nhau hoàn toàn.
- c) Mật khẩu có một chữ số 3.

HD. $n(\Omega) = 10^6$.

1.2. Một văn phòng có 3 máy in cùng hoạt động. Khả năng có sự cố trong một quý của mỗi máy in lần lượt tương ứng là 15%, 20% và 10%. Trong một quý, hãy tính khả năng:

- a) Cả 3 máy in đều gặp sự cố.
- b) Cả 3 máy in đều không gặp sự cố.
- c) Có ít nhất 1 máy in không gặp sự cố.
- d) Không quá hai máy in gặp sự cố.

HD. Dùng công thức nhân xác suất cho các biến cố độc lập.

1.3. Một hệ thống có hai bộ lọc bảo mật hoạt động độc lập. Xác suất để bộ lọc thứ nhất phát hiện một cuộc tấn công là 0,9 và xác suất để bộ lọc thứ hai phát hiện một cuộc tấn công là 0,8. Tính xác suất cả hai bộ lọc đều phát hiện được một cuộc tấn công.

HD. Dùng công thức nhân xác suất cho 2 biến cố độc lập.

Công thức xác suất đầy đủ, công thức Bayes:

1.4. Một công ty phần mềm có 3 loại máy chủ: máy chủ loại 1 chiếm 50%, loại 2 chiếm 30%, và loại 3 chiếm 20%. Xác suất máy chủ loại 1 gặp sự cố là 0,1; loại 2 là 0,2 và loại 3 là 0,3. Tính xác suất một máy chủ bất kỳ của công ty gặp sự cố.

ĐS. 0,17.

1.5. Một hệ thống chống virus có thể phát hiện 98% virus thật và báo động nhầm 1% khi không có virus. Giả sử xác suất xuất hiện virus trong máy tính là 0,5%. Nếu hệ thống báo có virus, tính xác suất máy tính thực sự bị nhiễm virus.

HD. Dùng công thức Bayes. Gọi V : "máy tính bị nhiễm vi rút". B : "Máy báo động phát hiện virus". $P(V|B) = \frac{P(V)P(B|V)}{P(B)}$. ĐS: 0,327.

- 1.6. * Một hệ thống giám sát mạng có hai phương pháp kiểm tra xâm nhập: phương pháp A và B. Phương pháp A phát hiện 95% xâm nhập với báo động nhầm 5%. Phương pháp B phát hiện 90% xâm nhập với báo động nhầm 2%. Xác suất xâm nhập thực sự là 0,1%. Tính xác suất có xâm nhập thực sự nếu cả hai phương pháp đều báo động.

HD. V : Hệ thống có thâm nhập; A, B : Phương pháp A, B báo động (tương ứng); $P(V) = 0,001$. Tính $P(A \cap B|V)$, $P(A \cap B|\bar{V})$; Từ đó tính $P(A \cap B)$. Áp dụng công thức Bayes: $P(V|A \cap B) = \frac{P(V)P(A \cap B|V)}{P(A \cap B)} = 0,461$.

Công thức xác suất nhị thức

- 1.7. Theo thống kê, tỉ lệ sinh viên năm cuối ở một trường đại học biết thành thạo về tin học văn phòng là 90%. Chọn ngẫu nhiên 10 sinh viên năm cuối của trường để kiểm tra trình độ tin học văn phòng. Tính xác suất để trong 10 sinh viên này thì có:

- Không sinh viên nào thành thạo tin học văn phòng.
- Từ 5 đến 7 sinh viên thành thạo tin học văn phòng.
- Ít nhất 2 sinh viên thành thạo tin học văn phòng.
- Số sinh viên thành thạo tin học văn phòng có khả năng nhất.

HD. Dùng công thức xác suất nhị thức với $n = 10, p = 0,9, q = 0,1$.

- 1.8. Tín hiệu thông tin được phát 3 lần với xác suất thu được của mỗi lần là 0,4.

- Tìm xác suất để nguồn thu nhận được thông tin đó.
- Nếu muốn xác suất thu được thông tin lên 0,9 thì phải phát bao nhiêu lần.

HD. a) $n = 3, p = 0,4$. Để nguồn thu nhận được thông tin thì ít nhất một lần trong 3 lần nhận được thông tin, tính $P(k \geq 1)$; b) $P(k \geq 1) = 0,9$, suy ra $n_{\text{mới}}$.

Biến ngẫu nhiên rời rạc, biến ngẫu nhiên liên tục và hàm mật độ xác suất:

- 1.9. Gieo một con xúc sắc cân đối đồng chất. Gọi X là số chấm ở mặt trên con xúc sắc

- Lập bảng phân bố xác suất của X .
- Viết biểu thức hàm phân phối xác suất $F(x)$. Vẽ đồ thị của $F(x)$.

- 1.10. Một lập trình viên có thể làm xong một chương trình trong 1, 2, hoặc 3 ngày với xác suất lần lượt là 0,3; 0,5 và 0,2. Tính kỳ vọng và phương sai số ngày cần để hoàn thành một chương trình.

HD. Lập bảng phân bố xác suất về số ngày X làm xong một chương trình, rồi áp dụng công thức tính kỳ vọng, phương sai.

1.11. Một người lập trình viết mã có 3 cấp độ lỗi: nhỏ (10%), vừa (20%) và lớn (70%). Chi phí sửa một lỗi mã lần lượt là 10, 20 và 50 đô la.

a) Tính chi phí trung bình khi sửa một lỗi mã (kỳ vọng).

b) Tính phương sai của biến ngẫu nhiên chi phí sửa một lỗi mã.

HD. Lập bảng phân bố xác suất về chi phí X (đô la) để sửa một lỗi, rồi áp dụng công thức tính kỳ vọng, phương sai.

1.12. Tuổi thọ của một loại côn trùng nào đó là biến ngẫu nhiên X (đơn vị: tháng) có hàm mật độ xác suất $p(x) = \begin{cases} Cx^2(4-x) & \text{nếu } 0 \leq x \leq 4, \\ 0 & \text{nếu trái lại.} \end{cases}$

a) Tìm hằng số C .

b) Tính xác suất sao cho côn trùng này sống lâu hơn 0,5 tháng tuổi.

c) Tính tuổi thọ trung bình của loại côn trùng này.

HD. a) Áp dụng công thức $1 = \int_{-\infty}^{\infty} p(x)dx$ để tìm C . b) $\mathbb{P}(X \geq 0,5)$. c) Tính $EX = \int_{-\infty}^{\infty} xp(x)dx$.

Một số phân phối xác suất thông dụng

1.13. Một lập trình viên thử nghiệm một hệ thống với 10 lần yêu cầu. Xác suất mỗi yêu cầu thành công là 0,8. Tính xác suất có 8 yêu cầu thành công.

HD. Công thức xác suất nhị thức với $k = 8$ và n, p đã có.

1.14. Một hệ thống phát hiện xâm nhập mạng có xác suất phát hiện thành công một cuộc tấn công là 0.9. Giả sử có 5 cuộc tấn công, tính xác suất hệ thống phát hiện ít nhất 4 cuộc.

HD. $n = ?$, $p = ?$, tính $P(k \geq 4)$.

1.15. Giả sử thời gian phản hồi của một máy chủ (tính bằng giây) là một biến ngẫu nhiên liên tục có phân phối đều từ 2 giây đến 10 giây. Tính xác suất rằng thời gian phản hồi nhỏ hơn 5 giây.

HD. Hàm mật độ xác suất của phân phối đều $p(x)$ với $a = 2, b = 10$. $P(< 2) = \int_2^? p(x)dx = \dots = 0,375$.

1.16. Giả sử thời gian giữa hai yêu cầu đến một máy chủ là biến ngẫu nhiên X có phân phối mũ với trung bình là $\mathbb{E}X = 3$ giây. Tính xác suất để thời gian giữa hai yêu cầu ít hơn 2 giây.

HD. $\mathbb{E}X = \frac{1}{\lambda}$ suy ra $\lambda = ?$. Hàm mật độ xác suất của phân phối đều $p(x)$. Tính $P(X < 2) = \int_?^? p(x)dx = \dots = 0,487$.

- 1.17. Giả sử rằng thời gian xử lý của một thuật toán là biến ngẫu nhiên có phân phối theo phân phối chuẩn với trung bình $\mu = 50\text{ms}$ và độ lệch chuẩn $\sigma = 5\text{ms}$. Tính xác suất để thời gian xử lý nằm trong khoảng từ 45 ms đến 55 ms.

HD. Áp dụng công thức tính xác suất của phân phối chuẩn để tính $P(45 \leq X \leq 55)$.

- 1.18. Trọng lượng một con bò trong một đàn (đơn vị kg) là biến X phân phối chuẩn $N(250; 40^2)$. Tính xác suất để khi chọn ngẫu nhiên một con ra cân thì trọng lượng X :

- a) Nhẹ hơn 175kg.
- b) Nặng hơn 300kg.
- c) Trong khoảng từ 260kg đến 270 kg.

HD. $EX = 250$; $DX = 40^2$. Dùng tính chất của phân phối chuẩn.

CHƯƠNG 2

LÝ THUYẾT MẪU VÀ BÀI TOÁN ƯỚC LƯỢNG THAM SỐ

2.1 Đám đông và mẫu

2.1.1 Đám đông và đặc tính nghiên cứu

Trong thực tế, nhiều khi cần quan tâm đến một số đặc điểm định tính hoặc định lượng của các phần tử thuộc về một tập hợp nào đó, chẳng hạn tuổi thọ của một thiết bị tin học, giá thành bán lẻ của một mặt hàng công nghệ nào đó tại các thời điểm khác nhau, tốc độ đường truyền internet, tỉ lệ cá bệnh trong ao, ... Tập hợp các phần tử cần nghiên cứu này được gọi là *đám đông*, *quần thể* hay *tổng thể* (Statistical population), kí hiệu là \mathcal{P} .

Việc tiến hành thu thập thông tin trên các phần tử của đám đông \mathcal{P} được gọi là *quan sát*. Đặc điểm cần quan tâm đó thay đổi từ phần tử này sang phần tử khác khi ta thực hiện các quan sát ngẫu nhiên trên một số phần tử của \mathcal{P} . Đặc điểm thay đổi đó của đám đông \mathcal{P} được coi như một biến ngẫu nhiên, ký hiệu X và được gọi là *biến ngẫu nhiên gốc đám đông* hay *biến ngẫu nhiên của đám đông*.

Đặc điểm của đám đông thường được nghiên cứu dưới hai phương diện:

- ♠ *Phương diện định lượng*: Khi ta cần quan tâm đến các giá trị về lượng của biến ngẫu nhiên như trọng lượng, năng suất, tuổi thọ, nồng độ chất, sức gió, nhiệt độ, ... và ta thường quan tâm đến hai đặc trưng
 - Kỳ vọng: $\mathbb{E}X = \mu$ là đặc trưng giá trị trung bình của đặc điểm định lượng cần quan tâm trên đám đông \mathcal{P} .
 - Phương sai: $\mathbb{D}X = \sigma^2$ là đặc trưng cho mức độ biến động giá trị của đặc điểm định lượng cần quan tâm trên đám đông \mathcal{P} .
- ♠ *Phương diện định tính*: Khi ta cần quan tâm đến một tính chất A nào đó trên đám đông \mathcal{P} . Các phần tử của \mathcal{P} hoặc có tính chất A hoặc không có tính chất A như: chất lượng sản phẩm, sự nảy mầm của một giống lúa, tỉ lệ chất độc hại trong nguồn nước, tỉ lệ người ủng hộ một ứng cử viên, ... Giá trị mà biến ngẫu nhiên X có thể

nhận được là

$$X = \begin{cases} 1 & \text{khi phần tử có tính chất A,} \\ 0 & \text{khi phần tử không có tính chất A.} \end{cases}$$

và ta thường quan tâm đến xác suất $\mathbb{E}X = p$.

2.1.2 Khái niệm mẫu và cách chọn mẫu

a) *Khái niệm mẫu:*

Vì những lý do như thời gian, chi phí tốn kém, làm hủy hoại đám đông,... ta không thể quan sát hết các phần tử của đám đông \mathcal{P} . Chính vì vậy, người ta chỉ lấy ra một số phần tử đại diện cho \mathcal{P} và nghiên cứu trên tập các phần tử này, tập hợp các phần tử đại diện cho \mathcal{P} được gọi là *mẫu thống kê* (Statistical sample; gọi tắt là mẫu). Phương pháp nghiên cứu trên mẫu đại diện cho đám đông đó được gọi là *phương pháp mẫu* và cách thức thực hiện quá trình lấy mẫu được gọi là *phương pháp lấy mẫu*.

Khi cần quan tâm đến đặc điểm là biến ngẫu nhiên X gốc đám đông \mathcal{P} , ta chọn ra mẫu có n phần tử, trong đó việc chọn phần tử thứ i là quá trình thực hiện một phép thử rút ngẫu nhiên một phần tử của \mathcal{P} , giá trị ngẫu nhiên này được gán cho biến ngẫu nhiên X_i . Với cách chọn này các biến ngẫu nhiên X_i độc lập với nhau và có cùng luật phân phối với biến ngẫu nhiên X . Mẫu này được gọi là mẫu ngẫu nhiên có kích thước n của đám đông \mathcal{P} , ký hiệu (X_1, X_2, \dots, X_n) . Giả sử tại lần lấy mẫu thứ i , giá trị mà X_i nhận được là x_i , khi đó bộ số (x_1, x_2, \dots, x_n) được gọi là mẫu cụ thể kích thước n (cỡ n).

Ví dụ 2.1.1. Thống kê về chiều cao X (đơn vị: *cm*) của 5 sinh viên của trường được chọn ngẫu nhiên để đo. Mẫu ngẫu nhiên (X_1, X_2, \dots, X_5) ; (X_i là b.n.n chỉ chiều cao của sinh viên thứ i , $i = \overline{1, 5}$). Giả sử một mẫu cụ thể thu được từ mẫu ngẫu nhiên trên là $(165, 172, 155, 163, 158)$.

b) *Cách chọn mẫu:*

Muốn kết luận thống kê rút ra sau khi khảo sát không bị sai lệch có hệ thống thì mẫu quan sát phải phản ánh trung thực đám đông \mathcal{P} , thông tin thu được phản ánh càng gần với tính chất của đám đông (tính chất đại diện cao). Mẫu không thể thiên về chọn các cá thể tạm gọi là "tốt" tức là cho các giá trị lớn hơn trung bình, hoặc "xấu" tức là thiên về các giá trị nhỏ hơn trung bình.

Có rất nhiều cách chọn mẫu vì việc chọn mẫu không những phải thỏa mãn yêu cầu chính là không thiên lệch mà còn phải phù hợp với điều kiện chuyên môn. Thuần túy về mặt thống kê có một số phương pháp chọn mẫu như:

• **Lấy mẫu ngẫu nhiên đơn giản (không định hướng):**

Là phương pháp lấy mẫu thỏa mãn các điều kiện: Mỗi lần chỉ được chọn một phần tử từ đám đông, khả năng được chọn của tất cả các phần tử trong đám đông là như nhau. Có hai cách thức tiến hành chọn đó là chọn không hoàn lại và có hoàn lại. Tuy nhiên khi đám đông có kích thước lớn nhiều so với cỡ mẫu thì hai cách chọn này có thể xem là như nhau.

Phương pháp lấy mẫu đơn giản có tính chất đại diện cho đám đông cao, tuy nhiên khó thực hiện và tốn thời gian, kinh phí khi chọn mẫu cỡ lớn.

• **Lấy mẫu ngẫu nhiên có định hướng:**

- *Lấy mẫu theo nhóm:* Là phương pháp chia đám đông thành các nhóm thuần nhất, từ mỗi nhóm này ta lấy ra một mẫu ngẫu nhiên đơn giản với một kích thước tương ứng. Tập hợp tất cả các phần tử thu được từ các mẫu ngẫu nhiên đơn giản lập nên mẫu ngẫu nhiên theo nhóm.
- *Lấy mẫu theo chùm:* Là phương pháp chia đám đông thành nhiều chùm (đám đông con), sao cho giữa các chùm có sự đồng đều về quy mô. Từ các chùm đó ta lấy ra một mẫu ngẫu nhiên đơn giản. Tập hợp tất cả các phần tử thu được từ các mẫu ngẫu nhiên đơn giản lập nên mẫu ngẫu nhiên theo chùm.

Phương pháp này dễ quy hoạch, tiết kiệm được thời gian và kinh phí nhưng sai số chọn mẫu cao hơn các phương pháp trên.

Ví dụ 2.1.2. Muốn tìm hiểu về thu nhập trong một năm của toàn bộ viên chức trong tỉnh Đồng Tháp. Ta không thể tìm hiểu thu nhập của tất cả viên chức trong tỉnh mà chọn ra một mẫu ngẫu nhiên kích thước n (giả sử $n = 500$ người) để tìm hiểu. Ta có một số cách chọn mẫu như sau:

Cách 1: Chọn mẫu ngẫu nhiên đơn giản. Chọn ngẫu nhiên (không hoàn lại) không phân biệt nghề nghiệp, tuổi tác, nơi ở, ... lần lượt cho đến khi nào đủ 500 người.

Cách 2: Chọn mẫu ngẫu nhiên theo nhóm. Chia đám đông này theo từng cơ cấu ngành nghề (giáo dục, quốc phòng, y tế, kinh doanh, ...). Khi đó trong mỗi ngành nghề có sự thuần nhất về mức lương (nếu có sự sai khác về thu nhập thì chủ yếu là do thâm niên và chức vụ công tác). Tập hợp lại các mẫu ngẫu nhiên đơn giản được chọn theo cơ cấu ngành nghề (mỗi ngành nghề ta có thể chọn số lượng khác nhau, ít hoặc nhiều hơn tùy thuộc vào số lượng của ngành nghề đó) ta được mẫu ngẫu nhiên cỡ 500.

Cách 3: Chọn mẫu ngẫu nhiên theo chùm. Chia đám đông này theo các huyện trong tỉnh. Giữa các huyện có sự tương đối đồng đều về quy mô (đầy đủ các thành phần). Tập hợp các mẫu ngẫu nhiên đơn giản ta được mẫu ngẫu nhiên cỡ 500.

2.1.3 Cách biểu diễn mẫu, hàm phân phối mẫu

Cho một mẫu ngẫu nhiên kích thước n $\{X_1, X_2, \dots, X_n\}$ gồm các biến ngẫu nhiên độc lập cùng phân phối và $\{x_1, x_2, \dots, x_n\}$ là một mẫu cụ thể thu được từ mẫu ngẫu nhiên trên.

★ Bảng tần số và bảng tần suất

Ta thấy rằng trong mẫu cụ thể có thể có một số kết quả trùng nhau. Tiến hành thu gọn mẫu này lại như sau: Đếm các kết quả khác nhau và sắp xếp theo thứ tự từ bé đến lớn, giả sử các kết quả khác nhau là $x_1 < x_2 < \dots < x_k, k \leq n$. Gọi n_i là *tần số* (frequency, số lần xuất hiện) của kết quả x_i , ta có *bảng tần số* (bảng tần số thu gọn) của mẫu như sau

X_i	x_1	x_2	...	x_k
Tần số n_i	n_1	n_2	...	n_k

trong đó, $n_1 + n_2 + \dots + n_k = n$.

Tương tự ta có bảng sau gọi là *bảng tần suất* của mẫu như sau

X_i	x_1	x_2	...	x_k
Tần suất f_i	$\frac{n_1}{n}$	$\frac{n_2}{n}$...	$\frac{n_k}{n}$

Chú ý, dễ thấy rằng nếu n giá trị của mẫu cụ thể đều khác nhau thì các tần số của mỗi giá trị đều bằng 1 và tần suất đều bằng $\frac{1}{n}$.

★ Bảng tần số ghép lớp

Người ta còn có cách biểu diễn mẫu số liệu thành bảng tần số dưới dạng lớp $(a_i; a_{i+1})$, độ dài của mỗi lớp là $h = a_{i+1} - a_i$. Tần số của mỗi lớp là số các kết quả thuộc vào lớp đó. Khi đó ta có *bảng tần số ghép lớp* như sau:

Lớp	$[a_1; a_2]$	$(a_2; a_3]$	$(a_3; a_4]$...	$(a_{k-1}; a_k]$
Tần số n_i	n_1	n_2	n_4	...	n_k

Cách chia lớp cho mẫu số liệu thực nghiệm được thực hiện như sau (xem [1]):

◇ Số lớp chia l : $l = \min\{k \in \mathbb{N} : 2^k \geq n\}$.

Ví dụ: Nếu $n = 55$ thì chọn $l = 6$ (vì $2^5 = 32 < 55$; $2^6 = 64 > 55$).

◇ Độ dài mỗi lớp: $h = \frac{x_{\max} - x_{\min}}{l}$.

◇ Trong hai lớp liên nhau $a_{i-1} \rightarrow a_i$ và $a_i \rightarrow a_{i+1}$ thì giá trị $x = a_i$ thuộc lớp thứ nhất.

Khi thực hành tính toán người ta thường lấy trung điểm của mỗi lớp (là trung bình cộng của giá trị đầu và giá trị cuối của mỗi lớp) làm phần tử đại diện cho lớp đó.

Ví dụ 2.1.3. Để khảo sát về chiều cao của sinh viên một trường dạy nghề người ta chọn ngẫu nhiên 30 sinh viên để đo chiều cao. Gọi X (đơn vị: cm) là biến ngẫu nhiên chỉ chiều cao của sinh viên, kết quả đo được như sau

155 161 170 152 155 160 162 160 155 170
 152 160 152 153 153 155 160 153 165 162
 165 152 162 168 152 159 159 163 153 163

Hãy lập bảng tần số của mẫu số liệu trên theo hai dạng.

Giải. ◇ Bảng tần số thu gọn của mẫu:

X	152	153	155	159	160	161	162	163	165	168	170
Tần số n_i	5	4	4	2	4	1	3	2	2	1	2

◇ Bảng tần số ghép lớp:

Số lớp $l = 5$; Độ dài mỗi lớp $h = \frac{170 - 152}{5} = 3,6$.

X	$[152; 155,6]$	$(155,6; 159,2]$	$(159,2; 162,8]$	$(162,8; 166,4]$	$(166,4; 170]$
n_i	13	2	8	4	3

★ Hàm phân phối mẫu

X là biến ngẫu nhiên gốc đám đông có hàm phân phối xác suất $F(x)$ chưa biết. Khi ta thực hiện n quan sát, gọi $F_n(x) = \frac{m_x}{n}$ là *hàm phân phối mẫu*, với m_x là số quan sát có giá trị x_i bé hơn x ($i = \overline{1, n}$).

Tính chất của hàm phân phối mẫu:

+ $0 \leq F_n(x) \leq 1$.

+ $F_n(x)$ đơn điệu tăng và liên tục bên trái.

Khi cỡ mẫu lớn thì $F_n(x)$ càng gần với phân phối xác suất của biến ngẫu nhiên X . Khi n đủ lớn ta có thể dùng $F_n(x)$ thay cho $F(x)$ chưa biết hoặc dựa vào $F_n(x)$ để dự đoán về dạng của $F(x)$ và đưa ra các đặc trưng liên quan.

Ví dụ 2.1.4. Bảng tần số từ mẫu thống kê điểm của 40 sinh viên như sau

X	4	5	6	7	8
n_i	5	10	12	8	5

và bảng tần suất

X	4	5	6	7	8
f_i	$\frac{5}{40}$	$\frac{10}{40}$	$\frac{12}{40}$	$\frac{8}{40}$	$\frac{5}{40}$

Khi đó, ta có hàm phân phối mẫu $F_n(x)$

$$F_n(x) = \begin{cases} 0 & \text{với } x \leq 4 \\ \frac{5}{40} & \text{với } 4 < x \leq 5 \\ \frac{15}{40} & \text{với } 5 < x \leq 6 \\ \frac{27}{40} & \text{với } 6 < x \leq 7 \\ \frac{35}{40} & \text{với } 7 < x \leq 8 \\ 1 & \text{với } x > 8. \end{cases}$$

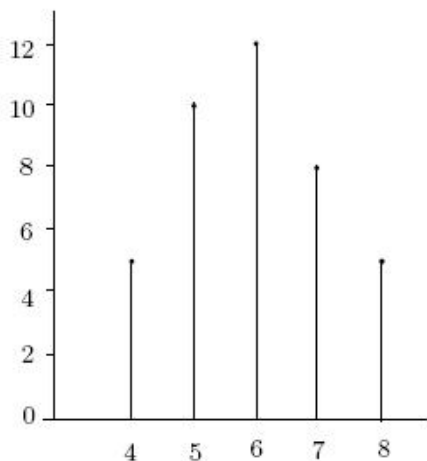
★ Đa giác tần suất và tổ chức đồ

+ Đối với số liệu chưa ghép lớp:

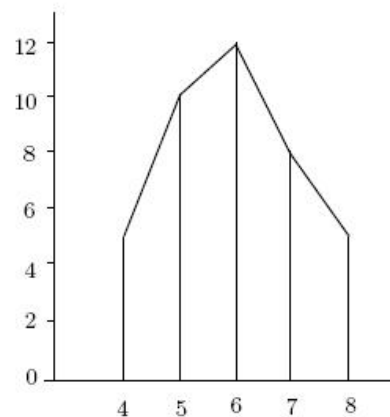
- Chấm trên mặt phẳng các điểm $(x_i, n_i), i = 1, 2, \dots, n$.
- Nối các điểm $(x_i, 0)$ với các điểm (x_i, n_i) ta được *biểu đồ tần số hình gậy*.
- Nối liên tiếp các điểm (x_i, n_i) và (x_{i+1}, n_{i+1}) ta được *biểu đồ đa giác tần số*.

Tương tự đối với biểu đồ đa giác tần suất.

Ví dụ 2.1.5. Minh họa số liệu thống kê điểm



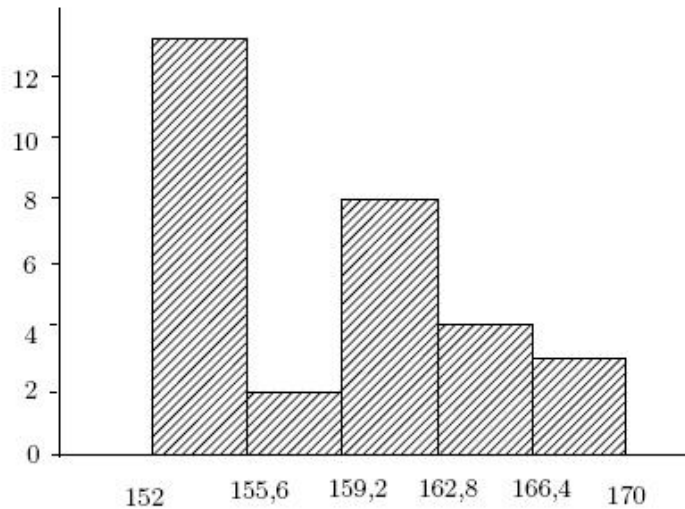
Biểu đồ tần số hình gậy



Biểu đồ đa giác tần số

+ Đối với số liệu đã ghép lớp:

- Trên mỗi lớp ta dựng hình chữ nhật có chiều cao bằng tần số (hay tần suất) tương ứng với lớp đó.
- Tô đậm hoặc kẻ chéo bằng các đường song song các hình chữ nhật này ta thu được *tổ chức đồ tần số* (hay tổ chức đồ tần suất).

Ví dụ 2.1.6. Minh họa số liệu thống kê chiều cao sinh viên

Biểu đồ đa giác tần số

2.2 Các đặc trưng của mẫu**2.2.1 Các đặc trưng của mẫu**

Cho mẫu ngẫu nhiên kích thước n (X_1, X_2, \dots, X_n) (của biến ngẫu nhiên X được quan sát từ đám đông \mathcal{P}), ta có các đặc trưng mẫu *trung bình mẫu* \bar{X} , *phương sai mẫu* \hat{S}^2 , *phương sai mẫu hiệu chỉnh* S^2 , *tỉ lệ mẫu* được định nghĩa như sau

- Theo đặc điểm định lượng:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i;$$

$$(\hat{S})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \stackrel{\text{chứng minh đơn giản}}{=} \overline{X^2} - (\bar{X})^2 \quad (\text{trong đó } \overline{X^2} = \frac{1}{n} \sum_{i=1}^n X_i^2);$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} (\hat{S})^2.$$

- Theo đặc điểm định tính: $X = \{0; 1\}$, $X = 1$ nếu biến cố A xuất hiện ("thành công"), $X = 0$ nếu biến cố A không xuất hiện ("thất bại").

Tỉ lệ mẫu $F = \frac{1}{n} \sum_{i=1}^n X_i = \frac{m}{n}$ (m là số phần tử có tính chất A), trong đó X_i bằng 0 hoặc bằng 1.

Ta có X_i là các biến ngẫu nhiên nên $\bar{X}, (\hat{S})^2, S^2, F$ cũng là các biến ngẫu nhiên.

Trong thực hành tính toán, với mẫu số liệu cụ thể kích thước n $\{x_1, x_2, \dots, x_n\}$ ta tiến hành thu gọn số liệu và biểu diễn mẫu thực nghiệm dưới dạng bảng tần số

X_i	x_1	x_2	\dots	x_k
Tần số n_i	n_1	n_2	\dots	n_k

a) Các số đặc trưng của mẫu cụ thể:

- Đối với mẫu định lượng:

- Trung bình mẫu thực nghiệm (mean):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{1}{n} (x_1 n_1 + x_2 n_2 + \dots + x_k n_k)$$

- Phương sai mẫu chưa hiệu chỉnh được ký hiệu là \hat{s}^2 , được tính bởi một trong hai công thức tương đương sau đây

$$\hat{s}^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_i = \overline{x^2} - (\bar{x})^2.$$

trong đó, $\overline{x^2} = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i$.

- Phương sai mẫu đã hiệu chỉnh (gọi tắt là phương sai mẫu; sample variance):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i - \bar{x})^2 n_i = \frac{n}{n-1} \hat{s}^2 = \frac{n}{n-1} [\overline{x^2} - (\bar{x})^2]$$

$s = \sqrt{s^2}$ được gọi là *độ lệch chuẩn mẫu* (standard deviation) hay độ lệch tiêu chuẩn, độ lệch tiêu chuẩn mẫu.

- Đối với mẫu định tính:

- Tỷ lệ mẫu thực nghiệm

$$f = \frac{m}{n} \quad (m \text{ là số phần tử có tính chất A trong mẫu}).$$

Nhận xét 2.2.1. - Nếu mẫu số liệu được biểu diễn dưới dạng bảng tần số ghép lớp thì ta lấy phần tử đại diện x_i cho lớp thứ i là trung điểm của lớp đó.

- Phương sai mẫu đánh giá mức độ phân tán của các số liệu của mẫu xung quanh giá trị trung bình. Phương sai mẫu càng lớn thì sự phân tán của số liệu càng lớn và ngược lại.

Ví dụ 2.2.2. Cho mẫu số liệu của biến ngẫu nhiên X (đơn vị kg) về trọng lượng của 15 SV nam được chọn ngẫu nhiên để đo, có kết quả được cho bởi bảng tần số

$X(kg)$	47	49	52	55	60
Tần số n_i	1	3	4	5	2

Hãy tính các số đặc trưng của mẫu số liệu trên.


Giải. Ta có $n = 15$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^5 x_i n_i = 53,1333;$$

$$\hat{s}^2 = \frac{1}{n} \sum_{i=1}^5 (x_i - \bar{x})^2 n_i = 13,7156 \Rightarrow \hat{s} = \sqrt{\hat{s}^2} = 3,7035.$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^5 (x_i - \bar{x})^2 n_i = 14,6952 \Rightarrow s = \sqrt{s^2} = 3,8334.$$

b) Hướng dẫn thực hành tính các số đặc trưng mẫu bằng máy tính cầm tay:

Xem phần Phụ lục trang 87. 

2.2.2 Phân phối của các đặc trưng mẫu

Cho $\{X_1, X_2, \dots, X_n\}$, là mẫu ngẫu nhiên kích thước n quan sát từ biến ngẫu nhiên X gốc đám đông \mathcal{P} , trong đó X_1, X_2, \dots, X_n là n b.n.n độc lập cùng phân phối. Giả sử b.n.n X có kỳ vọng $\mathbb{E}X = \mu$ và phương sai $\mathbb{D}X = \sigma^2$, tỉ lệ cá thể có tính chất A là $\mathbb{P}(A) = p$, theo luật số lớn ta có các kết quả sau:

a) Các số đặc trưng của các đặc trưng mẫu:

- Đối với trung bình mẫu: $\mathbb{E}(\bar{X}) = \mu; \quad \mathbb{D}(\bar{X}) = \frac{\sigma^2}{n}.$
- Đối với phương sai mẫu: $\mathbb{E}(\hat{S}^2) = \frac{(n-1)\sigma^2}{n}.$
- Đối với phương sai mẫu hiệu chỉnh: $\mathbb{E}(S^2) = \sigma^2.$
- Đối với tỉ lệ mẫu: $\mathbb{E}(F) = p; \quad \mathbb{D}(F) = \frac{p(1-p)}{n}.$

b) Phân phối của các đặc trưng mẫu: (xem [1])

Các thống kê là các hàm của các b.n.n nên cũng là b.n.n, do đó ta có thể khảo sát

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i; \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

như các biến ngẫu nhiên thông thường.

b1) Trường hợp biến ngẫu nhiên X có phân phối chuẩn $\mathbb{N}(\mu; \sigma^2)$:

+ Khi σ^2 đã biết, ta có

$$\bar{X} \sim \mathbb{N}(\mu, \frac{\sigma^2}{n}) \quad \text{khi đó} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim \mathbb{N}(0; 1).$$

$$\frac{n\hat{S}^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).$$

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1).$$

+ Khi σ^2 chưa biết, $n < 30$, ta có

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{S} \sqrt{n} \sim t(n-1).$$

b2) Trường hợp biến ngẫu nhiên X không có phân phối chuẩn $\mathbb{N}(\mu; \sigma^2)$:

Theo lý thuyết về luật số lớn và định lý giới hạn trung tâm ta có tính chất xấp xỉ gần đúng của phân phối xác suất khi n đủ lớn (thường lấy $n \geq 30$) như sau.

+ Khi σ^2 đã biết, $n \geq 30$, ta có

$$\bar{X} \simeq \mathbb{N}(\mu, \frac{\sigma^2}{n}); \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \simeq \mathbb{N}(0; 1).$$

Trong đó, dấu " \simeq " là xấp xỉ phân phối.

+ Khi σ^2 chưa biết, $n \geq 30$, ta có

$$\bar{X} \simeq \mathbb{N}(\mu, \frac{S^2}{n}); \quad \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{S} \sqrt{n} \simeq \mathbb{N}(0; 1).$$

+ Khi p đã biết, $np \geq 5$, $n(1-p) \geq 5$, n đủ lớn, ta có

$$\frac{(F - p)}{\sqrt{p(1-p)/n}} = \frac{(F - p)}{\sqrt{p(1-p)}} \sqrt{n} \simeq \mathbb{N}(0; 1).$$

Khi p chưa biết, n đủ lớn, ta có

$$\frac{(F - p)}{\sqrt{F(1-F)/n}} = \frac{(F - p)}{\sqrt{F(1-F)}} \sqrt{n} \simeq \mathbb{N}(0; 1).$$

Các tính chất được nêu ở mục b1), b2) có ứng dụng quan trọng trong lý thuyết ước lượng và kiểm định giả thuyết thống kê.

2.3 Ước lượng điểm

Để có nhận định hay đánh giá về một đặc điểm định lượng hoặc đặc điểm định tính nào đó của đám đông \mathcal{P} tức là đi tìm hiểu về một biến ngẫu nhiên X nào đó của \mathcal{P} , thì vấn đề ta cần quan tâm ở đây là tìm hiểu xem biến ngẫu nhiên X này có phân phối gì với tham số bao nhiêu hoặc kỳ vọng và phương sai của X là bao nhiêu... Trong trường hợp tổng quát ta giả sử biến ngẫu nhiên X có phân phối $F(\theta)$ với tham số θ , ở đây F là một phân phối đã biết còn θ là tham số chưa biết mà ta cần tìm hiểu và ước lượng (giả sử $X \sim \mathbb{B}(n; p)$ nhưng chưa biết p , $X \sim P(\lambda)$ nhưng chưa biết λ , $X \sim \mathbb{N}(\mu; \sigma^2)$ nhưng chưa biết μ hay σ^2, \dots).

Để ước lượng về tham số θ , trong thống kê người ta chọn ra một mẫu ngẫu nhiên có kích thước n gồm n biến ngẫu nhiên độc lập cùng phân phối X_1, X_2, \dots, X_n . Đối với mẫu ngẫu nhiên này người ta có hai hướng giải quyết như sau

- + *Hướng thứ nhất*: Tìm ra một tham số $\hat{\theta}$ chỉ phụ thuộc vào X_1, X_2, \dots, X_n để thay thế cho θ , bài toán dạng này được gọi là bài toán ước lượng điểm và $\hat{\theta}$ được gọi là ước lượng điểm của θ . Như vậy ta thay thế giá trị tuyệt đối chưa biết của đám đông \mathcal{P} bởi giá trị tương đối thu được từ mẫu ngẫu nhiên.
- + *Hướng thứ hai*: Tìm ra một khoảng $(\theta_1; \theta_2)$ chứa θ với xác suất lớn, bài toán dạng này được gọi là bài toán ước lượng khoảng và $(\theta_1; \theta_2)$ được gọi là khoảng ước lượng (khoảng tin cậy) của tham số θ .

Sau đây ta tìm hiểu về bài toán ước lượng điểm của tham số θ .

2.3.1 Tiêu chuẩn ước lượng điểm

Giả sử biến ngẫu nhiên X gốc đám đông \mathcal{P} có phân phối nào đó với tham số θ chưa biết mà ta cần quan tâm, sau khi khảo sát mẫu ta tính được các thống kê, dựa vào các thống kê để đưa ra một số $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ chỉ phụ thuộc vào các quan sát mà không phụ thuộc vào θ để thay thế θ gọi là ước lượng điểm của θ . Có nhiều ước lượng $\hat{\theta}$ cho tham số θ , do đó phải lựa chọn dựa trên rất nhiều tiêu chuẩn như:

- *Không chệch*: Hiểu một cách đơn giản là ước lượng không chứa sai số hệ thống, tức là không thiên về phía đưa ra các giá trị bé hơn θ hoặc thiên về việc đưa ra giá trị lớn hơn θ . Như vậy: $\mathbb{E}(\hat{\theta}) = \theta$.
- *Vững*: Khi tăng kích thước mẫu n lên vô hạn thì ước lượng $\hat{\theta}$ sẽ dần đến θ theo xác suất. Tức là, với mọi $\varepsilon > 0$ tùy ý thì $\lim_{n \rightarrow \infty} \mathbb{P}[|\hat{\theta} - \theta| < \varepsilon] = 1$.

- **Hợp lý tối đa:** Hàm $L(x, \theta) = \prod_{i=1}^n p(X_i, \theta)$ đạt cực đại tại $\hat{\theta}$, trong đó $L(x, \theta)$ là hàm hợp lý của X , $p(x, \theta)$ là hàm mật độ xác suất hoặc là hàm tính xác suất của biến ngẫu nhiên X .
- **Hiệu quả:** Là ước lượng không chệch và có phương sai bé nhất.
- **Chắc hay bền:** Không thay đổi nhiều khi trong mẫu có các số liệu quá nhỏ hay quá lớn, hoặc số liệu thu được không thỏa mãn giả thuyết phân phối chuẩn.

2.3.2 Ước lượng điểm cho kỳ vọng, xác suất và phương sai

Ta có một số kết quả về ước lượng điểm sau đây

a) Ước lượng điểm cho kỳ vọng:

Mệnh đề 2.3.1. Giả sử X là biến ngẫu nhiên gốc đám đông \mathcal{P} , có kỳ vọng μ cần ước lượng, khi đó trung bình mẫu \bar{X} chính là ước lượng không chệch của μ .

Ví dụ 2.3.2. Nếu X là biến ngẫu nhiên có phân phối chuẩn $N(\mu; \sigma^2)$ thì \bar{X} là ước lượng hiệu quả của μ .

b) Ước lượng điểm cho phương sai:

Mệnh đề 2.3.3. Giả sử X là biến ngẫu nhiên gốc đám đông \mathcal{P} , có phương sai $\mathbb{D}X = \sigma^2$ cần ước lượng, khi đó phương sai mẫu hiệu chỉnh S^2 chính là ước lượng không chệch của σ^2 .

Ví dụ 2.3.4. Nếu X là biến ngẫu nhiên có phân phối chuẩn $N(\mu; \sigma^2)$ thì \bar{X} và S^2 lần lượt là ước lượng hợp lý tối đa của μ và σ^2 .

c) Ước lượng điểm cho xác suất:

Mệnh đề 2.3.5. Giả sử X là biến ngẫu nhiên gốc đám đông \mathcal{P} , ta cần quan tâm đến một tính chất A có xác suất $p = \mathbb{P}(A) = \mathbb{E}X$ cần ước lượng, khi đó tỉ lệ mẫu f chính là ước lượng không chệch của xác suất p .

Ví dụ 2.3.6. Để khảo sát chỉ tiêu X của một loại sản phẩm, người ta chọn ra một mẫu ngẫu nhiên để quan sát, có bảng kết quả thực nghiệm như sau:

$X(cm)$	11-15	15-19	19-23	23-27	27-31	31-35	35-39
Số sản phẩm	8	9	20	16	16	13	18

Những sản phẩm có chỉ tiêu X từ 19 cm trở xuống được xếp vào loại B. Hãy ước lượng giá trị trung bình, phương sai của chỉ tiêu X và tỉ lệ các sản phẩm loại B.

Giải. Trung bình mẫu: $\bar{x} = 26,36(\text{cm})$; Phương sai mẫu hiệu chỉnh: $s^2 = 55,9903(\text{cm}^2)$;
Tỉ lệ mẫu của sản phẩm loại B: $f = \frac{17}{100} = 17\%$.

Ta có các ước lượng điểm như sau:

Giá trị trung bình của X là

$$\mathbb{E}X \approx \bar{x} = 26,36(\text{cm}).$$

Phương sai của X là

$$\mathbb{D}X \approx s^2 = 55,9903(\text{cm}^2).$$

Tỉ lệ sản phẩm loại B là

$$p \approx f = 17\%.$$

2.4 Ước lượng khoảng

Giả sử biến ngẫu nhiên X gốc đám đông \mathcal{P} có phân phối nào đó với tham số θ chưa biết mà ta cần quan tâm, sau khi khảo sát mẫu ta tính được các thống kê, từ đó đưa ra được khoảng $(\theta_1; \theta_2)$ chứa tham số θ (với xác suất lớn là P). Cận dưới θ_1 và cận trên θ_2 tính theo một quy tắc cụ thể dựa trên các thống kê của mẫu và dựa trên độ tin cậy P .

Sau khi chọn mẫu và xử lý số liệu ta đưa ra khoảng tin cậy $(\theta_1; \theta_2)$. Nếu θ thuộc vào $(\theta_1; \theta_2)$ thì khoảng tin cậy đưa ra đúng, ngược lại là sai. Như vậy mỗi khoảng tin cậy chỉ có thể đúng hoặc sai, xác suất đúng là P , xác suất sai là $\alpha = 1 - P$.

2.4.1 Khái niệm về khoảng tin cậy

Khoảng $(\theta_1; \theta_2)$ được gọi là một khoảng ước lượng (*khoảng tin cậy*) (confidence interval) của tham số θ với độ tin cậy $1 - \alpha$ nếu θ thuộc vào khoảng trên với xác suất lớn $1 - \alpha$, tức là

$$\mathbb{P}(\theta_1 < \theta < \theta_2) = 1 - \alpha,$$

trong đó $(\theta_1; \theta_2)$: Khoảng tin cậy;

$P = 1 - \alpha$: Độ tin cậy (confidence level);

$l = \theta_2 - \theta_1$: Độ dài khoảng ước lượng;

$\varepsilon = \frac{\theta_2 - \theta_1}{2}$: Độ chính xác (sai số) của ước lượng.

Giả sử một khoảng ước lượng có dạng $(a - \varepsilon; a + \varepsilon)$ thì độ chính xác của khoảng ước lượng là ε .

Để xây dựng quy tắc tính khoảng tin cậy phải nghiên cứu sự thay đổi của trung bình cộng \bar{X} và phương sai S^2 , coi đó là các biến ngẫu nhiên phụ thuộc vào mẫu chọn ra.

Sau đây là các quy tắc tìm khoảng ước lượng:

2.4.2 Khoảng tin cậy cho giá trị trung bình

Giả sử X là biến ngẫu nhiên gốc đám đông \mathcal{P} có kỳ vọng $\mathbb{E}X = \mu$ (chưa biết) và phương sai $\mathbb{D}X = \sigma^2$ (đã biết hoặc chưa biết), ta tìm khoảng ước lượng cho tham số μ theo các trường hợp của phương sai σ^2 sau đây:

a) Khi đã biết phương sai σ^2 :

Quy tắc thực hành:

+ B_1 : Tìm \bar{x} , σ , α .

+ B_2 : Xác định phân vị $u_{\frac{\alpha}{2}}$ (còn gọi là giá trị tới hạn, critical value). $u_{\frac{\alpha}{2}}$ là giá trị tra ngược từ bảng phân phối chuẩn tắc $N(0; 1)$, thỏa $\Phi(u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$ (tra từ trong ra ngoài bảng)).

Chẳng hạn, $\alpha = 0,05$ thì $\Phi(u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} = 1 - 0,025 = 0,9750 \xrightarrow{\text{tra bảng } \Phi} \Phi(1,96)$.

Như vậy ta có $\Phi(u_{\frac{\alpha}{2}}) = \Phi(1,96)$ nên suy ra $u_{\frac{\alpha}{2}} = 1,96$ (do Φ là hàm tăng trên \mathbb{R}).

+ B_3 : Tính độ chính xác $\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. Khi đó khoảng tin cậy $(\bar{x} - \varepsilon; \bar{x} + \varepsilon)$.

Ví dụ 2.4.1. Cân 36 con vịt được trọng lượng trung bình là $\bar{x} = 2,6\text{kg}$. Với độ tin cậy 95%, hãy ước lượng khoảng về trọng lượng trung bình của vịt nếu phương sai đã biết là 0,09.

Giải. $n = 36$; $\bar{x} = 2,6$; $\sigma = 0,3$; $1 - \alpha = 0,95 \Rightarrow \alpha = 0,05 \Rightarrow u_{\frac{\alpha}{2}} = 1,96$.

Độ chính xác

$$\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 0,098.$$

Vậy khoảng ước lượng của trọng lượng trung bình của vịt là

$$(\bar{x} - \varepsilon; \bar{x} + \varepsilon) = (2,502\text{kg}; 2,698\text{kg}).$$

Chú ý: Nếu kích thước mẫu $n < 30$ ta cần bổ sung điều kiện X tuân theo luật phân phối chuẩn, khi đó $\frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0; 1)$.

b) Khi chưa biết phương sai σ^2 , cỡ mẫu lớn $n \geq 30$

Quy tắc thực hành: Tương tự trường hợp đã biết phương sai σ^2 nhưng ta thay σ bởi s . Cụ thể, gồm các bước sau:

- + B_1 : Tìm \bar{x} , α , độ lệch chuẩn mẫu s .
- + B_2 : Xác định phân vị $u_{\frac{\alpha}{2}}$ (tra bảng phân phối $N(0; 1)$).
- + B_3 : Tính độ chính xác $\varepsilon = u_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$. Khi đó, khoảng ước lượng là $(\bar{x} - \varepsilon; \bar{x} + \varepsilon)$.

Ví dụ 2.4.2. Để kiểm tra máy cắt những thanh thép, người ta chọn ngẫu nhiên ra một số thanh thép để đo chiều dài X (đơn vị: cm), thu được mẫu số liệu sau

$X (cm)$	148	149	151	152	153
Số thanh	3	17	10	5	1

- a) Tính các số đặc trưng của mẫu số liệu: \bar{x}, s .
- b) Hãy ước lượng khoảng về chiều dài trung bình của các thanh thép với độ tin cậy 96%.

Giải. a) Ta có $\bar{x} = \frac{1}{n} \sum_i x_i n_i = 150$.

$$\overline{x^2} = \frac{1}{n} \sum_i x_i^2 n_i = \frac{810068}{36} = 22501,8889.$$

$$s^2 = \frac{n}{n-1} [\overline{x^2} - (\bar{x})^2] = 1,9429.$$

Suy ra $s = \sqrt{s^2} = 1,3939$.

b) $\bar{x} = 150$; $1 - \alpha = 0,96 \Rightarrow \alpha = 0,04 \Rightarrow u_{\frac{\alpha}{2}} = 2,06$ (vì $\Phi(u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} = 0,98 = \Phi(2,06)$).

Độ chính xác: $\varepsilon = u_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 0,4786$.

Vậy khoảng ước lượng của chiều dài trung bình là

$$(\bar{x} - \varepsilon; \bar{x} + \varepsilon) = (149,5214cm; 150,4786cm).$$

Ví dụ 2.4.3. Để có kết quả đánh giá về trọng lượng trung bình của học sinh nam ở một trường trung học phổ thông nào đó, người ta chọn ngẫu nhiên một số học sinh của trường này để đo trọng lượng. Kết quả đo được về trọng lượng X (đơn vị: kilôgam) của học sinh như sau

Trọng lượng X	44	45	46	48	50	53
Số em (n_i)	3	7	9	5	6	2

- a) Tính các số đặc trưng \bar{x}, s^2, s của mẫu số liệu trên.
- b) Với độ tin cậy 90%, hãy ước lượng khoảng về trọng lượng trung bình của học sinh của trường.

c) Khi chưa biết phương sai σ^2 , cỡ mẫu bé $n < 30$, X có phân phối chuẩn

Quy tắc thực hành:

+ B_1 : Tìm \bar{x} , α , độ lệch chuẩn mẫu s .

+ B_2 : Xác định phân vị $t_{(n-1; \frac{\alpha}{2})}$ (giá trị ở dòng $n - 1$, cột $\frac{\alpha}{2}$, bảng phân phối Student).

+ B_3 : Xác định khoảng tin cậy $(\bar{x} - \varepsilon; \bar{x} + \varepsilon)$ với $\varepsilon = t_{(n-1; \frac{\alpha}{2})} \frac{s}{\sqrt{n}}$.

Ví dụ 2.4.4. Để ước lượng năng suất của một giống ngô, người ta theo dõi 25 mảnh ruộng trồng ngô. Sau khi thu hoạch được năng suất trung bình $\bar{x} = 10,6$ (tạ/ha), độ lệch chuẩn mẫu $s = 2,082$. Giả sử biết được năng suất ngô có phân phối chuẩn. Với độ tin cậy 95%, hãy ước lượng năng suất trung bình của giống ngô này.

Giải. Ta có $n = 25 < 30$ (mẫu bé); $\bar{x} = 10,6$; $s = 2,082$;

$1 - \alpha = 0,95 \Rightarrow \alpha = 0,05 \Rightarrow t_{(n-1; \frac{\alpha}{2})} = t_{(24; 0,025)} = 2,064$.

Độ chính xác: $\varepsilon = t_{(n-1; \frac{\alpha}{2})} \frac{s}{\sqrt{n}} = 0,8582$.

Vậy khoảng ước lượng về năng suất ngô trung bình là $(\bar{x} - \varepsilon; \bar{x} + \varepsilon) = (9,74; 11,46)$.

Sinh viên tham khảo thêm về cách tính khoảng ước lượng không đối xứng sau đây.

✂ Chú ý: Các khoảng ước lượng nêu trên là các khoảng ước lượng đối xứng (hai phía). Để tìm khoảng ước lượng không đối xứng (một phía) dạng $(-\infty; \theta_0)$ (tương tự: $(\theta_0; +\infty)$) thì giá trị θ_0 chính là giá trị lớn nhất (tương tự: nhỏ nhất) trong công thức về khoảng ước lượng đối xứng đã trình bày trên nhưng thay mức phân vị $\frac{\alpha}{2}$ bởi α . Lưu ý,

+ u_α là giá trị thỏa $\Phi(u_\alpha) = 1 - \alpha$, tra từ bảng phân phối chuẩn tắc $N(0; 1)$.

+ $t_{(n-1; \alpha)}$ là giá trị tại dòng $n - 1$, cột α trong bảng phân phối Student.

Bài toán đi tìm giá trị lớn nhất (tối đa) hay nhỏ nhất (tối thiểu) của tham số θ cần ước lượng chính là bài toán đi tìm khoảng ước lượng một phía. Do vậy

Ví dụ 2.4.5. Nhắc lại Ví dụ 2.4.4 về mẫu 25 mảnh ruộng ngô. Câu hỏi đặt ra: Với độ tin cậy 95%, hãy ước lượng năng suất trung bình tối đa của giống ngô này.

Giải. Ta có $n = 25 < 30$ (mẫu bé); $\bar{x} = 10,06$; $s = 2,082$; $\alpha = 0,05 \Rightarrow t_{(n-1; \alpha)} = t_{(24; 0,05)} = 1,711$.

Năng suất trung bình cao nhất là $\mu_{\max} = \bar{x} + t_{(n-1; \alpha)} \frac{s}{\sqrt{n}} = 10,7725$.

2.4.3 Khoảng tin cậy cho tỉ lệ

Chọn mẫu ngẫu nhiên kích thước n từ đám đông \mathcal{P} , mẫu cụ thể có m cá thể loại A (là cá thể có tính chất A nào đó mà ta đang xét, ví dụ chất lượng đạt yêu cầu, hoa vàng, mắt xanh, bệnh, sản phẩm hỏng, ...). Gọi p là tỉ lệ cá thể loại A thực sự của đám đông \mathcal{P} , p là một tham số chưa biết, ta tiến hành ước lượng tham số p như sau:

a) **Quy tắc thực hành:**

+ B_1 : Tìm cỡ mẫu n , tần số mẫu m , tỉ lệ mẫu $f = \frac{m}{n}$.

+ B_2 : Tìm phân vị $u_{\frac{\alpha}{2}}$ (tra bảng $\mathbb{N}(0; 1)$), $u_{\frac{\alpha}{2}}$ là giá trị thỏa $\Phi(u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

+ B_3 : Độ chính xác (sai số): $\varepsilon = u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}$.

Khi đó, khoảng tin cậy cho tỉ lệ là $(f - \varepsilon; f + \varepsilon)$.

Ví dụ 2.4.6. Để biết tỉ lệ người tiêu dùng không thích một loại sản phẩm mới đưa ra thị trường người ta hỏi ý kiến 344 người về sản phẩm này, kết quả có 83 người cho biết là không thích sản phẩm đó.

a) Với mức tin cậy 90% hãy ước lượng tỉ lệ người không thích sản phẩm này.

b) Với mức tin cậy 92% hãy ước lượng tỉ lệ người tiêu dùng thích sản phẩm này.

Giải. a) Ta có $n = 344$; Tần số mẫu $m = 83$; tỉ lệ mẫu $f = \frac{m}{n} = 0,241$.

$1 - \alpha = 0,9 \Rightarrow \alpha = 0,1 \Rightarrow u_{\frac{\alpha}{2}} = 1,65$.

Độ chính xác $\varepsilon = u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} = 0,038$.

Vậy khoảng tin cậy cho tỉ lệ $(f - \varepsilon; f + \varepsilon) = (0,203; 0,279)$.

2.4.4 Khoảng tin cậy cho phương sai

a) **Trường hợp đã biết kỳ vọng μ :**

Giả sử $\{X_1, X_2, \dots, X_n\}$ là mẫu ngẫu nhiên được chọn từ biến ngẫu nhiên X của đám đông C . X có phân phối chuẩn $\mathbb{N}(\mu; \sigma^2)$ có kỳ vọng μ đã biết, với mẫu số liệu thực nghiệm

X	x_1	x_2	\dots	x_k
Tần số	n_1	n_2	\dots	n_k

Quy tắc thực hành: Tìm khoảng tin cậy của phương sai σ^2 theo các bước:

+ B_1 : Xác định phân vị $\chi^2_{(n; \frac{\alpha}{2})}$ và $\chi^2_{(n; 1 - \frac{\alpha}{2})}$ (tra bảng χ^2).

$$+ B_2: \text{Khoảng tin cậy: } \left(\frac{\sum_{i=1}^k n_i(x_i - \mu)^2}{\chi^2(n; \frac{\alpha}{2})}; \frac{\sum_{i=1}^k n_i(x_i - \mu)^2}{\chi^2(n; 1 - \frac{\alpha}{2})} \right).$$

b) Trường hợp chưa biết kỳ vọng μ :

Quy tắc thực hành:

+ B_1 : Xác định phân vị $\chi^2_{(n-1; \frac{\alpha}{2})}$ và $\chi^2_{(n-1; 1-\frac{\alpha}{2})}$ (tra bảng χ^2).

+ B_2 : Xác định khoảng tin cậy: $\left(\frac{(n-1)s^2}{\chi^2_{(n-1; \frac{\alpha}{2})}}; \frac{(n-1)s^2}{\chi^2_{(n-1; 1-\frac{\alpha}{2})}} \right)$.

Ví dụ 2.4.7. Để tham khảo độ chính xác của một dụng cụ đo độ dài người ta đo trên cùng một mục tiêu 30 lần bằng dụng cụ ấy. Kết quả nhận được $s^2 = 0,05$. Hãy tìm khoảng ước lượng cho độ chính xác (phương sai) của dụng cụ đo với độ tin cậy 95%.

HD. Trường hợp chưa biết μ .

2.4.5 Tìm cỡ mẫu khi cho biết độ chính xác của ước lượng

Đây là bài toán ngược của bài toán ước lượng tham số. Cho các thông tin về khoảng ước lượng hoặc về độ chính xác của ước lượng, yêu cầu tìm cỡ mẫu, độ tin cậy,...

Ta biết rằng trong thống kê, nếu mẫu ngẫu nhiên được chọn có kích thước càng lớn thì độ chính xác của ước lượng càng tốt, tức sai số càng nhỏ (với một mức ý nghĩa α nào đó). Bài toán đặt ra như sau: Tìm cỡ mẫu n để độ chính xác của khoảng ước lượng không vượt quá ε_0 . Như vậy cần tìm n để $\varepsilon \leq \varepsilon_0$, ta xét các trường hợp sau

a) Dạng ước lượng giá trị trung bình (kỳ vọng μ)

Khi đã biết phương sai, ta có $\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \varepsilon_0 \Rightarrow n \geq \left(\frac{u_{\frac{\alpha}{2}} \sigma}{\varepsilon_0} \right)^2$. Ta có thể chọn

$$n = \left\lceil \left(\frac{u_{\frac{\alpha}{2}} \sigma}{\varepsilon_0} \right)^2 \right\rceil + 1,$$

trong đó $[x]$ là phần nguyên của số thực x (là số nguyên $\leq x$ và gần x nhất), ví dụ $[2,95] = 2$; $[0,15] = 0$; $[-2,95] = -3$; $[-0,15] = -1$.

b) Dạng ước lượng tỉ lệ p :

Ta có $\varepsilon = u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} \leq \varepsilon_0 \Rightarrow n \geq \frac{(u_{\frac{\alpha}{2}})^2 f(1-f)}{\varepsilon_0^2}$. Dùng bất đẳng thức $f(1-f) \leq \frac{1}{4}$.

$f) \leq \frac{1}{4}, \forall f \in \mathbb{R}$, để n đủ tốt ta lấy $n \geq \frac{(u_{\frac{\alpha}{2}})^2}{4\varepsilon_0^2}$. Ta có thể chọn

$$n = \left\lceil \frac{u_{\frac{\alpha}{2}}^2}{4\varepsilon_0^2} \right\rceil + 1.$$

Các dạng ước lượng khác ta có cách tìm cỡ mẫu n tương tự. Bên cạnh bài toán tìm cỡ mẫu như trên ta còn có bài toán tìm độ tin cậy và một số bài toán ngược khác.

c) Các ví dụ:

Ví dụ 2.4.8. Chọn ngẫu nhiên 45 em học sinh ở một trường tiểu học để cân, ta tính được $\bar{x} = 36kg$. Giả sử kế thừa từ các nghiên cứu trước đó, ta đã biết được $\sigma = 2,5kg$.

a) Với độ tin cậy 95%, hãy ước lượng khoảng về trọng lượng trung bình của học sinh của trường này.

b) Để khoảng ước lượng về trọng lượng trung bình của học sinh có sai số không vượt quá 0,4 thì cần phải chọn mẫu gồm tối thiểu bao nhiêu học sinh để cân? Vậy phải chọn thêm tối thiểu bao nhiêu học sinh nữa để cân so với lúc đầu?

Giải. a) Ta có $n = 45$; $\sigma = 2,5kg$, $\bar{x} = 36kg$; $\alpha = 0,05$; $u_{\frac{\alpha}{2}} = 1,96$.

Độ chính xác $\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 0,7304$.

Khoảng ước lượng: $(\bar{x} - \varepsilon; \bar{x} + \varepsilon) = (35,2696kg; 36,7304kg)$.

b) Theo đề bài ta có $\varepsilon \leq 0,4$ hay $u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq 0,4$. Từ đó, suy ra $n \geq \left\{ \frac{u_{\frac{\alpha}{2}} \cdot \sigma}{0,4} \right\}^2$, chọn

$$n = \left\lceil \left\{ \frac{u_{\frac{\alpha}{2}} \cdot \sigma}{0,4} \right\}^2 \right\rceil + 1 = [150,0625] + 1 = 150 + 1 = 151 \text{ học sinh.}$$

Vậy phải chọn thêm $151 - 45 = 106$ học sinh.

Ví dụ 2.4.9. Một khách sạn lớn muốn ước lượng tỉ lệ khách có nhu cầu nghỉ trọ nhiều hơn một ngày. Với độ tin cậy 96%, và độ chính xác của khoảng ước lượng không vượt quá 0,05. Hãy cho biết phải chọn cỡ mẫu tối thiểu là bao nhiêu người trong hai trường hợp:

(a) Chưa có thông tin nào cho phép ước lượng này.

(b) Dựa vào tài liệu khảo sát trước đây, tỉ lệ này là $f = 25\%$.

Giải. Độ tin cậy $1 - \alpha = 0,96$; $\varepsilon = 0,05$. Tìm cỡ mẫu n :

a) Chưa có thông tin nào cho phép ước lượng này:

$$\alpha = 0,04 \Rightarrow u_{\frac{\alpha}{2}} = 2,05.$$

$$\text{Theo đề bài ta có } n = \left\lceil \frac{u_{\frac{\alpha}{2}}^2}{4\varepsilon^2} \right\rceil + 1 = [420,25] + 1 = 421.$$

Vậy phải chọn tối thiểu là 421 người.

b) Dựa vào tài liệu khảo sát trước đây, tỉ lệ này là $f = 25\%$:

$$\alpha = 0,04 \Rightarrow u_{\frac{\alpha}{2}} = 2,05.$$

Theo đề bài ta có

$$u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} \leq \varepsilon \Rightarrow n \geq \frac{u_{\frac{\alpha}{2}}^2 \cdot f(1-f)}{\varepsilon^2}$$

$$\text{Chọn } n = \left\lceil \frac{u_{\frac{\alpha}{2}}^2 \cdot f(1-f)}{\varepsilon^2} \right\rceil + 1 = [315,18] + 1 = 316.$$

Vậy phải chọn tối thiểu là 316 người.

BÀI TẬP CHƯƠNG 2

- 2.1. Một công ty phần mềm muốn ước lượng thời gian trung bình để hoàn thành một dự án phần mềm nhỏ. Họ lấy mẫu ngẫu nhiên 15 dự án và đo thời gian hoàn thành (tính theo ngày) như sau:

15, 22, 18, 20, 16, 25, 24, 19, 21, 23, 18, 20, 22, 17, 19.

Hãy ước lượng khoảng về thời gian trung bình để hoàn thành một dự án với độ tin cậy 95%.

HD. Tính \bar{x}, s . Bài toán ước lượng khoảng về giá trị trung bình với cỡ mẫu bé hơn 30, chưa biết phương sai, tra bảng phân phối t (Student) để tìm giá trị tới hạn $t_{(n-1, \frac{\alpha}{2})}$.

- 2.2. Một công ty công nghệ muốn ước lượng số lần trung bình một ứng dụng di động bị lỗi trong vòng một tháng. Họ thu thập dữ liệu từ 40 ứng dụng và nhận thấy số lần bị lỗi trung bình là 4 lần, với độ lệch chuẩn mẫu là 1,5 lần. Tính khoảng tin cậy 95% cho số lần bị lỗi trung bình mỗi tháng.

HD. Bài toán ước lượng khoảng về giá trị trung bình, chưa biết phương sai, cỡ mẫu lớn, tra bảng phân phối chuẩn để tìm giá trị tới hạn $u_{\frac{\alpha}{2}}$.

- 2.3. Một nhóm nghiên cứu muốn ước lượng thời gian trung bình mà người dùng dành để sử dụng một ứng dụng học trực tuyến. Họ khảo sát 50 người dùng và tìm thấy thời gian trung bình sử dụng ứng dụng là 45 phút, với độ lệch chuẩn mẫu là $s = 10$ phút. Tính khoảng tin cậy 95% cho thời gian trung bình sử dụng ứng dụng.

- 2.4. Lấy mẫu ngẫu nhiên 100 sinh viên để khảo sát về sử dụng công cụ quản lý phiên bản Git, có 85 sinh viên ngành khoa học máy tính sử dụng Git. Hãy tìm ước lượng khoảng cho tỉ lệ sinh viên sử dụng Git, với độ tin cậy 90%.

HD. Bài toán ước lượng khoảng về tỉ lệ, $f = m/n, m = ?, n = ?$

- 2.5. Để đánh giá trữ lượng cá trong hồ người ta đánh bắt 3000 con cá đánh dấu rồi thả xuống hồ. Sau đó bắt lại 1200 con thì thấy có 240 con có dấu. Với độ tin cậy 95% hãy ước lượng trữ lượng cá trong hồ.

HD. Gọi p là tỉ lệ cá đã bị đánh dấu trong hồ, N là trữ lượng cá trong hồ thì $p = 3000/N$. Cỡ mẫu $n = 1200, m = 240$. Từ đó ước lượng khoảng về tỉ lệ p và suy ra được N .

- 2.6. Giả sử khảo sát về kết quả kiểm tra năng lực đầu vào tiếng Anh của 50 sinh viên Trường Đại học Đồng Tháp thu được bảng số liệu sau:

Điểm x_i	[0 ; 2]	(2 ; 4]	(4 ; 6]	(6 ; 8]	(8 ; 10]
Số sinh viên n_i	8	10	12	14	6

- a) Hãy ước lượng khoảng tin cậy về điểm điểm trung bình kiểm tra năng lực đầu vào môn tiếng Anh của mỗi sinh viên với độ tin cậy 99% .
- b) Những sinh viên có điểm trên 6 là những sinh viên có năng lực tiếng Anh tốt. Hãy ước lượng tỉ lệ sinh viên học Anh văn tốt với độ tin cậy 95% .
- c) Giả sử sinh viên học Anh văn kém là có điểm từ 4 trở xuống. Hãy ước lượng tỉ lệ sinh viên học Anh văn kém với độ tin cậy 90% .

HD. Khi nhập số liệu vào máy tính, ta nhập vào phần tử đại diện lớp là trung điểm của mỗi lớp, ví dụ lớp $[0;2]$ ta nhập 1.

- 2.7. Để đánh giá hiệu quả của một loại thuốc, người ta đem sử dụng cho 1000 bệnh nhân thì có 820 người khỏi bệnh. Với độ tin cậy 95%, hãy ước lượng khoảng cho tỉ lệ chữa khỏi bệnh cho loại thuốc trên.
- 2.8. Cho X (%) và Y (kg/cm²) là hai chỉ tiêu chất lượng của một sản phẩm. Tiến hành kiểm tra một số sản phẩm, người ta thu được kết quả qua bảng số liệu sau:

$Y \setminus X$	30 - 35	35 - 40	40 - 45	45 - 50
130 - 135	3			
135 - 140	3	14	18	
140 - 145		11	20	
145 - 150		3	17	5
150 - 155			2	4

- a) Hãy ước lượng giá trị trung bình của chỉ tiêu Y của những sản phẩm có chỉ tiêu X trong khoảng $(35;40)$ với độ tin cậy 95%.
- b) Nếu muốn ước lượng giá trị trung bình của chỉ tiêu X với độ chính xác 0,5 thì độ tin cậy là bao nhiêu?

CHƯƠNG 3

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ, TƯƠNG QUAN, HỒI QUY TUYẾN TÍNH

3.1 Bài toán kiểm định giả thuyết thống kê

Khi khảo sát về một đám đông \mathcal{P} (hoặc nhiều đám đông) và xem xét một biến ngẫu nhiên X của đám đông \mathcal{P} , người ta có thể đưa ra các nhận định về phân phối của biến ngẫu nhiên này hoặc nếu đã biết phân phối thì nhận định về tham số θ của X . Các nhận định này có thể đúng hoặc sai (với mức ý nghĩa α nào đó cho trước). Để có kết luận thống kê là “chấp nhận” hay “bác bỏ” cho các nhận định này, ta tiến hành chọn mẫu ngẫu nhiên từ \mathcal{P} , tính tham số mẫu, chọn mức ý nghĩa α , sau đó xử lý số liệu và đưa ra kết luận thống kê.

3.1.1 Cặp giả thuyết thống kê

Khi nhận định về tham số θ của biến ngẫu nhiên gốc đám đông hay về phân phối của X , người ta luôn có thể đưa ra hai nhận định đối lập nhau, một nhận định được gọi là *giả thuyết*¹ (null hypothesis) (Kí hiệu là H_0) thì nhận định còn lại được gọi là *đối thuyết* (alternative hypothesis) (Kí hiệu là H_1). Cụ thể:

★ Nhận định về tham số θ , giả thuyết và đối thuyết có dạng:

+ Giả thuyết: $H_0 : \theta = \theta_0$;

+ Đối thuyết hai phía: $H_1 : \theta \neq \theta_0$.

Hoặc $H_1 : \theta > \theta_0$ gọi là đối thuyết *lớn hơn* (hoặc một phía phải, 1-đuôi (One-tailed) phải).

Hoặc $H_1 : \theta < \theta_0$ gọi là đối thuyết *bé hơn* (hoặc một phía trái, 1-đuôi trái).

★ Nhận định về hai tham số θ_1 và θ_2 , giả thuyết và đối thuyết có dạng:

+ Giả thuyết: $H_0 : \theta_1 = \theta_2$;

¹Một số tài liệu dùng từ "giả thiết".

+ Đối thuyết hai phía: $H_1 : \theta_1 \neq \theta_2$

Hoặc đối thuyết một phía $H_1 : \theta_1 > \theta_2$; $H_1 : \theta_1 < \theta_2$.

- ★ Ngoài ra còn có nhiều dạng nhận định khác mà giả thuyết và đối thuyết không liên quan đến tham số chưa biết (ví dụ kiểm định về sự phù hợp hay không phù hợp, độc lập hay phụ thuộc,...). Kiểm định thống kê dạng này gọi là kiểm định phi tham số.

Khi kiểm định giả thuyết thống kê, nếu không cẩn thận sẽ mắc phải sai lầm (rủi ro) trong hai trường hợp sau:

- *Sai lầm loại 1*²: Bác bỏ giả thuyết trong khi giả thuyết đúng, xác suất mắc sai lầm loại 1 được gọi là rủi ro loại 1.

- *Sai lầm loại 2*³: Chấp nhận giả thuyết trong khi giả thuyết sai, xác suất mắc sai lầm loại 2 được gọi là rủi ro loại 2.

Người ta thường tập trung chú ý vào sai lầm loại 1, trong thống kê phải không chế sao cho rủi ro loại này không vượt quá mức ý nghĩa α .

Như vậy nhiệm vụ của lý thuyết kiểm định là xử lý số liệu từ mẫu ngẫu nhiên được chọn để đưa ra kết luận thống kê là bác bỏ giả thuyết H_0 , chấp nhận đối thuyết H_1 hay ngược lại với mức ý nghĩa α cho trước. Giải quyết bài toán thống kê này theo hướng như dưới đây.

3.1.2 Tiêu chuẩn kiểm định giả thuyết thống kê

- ★ Bài toán kiểm định giả thuyết thống kê (cho 1 tham số θ):

Cho các thông tin của mẫu ngẫu nhiên được chọn từ đám đông \mathcal{P} , với mức ý nghĩa α , hãy kiểm định giả thuyết và đối thuyết sau

Giả thuyết $H_0 : \theta = \theta_0$;

Đối thuyết $H_1 : \theta \neq \theta_0$. (hoặc đối thuyết một phía $H_1 : \theta > \theta_0$, $H_1 : \theta < \theta_0$.)

Dạng bài toán kiểm định hai (hay nhiều) tham số và phi tham số là tương tự.

- ★ Hướng giải quyết cho bài toán kiểm định:

Để đưa ra kết luận thống kê với mức ý nghĩa α cho bài toán kiểm định, người ta tiến hành như sau:

- + Tìm phân vị (dựa vào mức ý nghĩa α và tính chất các phân phối của b.n.n trong bài toán).

²The error of the first kind (i.e., the conviction of an innocent person).

³The error of the second kind (acquitting a person who committed the crime).

- + Tìm ra miền tiêu chuẩn S còn gọi là *miền bác bỏ* (miền này có một tính chất nào đó liên quan đến phân vị vừa tìm).
- + Nếu các tham số mẫu (số liệu của mẫu) thỏa tính chất của miền S thì bác bỏ giả thuyết H_0 và chấp nhận đối thuyết H_1 (với mức ý nghĩa α), ngược lại thì chấp nhận giả thuyết và bác bỏ đối thuyết (với mức ý nghĩa α).

3.2 Kiểm định giả thuyết về giá trị trung bình

Giả sử biến ngẫu nhiên X có kỳ vọng $\mathbb{E}X = \mu$ (chưa biết) và phương sai $\mathbb{D}X = \sigma^2$ (đã biết hoặc chưa biết), xét bài toán kiểm định giả thuyết và đối thuyết sau:

- + Giả thuyết $H_0 : \mu = \mu_0$;
- + Đối thuyết 2 phía $H_1 : \mu \neq \mu_0$ (hoặc đối thuyết 1 phía $H_1 : \mu > \mu_0$; $H_1 : \mu < \mu_0$).

Xét bài toán trong hai trường hợp của phương sai:

3.2.1 Khi đã biết phương sai σ^2

a) *Bài toán kiểm định hai phía* $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$

Quy tắc thực hành:

- + B_1 : Lấy mẫu kích thước n , tính \bar{x} .
- + B_2 : Xác định phân vị $u_{\frac{\alpha}{2}}$ (tra bảng phân phối chuẩn tắc $\mathbb{N}(0; 1)$).
- + B_3 : Tính giá trị thực nghiệm (còn gọi là Test thống kê): $T_{tn} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$.
- + B_4 : So sánh: Nếu $|T_{tn}| \leq u_{\frac{\alpha}{2}}$ thì chấp nhận H_0 , bác bỏ H_1 (nếu ngược lại thì bác bỏ H_0 , chấp nhận H_1).

Ví dụ 3.2.1. Lợn nếu được nuôi bằng thức ăn A sẽ cho tăng trọng trung bình $32kg$ sau 4 tháng. Nuôi 100 con lợn bằng thức ăn này, sau 4 tháng người ta thấy tăng trọng trung bình là $30kg$, giả thuyết trọng lượng lợn là biến ngẫu nhiên có phân phối chuẩn với phương sai đã biết là $\sigma^2 = 25kg^2$. Với mức ý nghĩa 5% , hãy cho biết nuôi bằng thức ăn A thì sau 4 tháng lợn tăng trọng trung bình $32kg$ có đúng không, tức là kiểm định giả thuyết $H_0 : \mu = 32kg$ và đối thuyết $H_1 : \mu \neq 32kg$.

Giải. Bài toán kiểm định $H_0 : \mu = 32kg$; $H_1 : \mu \neq 32kg$

Ta có: $n = 100$, $\bar{x} = 30kg$, $\sigma = 5$; $\alpha = 0,05 \Rightarrow u_{\frac{\alpha}{2}} = 1,96$.

Giá trị thực nghiệm $T_{tn} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = -4$.

Vì $|T_{tn}| > u_{\frac{\alpha}{2}}$ nên bác bỏ H_0 , chấp nhận H_1 .

Vậy tăng trọng trung bình của lợn sau 4 tháng không phải là 32kg.

Giải thích: Về cơ sở của quy tắc kiểm định giá trị trung bình, sinh viên tham khảo ở phần phụ lục.

b) **Bài toán kiểm định một phía lớn hơn** $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$, **một phía bé hơn** $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$

Quy tắc thực hành:

+ Xác định phân vị u_{α} .

Trường hợp $H_1 : \mu > \mu_0$: Nếu $T_{tn} \leq u_{\alpha}$ thì chấp nhận H_0 , bác bỏ H_1 .

Trường hợp $H_1 : \mu < \mu_0$: Nếu $T_{tn} \geq -u_{\alpha}$ thì chấp nhận H_0 , bác bỏ H_1 .

Ví dụ 3.2.2. Sửa lại nội dung của ví dụ trên. Một người nhận định rằng lợn nếu được nuôi bằng thức ăn A sẽ cho tăng trọng trung bình là 30kg sau 4 tháng. Để kiểm tra điều này, người ta nuôi 100 con lợn bằng thức ăn A, sau 4 tháng người ta tính được tăng trọng trung bình của lợn là 28kg, giả thuyết trọng lượng lợn là biến ngẫu nhiên có phương sai $\sigma^2 = 25$. Với mức ý nghĩa 5%, hãy cho biết thực tế tăng trọng trung bình của lợn có thấp hơn so với nhận định của người này không?

Giải. Bài toán kiểm định $H_0 : \mu = 30kg$; $H_1 : \mu < 30kg$

Ta có: $n = 100$, $\bar{x} = 28kg$, $\sigma = 5$; $u_{\alpha} = 1,65$.

Giá trị thực nghiệm $T_{tn} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = -4$.

Vì $T_{tn} = -4 < -u_{\alpha} = -1,65$ nên bác bỏ H_0 , chấp nhận H_1 .

Vậy tăng trọng trung bình của lợn sau 4 tháng là thấp hơn 30kg.

Chú ý, thông thường đối với các câu hỏi: "có thay đổi không", "có như nhau không", "có khác nhau không", ... ta dùng đối thuyết khác " \neq ".

3.2.2 Khi chưa biết phương sai σ^2 , cỡ mẫu lớn $n \geq 30$

Tương tự như trường hợp đã biết phương sai σ^2 nhưng ta thay σ bởi s , tức là chỉ thay đổi giá trị thực nghiệm $T_{tn} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$.

Ví dụ 3.2.3. Chọn ngẫu nhiên 32 sinh viên nam của Khoa Giáo dục thể chất để đo thành tích nhảy xa X (đơn vị: mét), kết quả thu được

X	4,5	4,65	4,8	4,95	5,1	5,2	5,5	6,0
Số sinh viên	1	5	10	7	4	2	2	1

a) Tính các số đặc trưng mẫu \bar{x}, s^2, s .

b) Có nhận định cho rằng thành tích nhảy xa trung bình của sinh viên nam của khoa cao hơn 4,9 m. Với mức ý nghĩa 5%, hãy cho biết nhận định trên đúng hay sai?

ĐS: a) $\bar{x} = 4,94375; s = 0,3005$; b) Chấp nhận H_0 . Nhận định sai.

3.2.3 Khi chưa biết phương sai σ^2 , cỡ mẫu bé $n < 30$, X có phân phối chuẩn

Ta tiến hành theo các bước sau

a) **Bài toán kiểm định hai phía** $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$

Quy tắc thực hành:

+ B_1 : Lấy mẫu kích thước n , tính \bar{x} và s .

+ B_2 : Xác định phân vị $t_{(n-1; \frac{\alpha}{2})}$ (là giá trị ở dòng $n - 1$, cột $\frac{\alpha}{2}$, bảng phân phối Student).

+ B_3 : Tính giá trị thực nghiệm: $T_{tn} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$.

+ B_4 : So sánh: Nếu $|T_{tn}| \leq t_{(n-1; \frac{\alpha}{2})}$ thì chấp nhận H_0 , bác bỏ H_1 .

Ví dụ 3.2.4. Trong điều kiện chăn nuôi bình thường lượng sữa trung bình của một con bò sữa là 19 kg/ngày. Trong một đợt hạn, người ta theo dõi 25 con bò và đo được lượng sữa trung bình là 17,5kg/ ngày, độ lệch chuẩn mẫu 2,5kg. Giả thuyết lượng sữa có phân phối chuẩn. Với mức ý nghĩa 5%, hãy cho biết lượng sữa trung bình của bò có thay đổi trong đợt hạn hay không?

Giải. Bài toán kiểm định $H_0 : \mu = 19$; $H_1 : \mu \neq 19$.

$n = 25$, $\bar{x} = 17,5$, $s = 2,5$; $\alpha = 0,05 \Rightarrow t_{(n-1; \frac{\alpha}{2})} = 2,064$

Giá trị thực nghiệm

$$T_{tn} = \frac{(\bar{x} - \mu_0)}{s} \sqrt{n} = -3$$

Vì $|T_{tn}| = 3 > t_{(n-1; \frac{\alpha}{2})}$ nên bác bỏ H_0 , chấp nhận H_1 , tức là lượng sữa trung bình của bò trong đợt hạn không còn là 19 kg/ ngày nữa.

Bài toán kiểm định một phía lớn hơn: $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$, **một phía bé hơn** $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$

Quy tắc thực hành:

Trường hợp $H_1 : \mu > \mu_0$: Nếu $T_{tn} \leq t_{(n-1; \alpha)}$ thì chấp nhận H_0 , bác bỏ H_1 .

Trường hợp $H_1 : \mu < \mu_0$: Nếu $T_{tn} \geq -t_{(n-1; \alpha)}$ thì chấp nhận H_0 , bác bỏ H_1 .

Ví dụ 3.2.5. Sửa lại nội dung ví dụ trên. Trong điều kiện chăn nuôi bình thường lượng sữa trung bình của một con bò sữa là 19 kg/ngày. Trong một đợt hạn kéo dài, người ta theo dõi 25 con bò và đo được lượng sữa trung bình là 17,5kg/ ngày, độ lệch chuẩn mẫu $s = 2,5\text{kg}$. Giả thuyết lượng sữa có phân phối chuẩn. Với mức ý nghĩa 5%, hãy cho biết lượng sữa trung bình của bò có giảm sút trong đợt hạn hay không?

Giải. Bài toán kiểm định $H_0 : \mu = 19; \quad H_1 : \mu < 19$.

$n = 25, \quad \bar{x} = 17,5, \quad s = 2,5; \quad \alpha = 0,05; \quad t_{(n-1);\alpha} = 1,711$.

Giá trị thực nghiệm

$$T_{tn} = \frac{(\bar{x} - \mu_0)}{s} \sqrt{n} = -3$$

Vì $T_{tn} = -3 < -t_{(n-1);\alpha} = -1,711$ nên bác bỏ H_0 , chấp nhận H_1 .

Vậy lượng sữa trung bình của bò trong đợt hạn đã có giảm so với điều kiện chăn nuôi bình thường.

✠ **Bảng tổng hợp quy tắc kiểm định về giá trị trung bình μ :**

Điều kiện chấp nhận giả thuyết $H_0 : \mu = \mu_0$				
Trường hợp của σ	T_{tn}	$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$
1) Đã biết σ	$\frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$	$ T_{tn} \leq u_{\frac{\alpha}{2}}$	$T_{tn} \leq u_{\alpha}$	$T_{tn} \geq -u_{\alpha}$
2) Chưa biết $\sigma, n \geq 30$	$\frac{\bar{x} - \mu_0}{s} \sqrt{n}$	$ T_{tn} \leq u_{\frac{\alpha}{2}}$	$T_{tn} \leq u_{\alpha}$	$T_{tn} \geq -u_{\alpha}$
3) Chưa biết $\sigma, n < 30$ (X có phân phối chuẩn)	$\frac{\bar{x} - \mu_0}{s} \sqrt{n}$	$ T_{tn} \leq t_{(n-1);\frac{\alpha}{2}}$	$T_{tn} \leq t_{(n-1);\alpha}$	$T_{tn} \geq -t_{(n-1);\alpha}$

✠ **Chú ý:** Để ngắn gọn trong cách diễn đạt, phương pháp kiểm định giả thuyết như trong Mục 3.2.1, 3.2.2 được gọi là **kiểm định Z** cho kỳ vọng. Phương pháp kiểm định trong Mục 3.2.3 được gọi là **kiểm định T** (xem [7], trang 163).

3.3 Kiểm định giả thuyết về tỉ lệ

Gọi p là tỉ lệ cá thể có tính chất A (cá thể loại A) trong đám đông \mathcal{P} , tỉ lệ p chưa biết, bài toán kiểm định tỉ lệ p như sau:

3.3.1 Bài toán kiểm định hai phía

Giả thuyết và đối thuyết có dạng
$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Ta tiến hành lấy mẫu ngẫu nhiên kích thước n ; đếm số cá thể loại A trong mẫu (tần số mẫu), giả sử là m . Với mức ý nghĩa α , bài toán kiểm định gồm các bước sau:

Quy tắc thực hành :

+ B_1 : Tìm n , tần số mẫu m , tỉ lệ mẫu $f = \frac{m}{n}$.

+ B_2 : Xác định phân vị $u_{\frac{\alpha}{2}}$ (tra bảng phân phối chuẩn tắc $\mathbb{N}(0; 1)$).

+ B_3 : Tính giá trị thực nghiệm: $T_{tn} = \frac{f - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n}$.

+ B_4 : So sánh: Nếu $|T_{tn}| \leq u_{\frac{\alpha}{2}}$ thì chấp nhận H_0 bác bỏ H_1 .

Ví dụ 3.3.1. Một lô sản phẩm có tỉ lệ phế phẩm chưa biết. Kiểm tra ngẫu nhiên 50 sản phẩm từ lô thấy có 5 phế phẩm. Một người nhận định rằng tỉ lệ phế phẩm của lô là 8%. Với mức ý nghĩa 3%, hãy cho biết nhận định trên là đúng hay sai?

Giải. Bài toán kiểm định $H_0 : p = 0,08$; $H_1 : p \neq 0,08$.

$n = 50$, $m = 5$, $f = \frac{m}{n} = 0,1$; $\alpha = 0,03 \Rightarrow u_{\frac{\alpha}{2}} = 2,17$.

Giá trị thực nghiệm

$$T_{tn} = \frac{f - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} = 0,5213.$$

Vì $|T_{tn}| = 0,5213 < u_{\frac{\alpha}{2}} = 2,17$ nên chấp nhận H_0 , bác bỏ H_1 .

Vậy nhận định trên là đúng.

3.3.2 Bài toán kiểm định một phía

Kiểm định một phía lớn hơn $\begin{cases} H_0 : p = p_0 \\ H_1 : p > p_0 \end{cases}$ và một phía bé hơn $\begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases}$

Quy tắc thực hành:

+ Xác định phân vị u_α : Là giá trị thỏa $\Phi(u_\alpha) = 1 - \alpha$.

+ Tính giá trị thực nghiệm: $T_{tn} = \frac{f - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n}$.

Trường hợp $H_1 : p > p_0$: Nếu $T_{tn} \leq u_\alpha$ thì chấp nhận H_0 , bác bỏ H_1 .

Trường hợp $H_1 : p < p_0$: Nếu $T_{tn} \geq -u_\alpha$ thì chấp nhận H_0 , bác bỏ H_1 .

Ví dụ 3.3.2. Sửa lại nội dung ví dụ trên. Một lô sản phẩm có tỉ lệ phế phẩm chưa biết. Kiểm tra ngẫu nhiên 50 sản phẩm từ lô thấy có 5 phế phẩm. Một người nhận định rằng tỉ lệ phế phẩm của lô cao hơn 8%. Với mức ý nghĩa 3%, hãy cho biết nhận định trên là đúng hay sai?

Giải. Bài toán kiểm định $H_0 : p = 0,08$; $H_1 : p > 0,08$.

$n = 50, m = 5, f = \frac{m}{n} = 0,1; \alpha = 0,03 \Rightarrow u_\alpha = 1,88$ (vì $\Phi(u_\alpha) = 1 - \alpha = 0,97 = \Phi(1,88)$).

Giá trị thực nghiệm

$$T_{tn} = \frac{f - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} = 0,5212.$$

Vì $T_{tn} = 0,5212 < u_\alpha = 1,88$ nên chấp nhận H_0 , bác bỏ H_1 .

Vậy nhận định trên là sai.

3.4 Kiểm định (so sánh) hai tham số

3.4.1 Kiểm định (so sánh) hai giá trị trung bình

Trên hai đám đông $\mathcal{P}_1, \mathcal{P}_2$ ta tiến hành theo dõi một biến ngẫu nhiên định lượng X . Gọi X_1 là kết quả quan sát về biến ngẫu nhiên X trên \mathcal{P}_1 , X_2 là kết quả quan sát về biến ngẫu nhiên X trên \mathcal{P}_2 . Giả sử $\mathbb{E}X_1 = \mu_1, \mathbb{E}X_2 = \mu_2$ và $\mathbb{D}X_1 = \sigma_1, \mathbb{D}X_2 = \sigma_2$ (μ_1, μ_2 chưa biết).

Để so sánh μ_1 và μ_2 ta phải chọn ra mẫu để quan sát. Có hai cách chọn mẫu cho dạng này là chọn mẫu theo cặp và chọn mẫu độc lập. Trong phần này ta chỉ quan tâm, đến chọn mẫu độc lập. Từ $\mathcal{P}_1, \mathcal{P}_2$, ta chọn ngẫu nhiên ra hai mẫu độc lập, kích thước mẫu n_1 và n_2 có thể bằng nhau hoặc khác nhau. Bài toán kiểm định như sau

$$\begin{cases} H_0 : \mu_1 = \mu_2; \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \quad (\text{hoặc đôi thuyết một phía } H_1 : \mu_1 > \mu_2; \quad H_1 : \mu_1 < \mu_2)$$

Tương tự kiểm định kỳ vọng μ , xét các trường hợp sau của σ_1, σ_2 :

a) **Đã biết phương sai σ_1^2 và σ_2^2 :**

Quy tắc thực hành: (Đối với bài toán hai phía)

+ B_1 : Lấy hai mẫu kích thước n_1, n_2 , tính \bar{x}_1, \bar{x}_2 .

+ B_2 : Xác định phân vị $u_{\frac{\alpha}{2}}$ (tra bảng phân phối chuẩn tắc $N(0; 1)$).

+ B_3 : Tính giá trị thực nghiệm: $T_{tn} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

+ B_4 : So sánh: Nếu $|T_{tn}| \leq u_{\frac{\alpha}{2}}$ thì chấp nhận H_0 và bác bỏ H_1 .

Ví dụ 3.4.1. Giả sử chiều dài của một loại cá trong hai ao là biến ngẫu nhiên có phân phối chuẩn với độ lệch chuẩn lần lượt là $\sigma_1 = 2cm$ và $\sigma_2 = 2,2cm$. Lấy mẫu 100 con của ao 1 đo được $\bar{x}_1 = 8cm$; lấy mẫu 120 con ở ao 2 được $\bar{x}_2 = 8,5cm$. Với mức ý nghĩa 1%, hãy cho biết chiều dài trung bình của cá trong hai ao có khác nhau không (sự khác nhau này có ý nghĩa hay không hay chỉ là sự sai khác ngẫu nhiên)?

Giải. Bài toán kiểm định $H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2$

$n_1 = 100, \quad \bar{x}_1 = 8, \quad n_2 = 120, \quad \bar{x}_2 = 8,5, \quad \sigma_1 = 2, \quad \sigma_2 = 2,2.$

$\alpha = 0,01 \Rightarrow u_{\frac{\alpha}{2}} = (\text{vì } \phi(u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} = 1 - 0,01/2 = 0,995 = \Phi(2,58))$

Giá trị thực nghiệm

$$T_{tn} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = -1,7645.$$

Vì $|T_{tn}| = 1,7645 < u_{\frac{\alpha}{2}} = 2,58$ nên chấp nhận H_0 , bác bỏ H_1 .

Vậy chiều dài trung bình của cá trong hai ao là như nhau. Tức là sự khác nhau về trọng lượng cá ở hai ao là không có ý nghĩa, chỉ là sự khác nhau do chọn mẫu ngẫu nhiên.

Bài toán kiểm định một phía: lớn hơn $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$ và bé hơn $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$

Trường hợp $H_1 : \mu_1 > \mu_2$: Nếu $T_{tn} \leq u_{\alpha}$ thì chấp nhận H_0 , bác bỏ H_1 .

Trường hợp $H_1 : \mu_1 < \mu_2$: Nếu $T_{tn} \geq -u_{\alpha}$ thì chấp nhận H_0 , bác bỏ H_1 .

b) Chưa biết phương sai σ_1^2 và σ_2^2 , cỡ mẫu lớn ($n_1 \geq 30, n_2 \geq 30$):

Quy tắc thực hành tương tự trường hợp đã biết σ_1, σ_2 nhưng khi cỡ mẫu lớn ta thay σ_1 bởi s_1 , σ_2 bởi s_2 . Tức là ta chỉ cần thay đổi giá trị thực nghiệm

$$T_{tn} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Ví dụ 3.4.2. Để thí nghiệm về hai phương pháp chăn nuôi gà, người ta chọn mẫu và thử nghiệm nuôi theo hai phương pháp khác nhau. Biết trọng lượng gà là biến ngẫu nhiên có phân phối chuẩn, sau một tháng kết quả tăng trọng của thí nghiệm cho bởi bảng sau

Phương pháp nuôi	Số gà	Tăng trọng trung bình (kg)	Độ lệch chuẩn mẫu
I	$n_1 = 160$	$\bar{x}_1 = 1,25$	$s_1 = 0,3$
II	$n_2 = 110$	$\bar{x}_2 = 1,2$	$s_2 = 0,2$

Với mức ý nghĩa 5%, hãy cho biết:

a) Nuôi theo phương pháp I có tăng trọng trung bình cao hơn 1,15kg/tháng không?

b) Nuôi theo phương pháp I có hiệu quả hơn (tăng trọng trung bình cao hơn) nuôi theo phương pháp II không?

Kết quả: a) $H_1 : \mu > 1,15; u_\alpha = 1,65; T_{tn} = 4,2164$; Bác bỏ H_0 . b) $H_1 : \mu_1 > \mu_2; u_\alpha = 1,65; T_{tn} = 1,643$; Chấp nhận H_0 .

c) Chưa biết phương sai σ_1^2 và σ_2^2 , mẫu bé (ít nhất một trong hai số $n_1, n_2 < 30$):

Đây là bài toán còn nhiều vướng mắc về mặt lý thuyết, do đó ta chỉ xét trong trường hợp đơn giản hơn là thêm giả thuyết phụ: $\sigma_1 = \sigma_2$, hai biến ngẫu nhiên X_1, X_2 có phân phối chuẩn.

Quy tắc thực hành: (Đối với bài toán hai phía)

+ B_1 : Lấy hai mẫu kích thước n_1, n_2 , tính $\bar{x}_1, \bar{x}_2, s_1, s_2$.

+ B_2 : Xác định phân vị $t_{(n_1+n_2-1; \frac{\alpha}{2})}$ (tra bảng phân phối Student).

+ B_3 : Tính giá trị thực nghiệm:

$$T_{tn} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

+ B_4 : So sánh: Nếu $|T_{tn}| \leq t_{(n_1+n_2-1; \frac{\alpha}{2})}$ thì chấp nhận H_0 và bác bỏ H_1 .

Bài toán kiểm định một phía: Lớn hơn ($\mu_1 > \mu_2$) và bé hơn ($\mu_1 < \mu_2$).

Trường hợp $H_1 : \mu_1 > \mu_2$: Nếu $T_{tn} \leq t_{(n_1+n_2-1;\alpha)}$ thì chấp nhận H_0 , bác bỏ H_1 .

Trường hợp $H_1 : \mu_1 < \mu_2$: Nếu $T_{tn} \geq -t_{(n_1+n_2-1;\alpha)}$ thì chấp nhận H_0 , bác bỏ H_1 .

3.4.2 Kiểm định (so sánh) hai tỉ lệ

Gọi p_1, p_2 lần lượt là tỉ lệ cá thể có tính chất A (cá thể loại A) trong đám đông \mathcal{P}_1 và \mathcal{P}_2 , bài toán kiểm định giả thuyết thống kê như sau:

- + Giả thuyết $H_0 : p_1 = p_2$;
- + Đối thuyết 2 phía $H_1 : p_1 \neq p_2$;
- + Hoặc đối thuyết 1 phía $H_1 : p_1 > p_2$ hoặc $H_1 : p_1 < p_2$.

Ta tiến hành lấy hai mẫu ngẫu nhiên một cách độc lập trên \mathcal{P}_1 và \mathcal{P}_2 với kích thước mẫu tương ứng là n_1 và n_2 ; đếm số cá thể loại A trong mẫu 1 và mẫu 2, giả sử tương ứng là m_1, m_2 . Với mức ý nghĩa α , bài toán kiểm định gồm các bước sau:

a) Kiểm định hai phía:

Quy tắc thực hành:

+ B_1 : Tính tần suất $f_1 = \frac{m_1}{n_1}$, $f_2 = \frac{m_2}{n_2}$ và tần suất chung $f = \frac{m_1 + m_2}{n_1 + n_2}$.

+ B_2 : Xác định phân vị $u_{\frac{\alpha}{2}}$.

+ B_3 : Tính giá trị thực nghiệm: $T_{tn} = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$.

+ B_4 : So sánh: Nếu $|T_{tn}| \leq u_{\frac{\alpha}{2}}$ thì chấp nhận H_0 , bác bỏ H_1 .

Ví dụ 3.4.3. Chữa trị cho cùng một loại bệnh bởi hai loại thuốc khác nhau A và B. Dùng thuốc A cho 200 bệnh nhân thì có 150 người khỏi bệnh, dùng thuốc B cho 100 bệnh nhân thì có 72 người khỏi bệnh. Với mức ý nghĩa 5%, hãy cho biết tác dụng của hai loại thuốc trên (tỉ lệ chữa khỏi bệnh) có như nhau không?

Giải. Bài toán kiểm định $H_0 : p_1 = p_2$; $H_1 : p_1 \neq p_2$.

$n_1 = 200$, $m_1 = 150$, $f_1 = \frac{m_1}{n_1} = 0,75$; $n_2 = 100$, $m_2 = 72$, $f_2 = \frac{m_2}{n_2} = 0,72$.

$f = \frac{m_1 + m_2}{n_1 + n_2} = 0,74$; $\alpha = 0,05 \Rightarrow u_{\frac{\alpha}{2}} = 1,96$.

Ta có

$$T_{tn} = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0,5584.$$

Vì $|T_{tn}| = 0,5584 < u_{\frac{\alpha}{2}} = 1,96$ nên chấp nhận H_0 , bác bỏ H_1 .

Vậy tác dụng của hai loại thuốc trên là như nhau.

b) Kiểm định một phía: Lớn hơn ($H_1 : p_1 > p_2$) và bé hơn ($H_1 : p_1 < p_2$)

Trường hợp $H_1 : p_1 > p_2$: Nếu $T_{tn} \leq u_{\alpha}$ thì chấp nhận H_0 , bác bỏ H_1 .

Trường hợp $H_1 : p_1 < p_2$: Nếu $T_{tn} \geq -u_{\alpha}$ thì chấp nhận H_0 , bác bỏ H_1 .

Ví dụ 3.4.4. Quan sát về chiều cao của các cây con trong 2 vườn cây giống I và II. Nếu chiều cao của cây con từ 0,5m trở lên thì xem như cây đó đủ điều kiện để mang đến nơi trồng. Tại vườn cây I người ta kiểm tra 50 cây con thì có 35 cây đủ điều kiện trồng. Tại vườn cây II người ta quan sát 75 cây thì có 55 cây không đủ điều kiện trồng.

Với mức ý nghĩa 5%, hãy cho biết tỉ lệ cây đủ điều kiện trồng ở vườn cây I có thấp hơn ở vườn cây II hay không.

Kết quả: $H_1 : p_1 < p_2$; $f_1 = 0,7$; $f_2 = 0,7333$; $f = 0,72$; $u_{\alpha} = 1,65$; $T_{tn} = -0,4062 > -u_{\alpha}$; Chấp nhận H_0 .

3.5 Kiểm định phi tham số

3.5.1 Kiểm định một phân phối (kiểm định về sự phù hợp)

Để khảo sát một biến ngẫu nhiên X của đám đông \mathcal{P} , ta tiến hành chọn mẫu ngẫu nhiên kích thước n . Giả sử mẫu cụ thể gồm n kết quả thực nghiệm $\{x_1, x_2, \dots, x_n\}$ được chia thành k lớp dưới dạng bảng *tần số thực tế* (tức là bảng tần số thu được từ mẫu) như sau

X	L_1	L_2	L_3	\dots	L_k
Tần số n_i	n_1	n_2	n_3	\dots	n_k

Bảng *tần suất thực tế* của mẫu tương ứng là

X	L_1	L_2	L_3	\dots	L_k
Tần suất f_i	$f_1 = \frac{n_1}{n}$	$f_2 = \frac{n_2}{n}$	$f_3 = \frac{n_3}{n}$	\dots	$f_n = \frac{n_k}{n}$

(*)

Mặt khác, tiến hành xét một phân phối $F(x)$ nào đó có tính chất gần giống với tính chất thu được từ mẫu cụ thể trên, ở đây $F(x)$ là một phân phối đã biết có thể có hoặc không có tham số (ví dụ phân phối nhị thức $\mathbb{B}(n; p)$, phân phối chuẩn $\mathbb{N}(\mu; \sigma^2)$, ... hoặc phân phối xác suất theo một lý thuyết nào đó). Theo lý thuyết, nếu b.n.n X có phân phối $F(x)$ thì ta sẽ có bảng tần suất như sau gọi là bảng *tần suất lý thuyết*

X	L_1	L_2	L_3	\dots	L_k
Tần suất p_i	p_1	p_2	p_3	\dots	p_k

(**)

Vấn đề đặt ra ở đây là biến ngẫu nhiên X của đám đông mà ta đang xét có phân phối $F(x)$ hay không, tức là ta cần kiểm tra xem bảng tần suất thực tế (*) thu được từ mẫu có phù hợp với bảng tần suất lý thuyết (**) hay không.

Khi đó, bài toán kiểm định được đặt ra là:

$$\begin{cases} H_0 : X \sim F(x) \text{ (tần suất thực tế của } X \text{ phù hợp với tần suất lý thuyết)} \\ H_1 : X \not\sim F(x) \text{ (tần suất thực tế của } X \text{ không phù hợp với tần suất lý thuyết).} \end{cases}$$

a) Quy tắc thực hành:

+ B_1 : Tính các tần suất lý thuyết p_i .

Nếu $F(x)$ là phân phối rời rạc thì $p_i = \mathbb{P}(X = x_i)$.

Nếu $F(x)$ là phân phối liên tục thì $p_i = \mathbb{P}(x_i < X < x_{i+1})$, chọn $x_1 = -\infty$ và $x_2 = +\infty$.

+ B_2 : Tra bảng χ^2 , xác định phân vị $\chi^2_{(k-r-1;\alpha)}$ (k là số lớp, r là số tham số chưa biết của $F(x)$).

Nếu $X \sim \mathbb{B}(n; p)$, p chưa biết: Khi đó $r = 1$, ta ước lượng (ước lượng điểm) $p = \bar{x}$.

Nếu $X \sim P(\lambda)$, λ chưa biết: Khi đó $r = 1$, ta ước lượng $\lambda = \bar{x}$.

Nếu $X \sim \mathbb{N}(\mu; \sigma^2)$, μ và σ^2 chưa biết: Khi đó $r = 2$, ta ước lượng $\mu = \bar{x}$, $\sigma^2 = s^2$.

Nếu $X \sim U(a; b)$, μ và $(a; b)$ chưa biết: Khi đó $r = 2$, ta ước lượng a, b từ hệ thức $\frac{a+b}{2} = \bar{x}$, $\frac{(b-a)^2}{12} = s^2$.

+ B_4 : Tính giá trị thực nghiệm $T_{tn} = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$.

+ B_5 : So sánh: Nếu $T_{tn} \leq \chi^2_{(k-r-1;\alpha)}$ thì chấp nhận H_0 , bác bỏ H_1 . Ngược lại thì bác bỏ H_0 .

b) Chú ý: Để đảm bảo sai số hợp lý và thuận lợi cho việc tính toán, trong thực tế người ta thường chia lớp sau cho các tần số $n_i \geq 5, i = \overline{1, k}$. Như vậy nếu có các tần số “ < 5 ” thì ta cần gộp dồn các lớp đó lại đến khi tần số “ ≥ 5 ”.

Ví dụ 3.5.1. Trồng mỗi chậu hai cây hoa, số cây sống ghi trong bảng

X số cây sống trong chậu	0	1	2	Tổng
Số chậu n_i	248	190	62	500

Với mức ý nghĩa 5%, hãy cho biết X có phân phối nhị thức với tham số 2 và 0,3 hay không?

Giải. Bài toán kiểm định $H_0 : X \sim \mathbb{B}(2; 0, 3); \quad H_1 : X \sim \mathbb{B}(2; 0, 3)$.

Bảng tần suất theo lý thuyết

Số cây sống X	$n_1 = 0$	$n_2 = 1$	$n_3 = 2$
Tần suất lý thuyết p_i	$p_1 = C_2^0 \cdot 0 \cdot 3^0 \cdot 0,7^2$	$p_1 = C_2^1 \cdot 0,3 \cdot 0,7$	$p_3 = C_2^2 \cdot 0,3^2 \cdot 0,7^0$

Với $n = 500, k = 3$ lớp, $\alpha = 0,05, \chi_{(k-r-1;\alpha)}^2 = \chi_{(3-0-1;0,05)}^2 = 5,991$.

Giá trị thực nghiệm

$$T_{ln} = \frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2} + \frac{(n_3 - np_3)^2}{np_3} = 8,364.$$

Vì $T_{ln} > \chi_{(k-r-1,\alpha)}^2$ nên bác bỏ H_0 , chấp nhận H_1 . Vậy $X \sim \mathbb{B}(2; 0, 3)$.

Ví dụ 3.5.2. Tiến hành đo ngẫu nhiên 100 cây bạch đàn trong một lâm trường người ta thu được bảng số liệu sau đây. Với mức ý nghĩa 5%, hãy cho biết chiều cao của các cây bạch đàn của lâm trường có phân phối chuẩn không?

Chiều cao (mét)	Số cây	Chiều cao (mét)	Số cây
8,275-8,325	1	8,625-8,675	17
8,325-8,375	2	8,675-8,725	12
8,375-8,425	4	8,725-8,775	9
8,425-8,475	5	8,775-8,825	7
8,475-8,525	8	8,825-8,875	6
8,525-8,575	10	8,875-8,925	0
8,575-8,625	18	8,925-8,975	1

Giải. Bài toán kiểm định $H_0 : X \sim \mathbb{N}(\mu; \sigma^2); H_1 : X \sim \mathbb{N}(\mu; \sigma^2)$.

Ta có $\bar{x} = 8,63; s = 0,128$. Do μ và σ chưa biết nên ta chọn $\mu \approx 8,63$ và $\sigma \approx 0,128$.

Giữ nguyên các lớp có tần số “ ≥ 5 ”, gộp 3 lớp đầu lại, 3 lớp cuối lại, ta còn $k = 10$ lớp như sau.

Chiều cao X (mét)	Số cây	Chiều cao X (mét)	Số cây
$-\infty$ -8,425	7	8,625-8,675	17
8,425-8,475	5	8,675-8,725	12
8,475-8,525	8	8,725-8,775	9
8,525-8,575	10	8,775-8,825	7
8,575-8,625	18	8,825- $+\infty$	7

Nếu X có phân phối chuẩn $N(\mu, \sigma^2)$ thì tần suất theo lý thuyết được tính như sau

$$p_i = \mathbb{P}(x_i < X < x_{i+1}) = \Phi\left(\frac{x_{i+1} - \mu}{\sigma}\right) - \Phi\left(\frac{x_i - \mu}{\sigma}\right), \quad i = \overline{1, 10}. \quad (3.1)$$

Lần lượt thay các x_i , vào công thức (3.1) ta được

$p_1 = \mathbb{P}(x_1 < X < x_2) = \mathbb{P}(-\infty < X < 8,425) = \Phi\left(\frac{8,425 - \mu}{\sigma}\right) - \Phi\left(\frac{-\infty - \mu}{\sigma}\right) = \Phi\left(\frac{8,425 - 8,63}{0,128}\right) - 0 = \Phi(-1,6016) = 0,0546$. (tính $\Phi(x)$ bằng máy tính cầm tay hoặc tra bảng, nhưng tra bảng sẽ có nhiều sai số).

Tương tự, $p_2 = 0,05833$; $p_3 = 0,09306$; $p_4 = 0,12769$; $p_5 = 0,15071$; $p_6 = 0,153$; $p_7 = 0,13359$; $p_8 = 0,10034$; $p_9 = 0,06483$; $p_{10} = 0,06382$.

Với $k = 10$ lớp, số tham số chưa biết $r = 2$, ta có phân vị $\chi^2_{(k-r-1;\alpha)} = \chi^2_{(10-2-1;0,05)} = \chi^2_{(7;0,05)} = 14,0671$ (tra bảng phân phối χ^2).

$$\text{Giá trị thực nghiệm: } T_{in} = \sum_{i=1}^{10} \frac{(n_i - np_i)^2}{np_i} = \dots = 2,1992.$$

Vì $T_{in} < \chi^2_{(7;0,05)}$ nên chấp nhận H_0 , bác bỏ H_1 . Vậy chiều cao các cây bạch đàn trong lâm trường có phân phối chuẩn.

Ví dụ 3.5.3. (Sinh viên tự học) Người ta lai chéo hai giống cây khác nhau bởi hai cặp đặc tính Aa với Bb . Ở thế hệ đầu kết quả thu được khá thuần nhất. Ở thế hệ thứ hai xuất hiện bốn kiểu cây mà kiểu hình được đánh dấu bằng $A-B-$, $A-bb$, $aaB-$, $aabb$. Quan sát trên 160 cây thì thấy kiểu $A-B-$ có 100 cây, kiểu $A-bb$ có 18 cây, kiểu $aaB-$ có 24 cây, kiểu $aabb$ có 18 cây. Với mức ý nghĩa 5%, hãy cho biết kết quả trên có phù hợp với luật di truyền của Mendel không?

HD. H_0 : Phù hợp; H_1 : Không phù hợp. Cỡ mẫu $n = 160$. Theo luật di truyền của Mendel thì kiểu gen phải theo tỉ lệ $9 : 3 : 3 : 1$, tức là các tần suất lý thuyết phải là $p_1 = 9/16$; $p_2 = 3/16$; $p_3 = 3/16$; $p_4 = 1/16$;

$k = 4$; $r = 0$; $\chi^2(k-r-1; \alpha) = 7,8147$; $T_{in} = 13,51$; $T_{in} > \chi^2(k-r-1; \alpha)$. Bác bỏ H_0 . Không phù hợp.

3.5.2 Kiểm định về sự độc lập

a) **Bảng tương liên:** Xét hai biến ngẫu nhiên X và Y của đám đông \mathcal{P} . Giả sử biến ngẫu nhiên X chia thành k lớp S_1, S_2, \dots, S_k , biến Y chia thành l lớp L_1, L_2, \dots, L_l . Tiến hành chọn mẫu ngẫu nhiên để khảo sát đồng thời hai tính chất X và Y , ta thu được bảng số liệu thực nghiệm hai chiều sau đây gọi là *bảng tương liên*

$X \setminus Y$	L_1	L_2	...	L_l	Σ
S_1	n_{11}	n_{12}	...	n_{1l}	TH_1
S_2	n_{21}	n_{22}	...	n_{2l}	TH_2
...
S_k	n_{k1}	n_{k2}	...	n_{kl}	TH_k
Σ	TC_1	TC_2	...	TC_l	n

trong đó, n_{ij} là tần số của cặp (S_i, L_j) , $i = \overline{1, k}$, $j = \overline{1, l}$, tần số này được gọi là tần số thực tế vì chúng thu được qua thực tế.

Vấn đề đặt ra là hai biến ngẫu nhiên X và Y của đám đông \mathcal{P} độc lập hay phụ thuộc nhau. Bài toán kiểm định đặt ra

$$\begin{cases} H_0 : X \text{ và } Y \text{ độc lập nhau} \\ H_1 : X \text{ và } Y \text{ không độc lập (phụ thuộc)}. \end{cases}$$

b) Quy tắc thực hành:

+ B_1 : Xác định phân vị $\chi^2_{((k-1)(l-1); \alpha)}$ (là giá trị tại dòng $(k-1)(l-1)$, cột α , bảng phân phối χ^2)

+ B_2 : Tính giá trị thực nghiệm

$$T_{tn} = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - t_{ij})^2}{t_{ij}} = n \left[\sum_{i=1}^k \sum_{j=1}^l \frac{n_{ij}^2}{TH_i \cdot TC_j} - 1 \right] \quad (\text{với } t_{ij} = \frac{TH_i TC_j}{n}).$$

+ B_3 : So sánh: Nếu $T_{tn} \leq \chi^2_{((k-1)(l-1); \alpha)}$ thì chấp nhận H_0 , bác bỏ H_1 .

Ví dụ 3.5.4. Nghiên cứu ảnh hưởng thành phần thức ăn của bố mẹ (X) đối với giới tính Y của con cái ta được số liệu sau

$Y \setminus X$	Thiếu vitamin	Đủ vitamin
Trai	123	145
Gái	153	150

Hãy cho biết thành phần thức ăn có ảnh hưởng đến giới tính của con cái hay không, với mức ý nghĩa $\alpha = 5\%$

Giải. H_0 : " X và Y độc lập"; H_1 : " X và Y không độc lập".

$Y \setminus X$	Thiếu vitamin	Đủ vitamin	Σ
Trai	123	145	268
Gái	153	150	303
Σ	276	295	571

Ta có $k = 2, l = 2; \alpha = 0,05$, suy ra $\chi^2_{((k-1)(l-1); \alpha)} = \chi^2_{(1; 0,05)} = 3,8415$.

Giá trị thực nghiệm

$$T_{ln} = 571 \left[\frac{123^2}{276 \times 268} + \frac{145^2}{295 \times 268} + \frac{153^2}{276 \times 303} + \frac{150^2}{295 \times 303} - 1 \right] = 1,2048.$$

Vì $T_{ln} < \chi^2_{((2-1)(2-1); \alpha)}$ nên chấp nhận H_0 , bác bỏ H_1 .

Vậy thành phần thức ăn của bố mẹ độc lập với giới tính của con cái.

✚ Chú ý: Khi bảng số liệu có nhiều tần số (xem các bài tập cuối chương) thì việc tính toán trên máy tính cầm tay có thể bị tràn màn hình. Để khắc phục điều này, ta có thể tính riêng từng dòng (hoặc từng cột) và lần lượt lưu vào các biến nhớ A, B, C của máy, sau đó tính

$$T_{ln} = n \times [A + B + C - 1].$$

3.6 Tương quan, hồi quy tuyến tính

3.6.1 Mở đầu về tương quan tuyến tính

Trên cùng một đám đông \mathcal{P} có hai đặc điểm định lượng cần nghiên cứu, đó là hai biến ngẫu nhiên gốc của đám đông \mathcal{P} tương ứng lần lượt là X và Y .

Ta nói X và Y có *tương quan tuyến tính*⁴, nếu Y có thể biểu diễn xấp xỉ qua X dưới dạng phương trình tuyến tính (phương trình đường thẳng)

$$Y = AX + B, \text{ với } A, B \text{ là hằng số nào đó.}$$

Xấp xỉ trên có thể có sai số ít hoặc nhiều. Bài toán đặt ra ở đây là tìm hiểu mức độ tương quan tuyến tính giữa hai biến ngẫu nhiên X, Y và tìm biểu thức biểu diễn sự liên hệ giữa chúng, tức là tìm các hằng số A và B sao cho sai số bé nhất.

3.6.2 Hệ số tương quan tuyến tính thực nghiệm

Hệ số tương quan giữa hai biến ngẫu nhiên X và Y là

$$\rho(X, Y) = \frac{\mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y}{\sqrt{\mathbb{D}X\mathbb{D}Y}}$$

⁴Linear correlation.

Đó là số đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên X và Y , nhưng nếu chưa biết phân phối xác suất thì hệ số tương quan lý thuyết $\rho(X, Y)$ chưa xác định được. Do đó ta tìm cách ước lượng $\rho(X, Y)$ bởi một giá trị thu được từ mẫu quan sát, giá trị đó được gọi là *hệ số tương quan*.

Giả sử $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ là mẫu thu được từ vectơ ngẫu nhiên (X, Y) . Hệ số tương quan mẫu là biến ngẫu nhiên

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$

Hệ số tương quan thực nghiệm: Giả sử X_i, Y_i lần lượt nhận các giá trị thực nghiệm $x_i, y_i, i = 1, n$. Khi đó, hệ số tương quan mẫu thực nghiệm được tính theo công thức

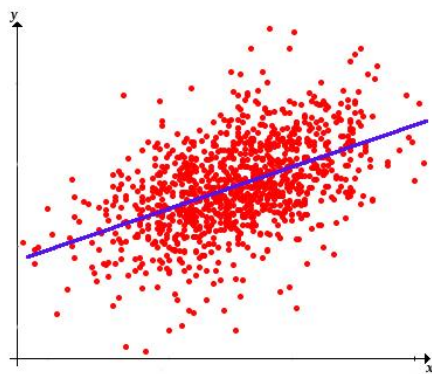
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}.$$

Viết gọn lại ta có công thức tương đương,

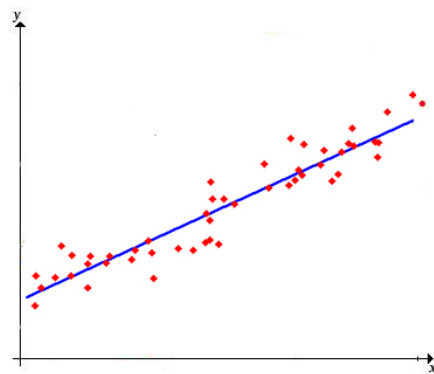
$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\widehat{s_x} \widehat{s_y}}$$

trong đó $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$.

Hệ số tương quan mẫu có tính chất $|r| \leq 1$. Biểu diễn các cặp (x_i, y_i) của mẫu lên một mặt phẳng tọa độ tạo thành đám mây điểm thể hiện mối quan hệ giữa X và Y .



$r \approx 0$



$|r| \approx 1$

- Nếu đám mây điểm có xu hướng tập trung quanh một đường thẳng nào đó (có hệ số góc khác 0) thì $|r|$ càng gần 1 và ta có thể kết luận X, Y có quan hệ gần với quan hệ tuyến tính (tương quan tuyến tính).
- Nếu đám mây điểm phân tán một cách không có quy luật thành hình êlip hay hình vuông thì r gần bằng 0.

Chú ý, xem hướng dẫn cách tính trực tiếp hệ số tương quan mẫu r_{xy} (trang 87).



3.7 Phương trình hồi quy tuyến tính thực nghiệm

Phương trình hồi quy tuyến tính thực nghiệm của Y theo X có dạng $y = ax + b$ với a và b được xác định bởi công thức:

$$\begin{cases} a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}; \\ b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n}. \end{cases}$$

Viết gọn lại ta có công thức tương đương

$$\begin{cases} a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\widehat{s}_x^2}, \\ b = \bar{y} - a\bar{x}, \end{cases} \text{ trong đó, } \overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i.$$

Chú ý, xem hướng dẫn cách tính trực tiếp các hệ số a, b (trang 87).



Ví dụ 3.7.1. Cho bảng số liệu của một công ty về mức doanh thu X (tỉ đồng) và số tiền dành cho quảng cáo Y (triệu đồng) của một số tháng như sau:

X	5	7	8	11	9
Y	45	60	75	90	80

- Hãy xác định hệ số tương quan mẫu r_{xy} .
- Tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x .
- Nếu doanh thu của một tháng nào đó là 10 tỉ đồng, hãy dự báo chi phí quảng cáo của công ty tháng đó là bao nhiêu.

Giải. a) Ta có $n = 5$,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 8; \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 70; \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i = 591.$$

$$\hat{s}_x^2 = \overline{x^2} - (\bar{x})^2 = 4 \Rightarrow \hat{s}_x = 2.$$

$$\hat{s}_y^2 = \overline{y^2} - (\bar{y})^2 = 250 \Rightarrow \hat{s}_y = 15,8114.$$

Vậy hệ số tương quan mẫu thực nghiệm $r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{s}_x \hat{s}_y} = 0,9803$. Vì giá trị của $|r_{xy}| = 0,9803$ gần 1 nên chi phí dành cho quảng cáo và doanh thu có tương quan chặt.

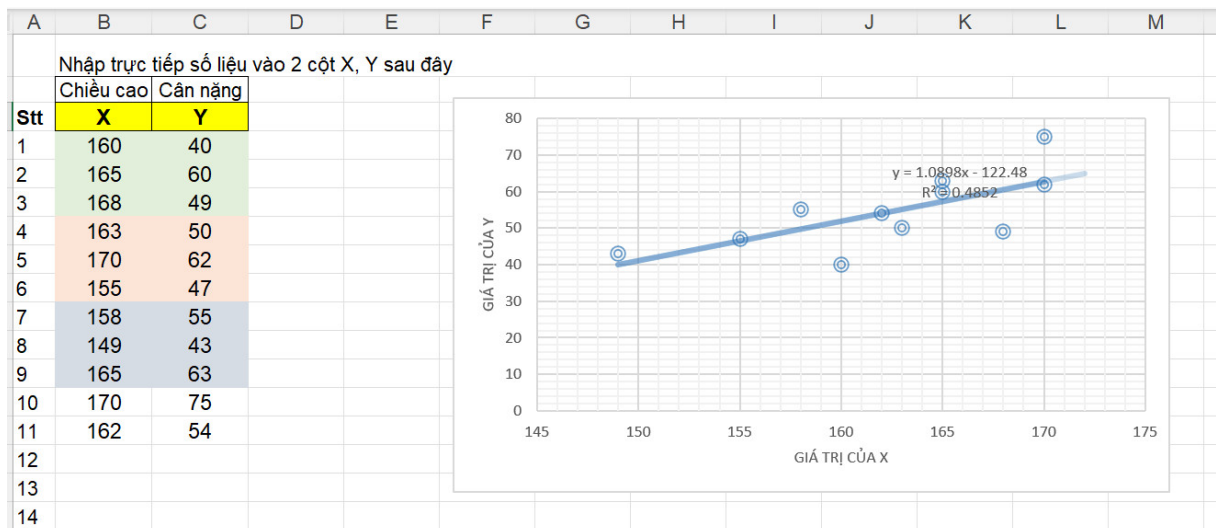
b) Phương trình hồi quy tuyến tính thực nghiệm của y theo x là $y = ax + b$ với

$$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{s}_x^2} = 7,75; \quad b = \bar{y} - a\bar{x} = 8.$$

Vậy ta có $y = 7,75x + 8$.

c) Ta có $x = 10$ suy ra $y = 7,75x + 8 = 85,5$. Vậy chi phí quảng cáo của tháng đó vào khoảng 85,5 triệu đồng.

Ghi chú: Phần mềm Excel có chức năng hiển thị đường thẳng xu hướng (trendline) $y = ax + b$, đường thẳng này thay đổi tự động mỗi khi ta nhập từng cặp giá trị (x_i, y_i) mới vào mẫu dữ liệu thực nghiệm.



BÀI TẬP CHƯƠNG 3

Kiểm định về một giá trị trung bình, một tỉ lệ:

- 3.1. Theo hợp đồng bán gạo, gạo được đóng gói trong bao 50kg. Kiểm tra ngẫu nhiên 16 bao được $\bar{x} = 49,75\text{kg}$. Giả sử phương sai của các bao gạo đã biết là $\sigma^2 = 6.4$. Giả thuyết rằng trọng lượng gạo đóng vào bao là biến ngẫu nhiên có phân phối chuẩn. Hỏi hợp đồng có được bên bán thực hiện đúng hay không, với mức ý nghĩa 5%?

HD. $H_0 : \mu = 50, H_1 : \mu \neq 50$. Đã biết phương sai σ^2 . Tìm giá trị tới hạn, tra bảng phân phối chuẩn.

- 3.2. Một trường đại học khẳng định rằng thời gian trung bình mà sinh viên ngành khoa học máy tính dành để học lập trình mỗi tuần là 10 giờ. Một nhóm nghiên cứu đã khảo sát ngẫu nhiên 36 sinh viên và thu được kết quả thời gian học tập trung bình là 11,2 giờ với độ lệch chuẩn mẫu là 2,5 giờ.

Kiểm tra xem có sự khác biệt về thời gian trung bình mà sinh viên ngành khoa học máy tính dành để học lập trình có ý nghĩa thống kê hay không, với mức ý nghĩa 10%.

HD. $H_0 : \mu = 10, H_1 : \mu \neq 10$. Chưa biết phương sai σ^2 , cỡ mẫu lớn. Tìm giá trị tới hạn, tra bảng phân phối chuẩn tắc.

- 3.3. Một giảng viên tin học cho rằng thời gian trung bình để sinh viên hoàn thành một bài tập lập trình là 3 giờ. Một khảo sát với 25 sinh viên cho thấy thời gian hoàn thành trung bình là 3,5 giờ với độ lệch chuẩn mẫu là 0,8 giờ.

Sử dụng kiểm định T với mức ý nghĩa 1% để kiểm tra xem có bằng chứng cho thấy thời gian hoàn thành trung bình là cao hơn 3 giờ hay không.

HD. $H_0 : \mu = 3, H_1 : \mu > 3$. Chưa biết phương sai σ^2 , cỡ mẫu nhỏ. Tìm giá trị tới hạn, tra bảng phân phối t (Student).

- 3.4. Một khảo sát đã chỉ ra rằng 60% sinh viên ngành công nghệ thông tin tham gia vào các khóa học trực tuyến ngoài giờ học chính. Trong một mẫu ngẫu nhiên gồm 200 sinh viên, có 130 người tham gia các khóa học trực tuyến.

Hãy sử dụng kiểm định Z để kiểm tra giả thuyết rằng tỉ lệ sinh viên tham gia các khóa học trực tuyến là 60%, với mức ý nghĩa 5%.

HD. Kiểm định về tỉ lệ, với $H_0 : p = 0.6, H_1 : p \neq 0.6. f = m/n, n = ?, m = ?$.

- 3.5. Công ty xe buýt nói rằng cứ trung bình 5 phút lại có một chuyến xe. Chọn ngẫu nhiên 8 thời điểm và ghi lại thời gian giữa hai chuyến xe buýt ta thu được số liệu sau:

5,3 4,5 4,8 5,1 4,3 4,8 4,9 4,7

Với mức ý nghĩa 4%, nhận định xem công ty xe buýt nói có đúng không? Giả sử rằng thời gian chờ để có xe buýt khác đến trạm (tính từ xe buýt liền trước) là biến ngẫu nhiên có phân phối chuẩn.

- 3.6. Một nghiên cứu chỉ ra rằng 40% người dùng internet sử dụng phần mềm chống virus. Một khảo sát với 500 người dùng internet cho thấy 230 người sử dụng phần mềm chống virus.

Sử dụng kiểm định Z để kiểm tra giả thuyết rằng tỉ lệ người dùng sử dụng phần mềm chống virus có cao hơn mức 40% hay không, với mức ý nghĩa 5%.

Kiểm định về một hai giá trị trung bình, hai tỉ lệ:

- 3.7. Một công ty muốn so sánh năng suất làm việc của hai nhóm lập trình viên sử dụng các ngôn ngữ lập trình khác nhau. Nhóm A sử dụng Python, nhóm B sử dụng Java. Kết quả khảo sát cho thấy nhóm A ($n = 15$) có thời gian trung bình hoàn thành nhiệm vụ là 5 giờ với độ lệch chuẩn mẫu 1,2 giờ, trong khi nhóm B ($n = 15$) có thời gian trung bình là 6 giờ với độ lệch chuẩn mẫu 1,5 giờ.

Sử dụng kiểm định t để kiểm tra xem có sự khác biệt có ý nghĩa thống kê giữa hai nhóm không với mức ý nghĩa 5%.

- 3.8. Một nghiên cứu so sánh hiệu suất làm việc của các lập trình viên khi làm việc tại văn phòng và làm việc từ xa. Kết quả khảo sát từ 30 lập trình viên làm việc tại văn phòng có thời gian trung bình hoàn thành công việc là 8 giờ với độ lệch chuẩn mẫu 1,2 giờ, trong khi đó 30 lập trình viên làm việc từ xa có thời gian trung bình hoàn thành công việc là 7,5 giờ với độ lệch chuẩn mẫu 1,0 giờ.

Kiểm tra xem có sự khác biệt có ý nghĩa thống kê giữa hai giá trị trung bình không với mức ý nghĩa 5%.

- 3.9. Một công ty công nghệ muốn so sánh tỉ lệ nhân viên tham gia các khóa học phát triển kỹ năng giữa hai bộ phận: IT và Marketing. Trong 100 nhân viên IT, 70 người tham gia khóa học, trong khi 100 nhân viên Marketing chỉ có 50 người tham gia. Hãy kiểm định để kết luận rằng tỉ lệ tham gia khóa học của bộ phận IT có cao hơn so với bộ phận Marketing hay không (với $\alpha = 0,05$).

- 3.10. Một cuộc khảo sát cho thấy rằng 70% nhân viên IT sử dụng máy tính xách tay và 60% nhân viên IT sử dụng máy tính bàn. Một khảo sát ngẫu nhiên với 100 nhân viên mỗi nhóm cho kết quả: 75 nhân viên sử dụng máy tính xách tay và 58 nhân viên sử dụng máy tính bàn.

Hãy sử dụng kiểm định Z để kiểm tra xem có sự khác biệt có ý nghĩa thống kê giữa tỉ lệ sử dụng máy tính xách tay và máy tính bàn hay không, với mức ý nghĩa 5%.

Kiểm định về sự phù hợp, luật phân phối:

- 3.11. Tỷ lệ nhóm máu trong dân số theo hằng số sinh học như sau: 18% nhóm máu A; 28% nhóm máu B; 5% nhóm máu AB và 49% nhóm máu O. Một mẫu 500 người được kiểm tra nhóm máu và cho kết quả sau:

Nhóm máu	A	B	AB	O
Số người	75	150	15	260

Với mức ý nghĩa 5%, hãy cho biết tỷ lệ nhóm máu có phù hợp với quy luật trên hay không?

HD. Kiểm định về sự phù hợp.

- 3.12. Một cuộc khảo sát được thực hiện để xem liệu có mối quan hệ giữa việc sở hữu máy tính xách tay và kết quả học tập của sinh viên ngành công nghệ thông tin. Dữ liệu được thu thập như sau:

Tình trạng về laptop	Kết quả học tập tốt	Kết quả học tập kém
Có laptop	80	20
Không laptop	30	70

Hãy sử dụng kiểm định Chi-square (χ^2) để kiểm tra giả thuyết rằng việc sở hữu máy tính xách tay không liên quan đến kết quả học tập, với mức ý nghĩa 10%.

- 3.13. Một nghiên cứu khảo sát mối quan hệ giữa việc sử dụng phần mềm mã nguồn mở và việc đóng góp mã nguồn mở. Dữ liệu thu thập được như sau:

Đóng góp mã nguồn mở	Sử dụng phần mềm mã nguồn mở	Không sử dụng phần mềm mã nguồn mở
Có	100	20
Không	50	80

Cho biết việc sử dụng phần mềm mã nguồn mở có liên quan đến việc đóng góp mã nguồn mở hay không, với mức ý nghĩa 5%.

- 3.14. Một nghiên cứu khảo sát mối quan hệ giữa ngành học và sự quan tâm đến việc học máy (machine learning). Dữ liệu thu được như sau:

Ngành học	Quan tâm	Không quan tâm
Khoa học máy tính	50	10
Công nghệ thông tin	30	20

Hãy sử dụng kiểm định Chi-square (χ^2) để kiểm tra xem sự quan tâm đến việc học máy có liên quan đến ngành học hay không, với mức ý nghĩa 5%.

Tương quan, hồi quy tuyến tính:

- 3.15. Một công ty phát triển ứng dụng di động muốn dự đoán số lượt tải ứng dụng dựa trên số lượt xem quảng cáo. Dữ liệu thu thập từ 10 chiến dịch quảng cáo như sau:

Số lượt xem quảng cáo (x) Số lượt tải ứng dụng (y)

1000 50

2000 80

1500 60

2500 100

3000 120

4000 150

3500 140

4500 170

5000 200

5500 220

a) Tìm hệ số tương quan mẫu r_{xy} (ghi ra công thức khi tính toán).

b) Viết phương trình hồi quy tuyến tính dự đoán số lượt tải ứng dụng (y) dựa trên số lượt xem quảng cáo (x).

- 3.16. Một công ty phát triển phần mềm muốn dự đoán số giờ làm việc dựa trên số dòng mã đã viết. Dữ liệu thu thập từ 10 lập trình viên như sau:

Số dòng mã (x) Số giờ làm việc (y)

100 3

200 5

150 4

300 6

250 5

400 8

350 7

450 9

500 10

550 11

a) Tìm hệ số tương quan mẫu r_{xy} (ghi ra công thức khi tính toán).

b) Viết phương trình hồi quy tuyến tính dự đoán số giờ làm việc (y) dựa trên số dòng mã (x).

PHỤ LỤC

Phụ lục 1: Dùng máy tính cầm tay CASIO fx570VN Plus, CASIO fx580VNX hỗ trợ xử lý số liệu thống kê

A. GIÁ TRỊ HÀM PHÂN PHỐI CHUẨN:

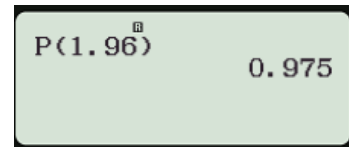
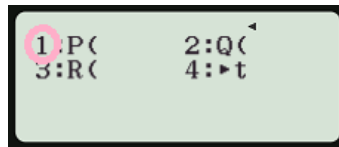
Tính giá trị của hàm phân phối chuẩn tắc $N(0, 1) : \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt, x \in \mathbb{R}$.

- Thao tác tính $\Phi(1.96)$ trên máy Casio fx 580VN X:

Bước 1: Vào chế độ thống kê **Menu** **6** **=** **1** **ON**.

Bước 2: **OPTN**, bấm mũi tên xuống **▼**, chọn **4** (Norm Distr), chọn **1**, nhập **1.96** **=**

Kết quả: 0.975.



Chú ý, trong các chức năng “P(, Q(, R(” nêu trên thì, với $a \in \mathbb{R}$, ta có:

$$P(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-\frac{t^2}{2}} dt. \text{ Ví dụ: } P(1.96) = 0.975.$$

$$Q(a) = \frac{1}{\sqrt{2\pi}} \int_0^a e^{-\frac{t^2}{2}} dt. \text{ Ví dụ: } Q(1.96) = 0.475$$

$$R(a) = 1 - P(a) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-\frac{t^2}{2}} dt. \text{ Ví dụ: } R(1.96) = 0.024998.$$

- Thao tác tính $\Phi(1.96)$, trên máy Casio fx 570VN Plus:

Bước 1: Vào chế độ thống kê, **Mode** **3** **1** **ON**.

Bước 2: **Shift** **STAT** **5** **1**, nhập **1.96** **=**.

Kết quả: 0.975.

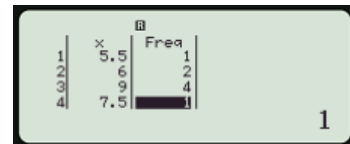
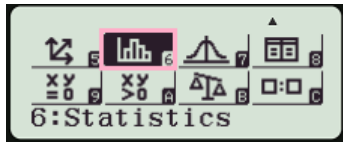
B. TÍNH CÁC SỐ ĐẶC TRƯNG MẪU

Tính các số đặc trưng mẫu $\bar{x}, \hat{s}^2, \hat{s}, s^2, s$ của mẫu số liệu thực nghiệm được biểu diễn dưới dạng bảng tần số như sau

X_i	x_1	x_2	...	x_k
Tần số n_i	n_1	n_2	...	n_k

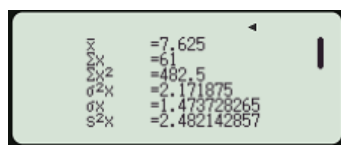
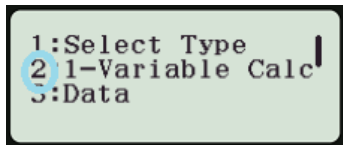
• **Thao tác trên máy Casio fx 580VN X:**

Bước 1: Vào chế độ thống kê **Menu** **6** **=** **1**. Nhập số liệu vào, di chuyển mũi tên phải để nhập cột tần số. Sau khi nhập xong, nếu dò lại số liệu đã chính xác thì bấm **ON**.



Chú ý, nếu máy không hiện ra cột tần số Freq thì chỉnh lại cài đặt bằng thao tác **Shift** **Setup** **▼** **3** **1**.

Bước 2: Gọi dữ liệu ra **OPTN**, chọn **2**.



• **Thao tác trên máy Casio fx 570VN Plus:**

Bước 1 (Vào chế độ thống kê **STAT**):

Thao tác **Mode** **3** **1** (chọn STAT, 1-Var).

Màn hình hiện ra hai cột: Cột nhập giá trị x_i và cột nhập tần số n_i

	X	Freq
1	—	—
2	—	—
⋮	—	—

○ **Chú ý:** Nếu máy không hiện ra cột tần số Freq như trên thì chỉnh lại cài đặt bằng thao tác **Shift** **Setup** **▼** **4** **1** (chọn STAT, chọn ON).

Bước 2 (Nhập dữ liệu vào máy):

Tại con trỏ, ta nhập giá trị của x và ấn **=**. Dùng phím di chuyển con trỏ qua lại **◀** và lên xuống **▲** **▼** để nhập và sửa số liệu. Khi thấy các số liệu đã hoàn toàn đúng, ấn **ON**, lúc này máy đã lưu lại số liệu vừa nhập, hoặc nếu bạn tắt máy thì số liệu vẫn còn lưu lại.

Bước 3 (Gọi dữ liệu ra):

Thao tác **Shift** **STAT** **4** (Var), máy hiện ra

1: n	2: \bar{x}	chọn →	1 [=] để tính cỡ mẫu n
3: σ_x	4: s_x		2 [=] để tính trung bình mẫu \bar{x}
			3 [=] để tính độ lệch mẫu chưa hiệu chỉnh \hat{s}
			4 [=] để tính độ lệch mẫu đã hiệu chỉnh s

○ *Chú ý:*

Để tính \hat{s}^2 , ấn [Shift] [STAT] 4 3 x^2 (phím bình phương) [=].

Để tính s^2 , ấn [Shift] [STAT] 4 4 x^2 [=]

Tương tự đối với thao tác [Shift] [STAT] 3 (Sum), máy hiện ra

1: $\sum x^2$	2: $\sum x$	chọn →	1 để tính $\sum_i x_i^2 n_i$
			2 để tính $\sum_i x_i n_i$

C. Ví dụ:

Cho mẫu số liệu về trọng lượng X (đơn vị kg) của 15 sinh viên nam trong khoa được chọn ngẫu nhiên để đo, kết quả về trọng lượng được cho bởi bảng tần số như sau

$X(kg)$	47	49	52	55	60
Tần số n_i	1	3	4	5	2

Hãy tính các số đặc trưng mẫu $\bar{x}, \hat{s}^2, \hat{s}, s^2, s$ của mẫu số liệu thực nghiệm trên.

Giải. Ta có, $n = 15$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^5 x_i n_i = 53,1333;$$

$$\hat{s}^2 = \frac{1}{n} \sum_{i=1}^5 (x_i - \bar{x})^2 n_i = 13,7156 \Rightarrow \hat{s} = \sqrt{\hat{s}^2} = 3,7035$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^5 (x_i - \bar{x})^2 n_i = 14,6953 \Rightarrow s = \sqrt{s^2} = 3,8334.$$

Yêu cầu các bạn thao tác lại trên máy để kiểm tra kết quả của mình.

D. TƯƠNG QUAN, HỒI QUY TUYẾN TÍNH:

Tìm hệ số tương quan mẫu r_{xy} , lập phương trình hồi quy tuyến tính của y theo x đối với bảng số liệu thu được của hai biến ngẫu nhiên X, Y như sau

Không có tần số:	X	x_1	x_2	...	x_k	Có tần số:	X	x_1	x_2	...	x_k
	Y	y_1	y_2	...	y_k		Y	y_1	y_2	...	y_k
							Tần số n_i	n_1	n_2	...	n_k

• Thao tác trên máy Casio fx 570VN Plus:

Bước 1 (Vào chế độ hồi quy tuyến tính **REG**):

Mode **3** (chọn STAT) **2** (chọn $A + Bx$).

Màn hình hiện ra ba cột: Cột nhập giá trị x_i , y_i và n_i

	X	Y	Freq
1	—	—	—
2	—	—	—
\vdots	—	—	—

Chú ý, nếu không hiện ra cột tần số Freq như trên thì cài đặt lại bằng thao tác **Shift** **Setup** **▼** **4** (chọn STAT) **1** (chọn ON).

Bước 2 (Nhập dữ liệu vào máy):

Nhập từng giá trị vào máy và ấn $\boxed{=}$ để lưu. Dùng phím di chuyển con trỏ qua lại $\boxed{\blacktriangleright}$ $\boxed{\blacktriangleleft}$ và lên xuống $\boxed{\blacktriangleup}$ $\boxed{\blacktriangledown}$ để nhập và sửa số liệu. Khi thấy các số liệu đã hoàn toàn đúng, ấn **ON**, lúc này máy đã lưu lại số liệu vừa nhập.

Bước 3 (Gọi dữ liệu ra):

Shift	STAT 3	(Sum)	Shift	STAT 4	(Var)	Shift	STAT 5	(Reg)
1:	$\sum x^2$	2: $\sum x$	1:	n	2: \bar{x}	1:	A	2: B
3:	$\sum y^2$	4: $\sum y$	3:	\hat{s}_x	4: s_x	3:	r_{xy}	4: \hat{x}
5:	$\sum xy$	6: $\sum x^3$	5:	\bar{y}	6: \hat{s}_y	5:	\hat{y}	
7:	$\sum x^2 y$	8: $\sum x^4$	7:	s_y				

• **Ví dụ:** Cho bảng số liệu của một công ty về mức doanh thu X (tỉ đồng) và số tiền dành cho quảng cáo Y (triệu đồng) của một số tháng như sau

X	5	7	8	11	9
Y	45	60	75	90	80

a) Hãy xác định hệ số tương quan mẫu r_{xy} .

b) Tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x .

Giải. a) $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 8$; $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 70$; $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i = \frac{1}{5} \cdot 2955 = 591$.

$$\hat{s}_x^2 = \overline{x^2} - (\bar{x})^2 = 4 \Rightarrow \hat{s}_x = 2;$$

$$\hat{s}_y^2 = \overline{y^2} - (\bar{y})^2 = 250 \Rightarrow \hat{s}_y = 15.8114$$

$$\text{Vậy } r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{s}_x \hat{s}_y} = 0,9803.$$

b) Phương trình hồi quy tuyến tính của y theo x là $y = ax + b$ (phương trình của máy $y = A + Bx$)

$$\begin{cases} a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{s}_x^2}; \\ b = \bar{y} - b\bar{x} \end{cases} \Rightarrow \begin{cases} a = 7,75 & (\text{giá trị } B \text{ trên máy tính}) \\ b = 8 & (\text{giá trị } A \text{ trên máy tính}) \end{cases}$$

Vậy phương trình hồi quy tuyến tính thực nghiệm của y theo x là : $y = 7,75x + 8$.

Phụ lục 2: Bảng phân phối chuẩn tắc $N(0, 1)$: $\mathbb{P}[X < x] = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,88849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4,0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Phụ lục 3: Bảng phân phối t (Student): $X \sim t(n) \Rightarrow \mathbb{P}[X > t_{(n,\alpha)}] = \alpha$.

n	Mức ý nghĩa α									
	0,4	0,3	0,2	0,1	0,05	0,025	0,02	0,01	0,005	0,001
1	0,325	0,727	1,376	3,078	6,314	12,706	15,895	31,821	63,675	318,309
2	0,289	0,617	1,061	1,886	2,920	4,303	4,489	6,965	9,925	22,327
3	0,277	0,584	0,978	1,638	2,353	3,182	3,482	4,541	5,841	10,215
4	0,271	0,569	0,941	1,533	2,132	2,766	2,999	3,747	4,604	7,173
5	0,267	0,559	0,920	1,467	2,015	2,571	2,757	3,365	4,032	5,893
6	0,265	0,553	0,906	1,440	1,943	2,477	2,612	3,143	3,707	5,208
7	0,263	0,549	0,896	1,415	1,895	2,365	2,517	2,998	3,499	4,785
8	0,262	0,546	0,889	1,397	1,860	2,306	2,449	2,896	3,355	4,501
9	0,261	0,543	0,883	1,383	1,833	2,262	2,398	2,821	3,250	4,297
10	0,260	0,542	0,879	1,372	1,812	2,228	2,359	2,764	3,169	4,144
11	0,260	0,540	0,876	1,363	1,796	2,201	2,328	2,718	3,106	4,025
12	0,259	0,539	0,873	1,356	1,782	2,179	2,303	2,861	3,055	3,930
13	0,259	0,538	0,870	1,350	1,771	2,160	2,282	2,650	3,012	3,852
14	0,258	0,537	0,868	1,345	1,761	2,145	2,264	2,624	2,977	3,787
15	0,258	0,536	0,866	1,341	1,753	2,131	2,249	2,602	2,947	3,733
16	0,258	0,535	0,865	1,337	1,746	2,120	2,235	2,583	2,921	3,686
17	0,257	0,534	0,863	1,333	1,740	2,110	2,224	2,567	2,898	3,646
18	0,257	0,534	0,862	1,330	1,734	2,101	2,214	2,552	2,878	3,610
19	0,257	0,533	0,861	1,328	1,729	2,093	2,205	2,539	2,861	3,579
20	0,257	0,533	0,860	1,325	1,725	2,086	2,197	2,528	2,845	3,552
21	0,257	0,532	0,859	1,323	1,721	2,080	2,189	2,518	2,831	3,527
22	0,256	0,532	0,858	1,321	1,717	2,074	2,183	2,508	2,819	3,505
23	0,256	0,532	0,858	1,319	1,714	2,069	2,177	2,500	2,807	3,485
24	0,256	0,531	0,857	1,318	1,711	2,064	2,172	2,492	2,797	3,467
25	0,256	0,531	0,856	1,316	1,708	2,060	2,167	2,485	2,787	3,450
26	0,256	0,531	0,856	1,315	1,706	2,056	2,162	2,479	2,779	3,435
27	0,256	0,531	0,855	1,314	1,703	2,052	2,158	2,473	2,771	3,421
28	0,256	0,530	0,855	1,313	1,701	2,048	2,154	2,467	2,763	3,408
29	0,256	0,530	0,854	1,311	1,699	2,045	2,150	2,462	2,756	3,396
30	0,256	0,530	0,854	1,310	1,697	2,042	2,147	2,457	2,750	3,385
40	0,255	0,529	0,851	1,303	1,684	2,021	2,123	2,423	2,704	3,307
50	0,255	0,528	0,849	1,299	1,676	2,009	2,109	2,403	2,678	3,261
60	0,254	0,527	0,848	1,296	1,671	2,000	2,099	2,390	2,660	3,232
70	0,254	0,527	0,847	1,294	1,667	1,994	2,093	2,381	2,648	3,211
80	0,254	0,526	0,846	1,292	1,664	1,990	2,088	2,374	2,639	3,195
90	0,254	0,526	0,846	1,291	1,662	1,987	2,084	2,368	2,632	3,183
100	0,254	0,526	0,845	1,290	1,660	1,984	2,081	2,364	2,626	3,174

Khi bậc tự do n càng lớn thì $t_{(n-1;\alpha)}$ gần bằng u_α .

Phụ lục 4: Bảng phân phối chi (khi) bình phương: $X \sim \chi^2(n) \Rightarrow \mathbb{P}[X > \chi^2_{(n,\alpha)}] = \alpha$.

n	Mức ý nghĩa α								
	0,99	0,95	0,90	0,50	0,10	0,05	0,02	0,01	0,001
1	0,0002	0,0039	0,0158	0,4549	2,7055	3,8415	5,4119	6,6349	10,8276
2	0,0201	0,1026	0,2107	1,3863	4,6052	5,9915	7,8240	9,2103	1,8155
3	0,1448	0,3518	0,5844	2,3660	6,2514	7,8147	9,8374	11,3449	16,2662
4	0,2971	0,7107	1,0636	3,3567	7,7794	9,4877	11,6678	13,2767	18,4668
5	0,5543	1,1455	1,6103	4,3515	9,2364	11,0705	13,3882	15,0863	20,5150
6	0,8721	1,6354	2,2041	5,3481	10,6446	12,5916	15,0332	16,8119	22,4577
7	1,2390	2,1673	2,8331	6,3458	12,0170	14,0671	16,6224	18,4753	24,3219
8	1,6465	2,7326	3,4895	7,3441	13,3616	15,5073	18,1682	20,0902	26,1245
9	2,0879	3,3251	4,1682	8,3428	14,6837	16,9190	19,6790	21,6660	27,8772
10	2,5582	3,9403	4,8652	9,3418	15,9872	18,3070	21,1608	23,2093	27,8772
11	3,0535	4,5748	5,5778	10,3410	17,2750	19,6751	22,6179	24,7250	31,2641
12	3,5706	5,2260	6,3038	11,3403	18,5493	21,0261	24,0540	26,2170	32,9095
13	4,1069	5,8919	7,0415	12,3398	19,8119	22,2620	25,4715	27,6882	34,5282
14	4,6604	6,5706	7,7895	13,3393	21,0641	23,6848	26,8728	29,1412	36,1233
15	5,2293	7,2609	8,5468	14,3389	22,3071	24,9958	28,2595	30,5779	37,6973
16	5,8122	7,9616	9,3122	15,3385	23,5418	26,2962	29,6332	31,9999	39,2524
17	6,4078	8,6718	10,0852	16,3382	24,7690	27,5871	30,9950	33,4087	40,7902
18	7,0149	9,3905	10,8649	17,3379	25,9894	28,8693	32,3462	34,8053	42,3124
19	7,6327	10,1170	11,6509	18,3377	27,2036	30,1435	33,6874	36,1909	43,8202
20	8,2604	10,8508	12,4426	19,3374	28,4120	31,4140	35,0196	37,5662	45,3147
21	8,8972	11,5913	13,2396	20,3372	29,6151	32,6706	36,3434	38,9322	46,7990
22	9,4525	12,3380	14,0415	21,3370	30,8133	33,9244	37,6595	40,2894	48,2679
23	10,1957	13,0905	14,8480	22,3369	32,0069	35,1725	38,9693	41,6384	49,7282
24	10,8564	13,8484	15,6587	23,3367	33,1962	36,4150	40,2704	42,9798	51,1786
25	11,5240	14,6114	16,4734	24,3366	34,3816	37,6525	41,5661	44,3141	52,6197
26	12,1981	15,3792	17,2919	25,3365	35,5632	38,8851	42,8558	45,6417	54,5020
27	12,8785	16,1514	18,1139	26,3363	36,7412	40,1133	44,1400	46,9629	55,4760
28	13,5647	16,9279	18,9392	27,3362	37,9159	41,3371	45,4188	48,2782	56,8923
29	14,2565	17,7084	19,7677	28,3361	39,0875	42,5570	46,6927	49,5879	58,3012
30	14,9535	18,4927	20,5992	29,3360	40,2560	43,7730	47,9618	50,8922	59,7031

Phụ lục 5: Đề thi tham khảo

ĐỀ THI KẾT THÚC HỌC PHẦN (1)

Đề tham khảo. Thời gian làm bài: 90 phút.

Các kết quả tính toán gần đúng cần được làm tròn đến 3 chữ số thập phân.

Câu 1 (2.5 điểm). Một lô hàng có 20 sản phẩm trong đó có 5 sản phẩm loại A, 7 sản phẩm loại B, 8 sản phẩm loại C. Chọn ngẫu nhiên 3 sản phẩm từ lô hàng.

- a) Có 1 sản phẩm loại A, 2 sản phẩm loại B.
- b) Gọi X là số sản phẩm loại A nhận được. Lập bảng phân phối xác suất của X .

Câu 2 (2.0 điểm). Giả sử theo thống kê người ta biết được xác suất để một loại xe nào đó ngẫu nhiên bị cháy nổ trong một năm là 0.001. Bán bảo hiểm cháy nổ thời hạn một năm cho các xe loại này với giá 32000 đồng/xe. Nếu xe bị cháy nổ phù hợp với quy định bồi thường thì công ty bảo hiểm sẽ trả cho chủ xe 25000000 đồng/xe (bị cháy nổ). Gọi X là số tiền công ty thu được khi bán bảo hiểm cháy nổ (thời hạn một năm) cho một xe loại này.

Tính số tiền trung bình EX mà công ty thu được khi bán bảo hiểm cháy nổ (thời hạn một năm) cho một xe loại này.

Câu 3 (3.0 điểm). Chọn ngẫu nhiên 55 cây con giống của một loại cây một năm tuổi trong vườn ươm để đo chiều cao. Ta có mẫu số liệu về chiều cao X (đơn vị: m) như sau

$X (m)$	$[0 ; 0.2)$	$[0.2 ; 0.4)$	$[0.4 ; 0.6)$	$[0.6 ; 0.8)$	$[0.8 ; 1)$
Số cây con	7	9	20	12	7

- a) Tính các số đặc trưng mẫu: \bar{x} , s^2 , s (ghi đầy đủ công thức).
- b) Những cây con cao từ 0.4 m trở xuống được xem là cây con kém phát triển. Có nhận định cho rằng tỉ lệ các cây con kém phát triển trong vườn là lớn hơn 25%. Với mức ý nghĩa 10%, hãy cho biết nhận định trên có đúng hay không?

Câu 4 (2.5 điểm). Để kiểm tra trọng lượng (đơn vị: kg) của các bao gạo trong kho, người ta chọn

Column1	
Mean	49.93353125
Standard Error	0.058980331
Median	50
Mode	50
Standard Deviation	0.333643134
Sample Variance	0.111317741
Kurtosis	1.413488669
Skewness	-0.8584145
Range	1.54
Minimum	49.06
Maximum	50.6
Sum	1597.873
Count	32
Largest(1)	50.6
Smallest(1)	49.06

ngẫu nhiên một số bao gạo để cân. Xử lý thống kê mẫu số liệu bằng công cụ thống kê Data Analysis của phần mềm Microsoft Excel ta có bảng bên.

a) Hãy cho biết giá trị của cỡ mẫu n , trung bình mẫu \bar{x} , phương sai mẫu (hiệu chỉnh) s^2 và độ lệch chuẩn mẫu s (không cần ghi công thức). Từ đó, hãy tìm khoảng ước lượng về trọng lượng trung bình của các bao gạo trong kho với độ tin cậy 96%.

b) Hãy kiểm định thống kê xem trọng lượng trung bình của các bao gạo trong kho có thấp hơn 50 kg hay không? $\alpha = 5\%$.

Ghi chú: Sinh viên được sử dụng tài liệu khi làm bài.

ĐỀ THI KẾT THÚC HỌC PHẦN (2)

Đề tham khảo. Thời gian làm bài: 90 phút.

Các kết quả tính toán gần đúng được làm tròn đến 3 chữ số thập phân.

Câu 1 (2.0 điểm). Có 25 thùng hàng laptop đã qua sử dụng, trong đó có 15 thùng hàng hiệu Dell và 10 thùng hàng hiệu Asus. Mỗi thùng hàng đều có nhiều laptop đã qua sử dụng. Tỷ lệ máy tốt (còn sử dụng được) đối với các thùng hàng hiệu Dell là 80%, hiệu Asus là 87%. Chọn ngẫu nhiên một thùng hàng rồi chọn ngẫu nhiên trong đó ra một laptop để kiểm tra, tính xác suất gặp phải máy tính tốt.

Câu 2 (2.5 điểm). Tuổi thọ của một loại côn trùng nào đó là biến ngẫu nhiên X (đơn vị: tháng) có hàm mật độ xác suất

$$p(x) = \begin{cases} kx^2(6-x) & \text{nếu } 0 \leq x \leq 6 \\ 0 & \text{nếu trái lại.} \end{cases} \quad (C \text{ là hằng số})$$

Tìm hằng số k và tuổi thọ trung bình EX của loại côn trùng này.

Câu 3 (3.5 điểm). Chọn ngẫu nhiên 60 quả xoài của một loại xoài trong vườn. Ta có mẫu số liệu về trọng lượng X (đơn vị kg) như sau

X (kg)	0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8 - 1.0
Số quả	9	15	24	10	2

a) Với độ tin cậy 95% hãy tìm khoảng ước lượng về trọng lượng trung bình của quả xoài loại này trong vườn.

b) Những quả nặng từ 0.6 kg trở lên được xem là quả to. Có nhận định cho rằng tỉ lệ các quả to trong vườn là thấp hơn 25%. Với mức ý nghĩa 4%, hãy cho biết nhận định trên có đúng không?

Câu 4 (2.0 điểm). Để nghiên cứu về mối quan hệ giữa chu vi lồng ngực X (đơn vị: cm) và trọng lượng Y (đơn vị: kg) của một giống lợn, người ta chọn ngẫu nhiên 10 con lợn trong trang trại để đo. Bảng số liệu thu được như sau

X (cm)	61	76.2	91.5	83.8	94.4	66.6	101.6	104.7	119.4	114.3
Y (kg)	22.3	44.9	72.7	59.5	81.7	31.4	96	100.4	123.5	116

a) Tìm hệ số tương quan mẫu thực nghiệm r_{xy} .

b) Tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x . Hãy dự báo trọng lượng của một con lợn có chu vi lồng ngực là 82 cm.

Ghi chú: Sinh viên được sử dụng tài liệu khi làm bài.

Phụ lục 6: Minh họa sử dụng gói công cụ Data Analysis trong Excel

	A	B	C	D	E	F	G
4				Minh họa sử dụng chức năng Descriptive Statistics (Thống kê mô tả), gói công cụ Data Analysis trong Excel			
5		Số liệu cân nặng					
6		(kg)		<i>Column1</i>			<i>Giải thích ý nghĩa</i>
7		56					
8		52	Mean		53.3		Trung bình mẫu, lệnh Excel tính riêng là average (B7:B21)
9		56	Standard Error		1.005698052		Sai số chuẩn $e = s/\sqrt{n} = E12/\sqrt{E20}$
10		60	Median		52		Trung vị Q2, lệnh Excel median (B7:B21)
11		49.5	Mode		52		Mốt (Yếu vị), lệnh Excel mode (B7:B21)
12		52	Standard Deviation		3.895051806		Độ lệch chuẩn mẫu s , Excel stdev (B7:B21)
13		47	Sample Variance		15.17142857		Phương sai mẫu (hiệu chỉnh) s^2 , lệnh Excel var (B7:B21)
14		49	Kurtosis		-0.583466535		Độ nhọn, lệnh Excel kurt (B7:B21)
15		52	Skewness		0.266819661		Độ xiên, độ lệch, lệnh Excel skew (B7:B21)
16		52	Range		13		Khoảng biến thiên = $X_{\max} - X_{\min}$
17		49	Minimum		47		X_{\min} , lệnh min (B7:B21)
18		54.5	Maximum		60		X_{\max} , lệnh max (B7:B21)
19		56	Sum		799.5		Tổng = $X_1 + X_2 + \dots + X_n$, lệnh Excel sum (B7:B21)
20		60	Count		15		Cỡ mẫu n , lệnh Excel count (B1:B21)
21		54.5	Confidence Level(95.0%)		2.157007794		Sai số epsilon của khoảng ước lượng kỳ vọng (mean-epsilon; mean+epsilon), với độ tin cậy $P=1-\alpha=95\%$, $\epsilon = t(\alpha, n-1) * e = t(0.05, 15-1) * s / \sqrt{n}$. Lệnh Excel: T.INV.2T (0.05,14)*E9 = T.INV.2T (0.05,14)*E12/ sqrt (E20)

TÀI LIỆU THAM KHẢO

- [1] Lê Sỹ Đồng, *Xác suất thống kê và ứng dụng*, NXB Giáo dục, 2004.
- [2] Lê Sỹ Đồng, *Bài tập xác suất thống kê và ứng dụng*, NXB Giáo dục, 2010.
- [3] Phạm Văn Kiều, *Xác suất thống kê*, NXB Đại học Sư phạm, 2005.
- [4] Lê Trung Hiếu, Huỳnh Ngọc Cẩm, Võ Thị Lệ Hằng, *Bài giảng Xác suất thống kê*, Bài giảng lưu hành nội bộ Trường Đại học Đồng Tháp, 2023.
- [5] Đào Hữu Hồ, *Xác suất thống kê*, NXB ĐHQG Hà Nội, 2004.
- [6] Đào Hữu Hồ, *Hướng dẫn giải các bài toán xác suất thống kê*, NXB ĐHQG Hà Nội, 2009.
- [7] Đỗ Đức Thái, Nguyễn Tiến Dũng, *Nhập môn hiện đại xác suất và thống kê*, NXB Đại học sư phạm, 2010.

Một số tài liệu tham khảo thêm: Truy cập trực tiếp tại link <https://tinyurl.com/rbzy52jz>, hoặc quét mã QR sau đây để vào link.

