MỘT SỐ KÝ HIỆU

1. Ω: Không gian mẫu.
2. w: Biến cố sơ cấp.
3. $P(A)$: Xác suất biến cố A.
4. $\mu_X = E(X)$: Kỳ vọng (trung bình) của biến cố X.
5. $\overline{X}$: Trung bình mẫu của X.
6. $\sigma_X^2 = var(X) = D(X)$: Phương sai của biến cố X.
7. S_X^2: Phương sai ngẫu cóa hiệu chỉnh của X.
8. X: Biến ngẫu nhiên X.
9. $X \sim B(n;p)$: Biến ngẫu nhiên X có phân phối nhị thức.
10. $X \sim H(N,K,n)$: Biến ngẫu nhiên X có phân phối siêu bội.
11. $X \sim P(\mu)$: Biến ngẫu nhiên X có phân phối Poisson.
12. $X \sim U[a,b]$: Biến ngẫu nhiên X có phân phối đều.
13. $X \sim Exp(\lambda)$: Biến ngẫu nhiên X có phân phối mũ.
14. $X \sim N(0,1)$: Biến ngẫu nhiên X có phân phối chuẩn tắc.
15. $X \sim N(\mu,\sigma^2)$: Biến ngẫu nhiên X có phân phối chuẩn.
16. $X \sim \Gamma(\alpha,\beta)$: Biến ngẫu nhiên X có phân phối Gamma.
17. $X \sim \chi^2(r)$: Biến ngẫu nhiên X có phân phối Chi bình phương.
18. $X \sim St(n)$: Biến ngẫu nhiên X có phân phối Student.
19. $X \sim F(n,m)$: Biến ngẫu nhiên X có phân phối Fisher.
20. δ: Lượng tăng giảm tuyệt đối liên hoàn.
21. Δ: Lượng tăng giảm tuyệt đối định gốc.
22. Σ: Ký hiệu tổng.
23. Π: Ký hiệu tích.
24. H_0: Giả thuyết H_0.
25. H_1: Đối thuyết (nghịch thuyết) H_1.

Chương 1 **BIẾN CỐ NGẪU NHIÊN VÀ XÁC SUẤT**

Mục tiêu chương 1

Chương này giúp sinh viên:

- Phân biệt được sự kiện ngẫu nhiên (đối tượng môn xác suất) và sự kiện tất định (đối tượng của vật lý và hóa học). Nắm được các khái niệm về phép thử, không gian mẫu, biến cố và biến cố sơ cấp cũng như các biến cố đặc biệt.
- Hiểu được thế nào là xác suất và biết một số định nghĩa về xác suất.
- Biết và áp dụng được công thức xác suất đầy đủ và công thức xác suất Bayes.
- Biết áp dụng công thức Bernoulli để tính xác suất.

1.1. Phép thử và các loại biến cố

1.1.1. Sự kiện ngẫu nhiên và phép thử

Sự kiện ngẫu nhiên là những sự kiện dù được thực hiện trong cùng một điều kiện như nhau vẫn có thể cho nhiều kết quả khác nhau. Chẳng hạn, tung một con xúc xắc, ta không thể chắc chắn rằng mặt nào sẽ xuất hiện; lấy ra một sản phẩm từ một lô hàng gồm cả hàng chính phẩm lẫn phế phẩm, ta không chắc chắn sẽ nhận được hàng chính phẩm hay phế phẩm. Sự kiện ngẫu nhiên là đối tượng khảo sát của lý thuyết xác suất.

Mỗi lần cho xảy ra một sự kiện ngẫu nhiên được gọi là thực hiện một *phép thử*, còn sự kiện có thể xảy ra trong kết quả của phép thử đó gọi là *biến cố*. Khi đó, dù ta không thể dự đoán được kết quả nào sẽ xảy ra nhưng thường ta có thể liệt kê tất cả các kết quả có thể xảy ra. Tập hợp tất cả các kết quả có thể xảy ra của một phép thử được gọi là *không gian mẫu*. Ký hiệu Ω.

Ví dụ 1.1. Xét phép thử "tung một con xúc xắc". Ta có thể nhận được mặt 1 chấm, mặt 2 chấm, …, mặt 6 chấm. Vậy không gian mẫu có thể liệt kê và ký hiệu như sau:

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

1.1.2. Các loại biến cố

- Biến cố chắc chắn là biến cố luôn luôn sẽ xảy ra khi thực hiện một phép thử, biến cố chắc chắn thường ký hiệu là U.

- Biến cố không thể có là biến cố không bao giờ xảy ra khi thực hiện một phép thử, biến cố không thể có thường ký hiệu là V.

- Biến cố ngẫu nhiên là biến cố có thể xảy ra hoặc không xảy ra khi thực hiện phép thử. Ta có thể xem biến cố ngẫu nhiên là một tập con của không gian mẫu, các biến cố ngẫu nhiên thường ký hiệu là $A, B, C, ... \subset \Omega$.

- Biến cố sơ cấp là một kết quả (kết cục) của không gian mẫu, ký hiệu $w \in \Omega$. Do đó, không gian mẫu là tập hợp tất cả các biến cố sơ cấp.

Ví dụ 1.2. Thực hiện phép thử tung một con xúc xắc. Ta có:

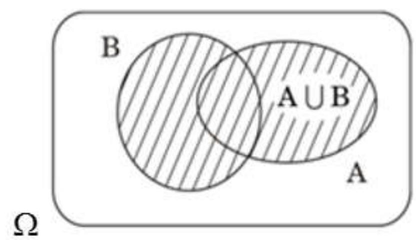
- Không gian mẫu: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Biến cố: "nhận được mặt có số chấm ≤ 6" là biến cố chắc chắn.
- Biến cố: "nhận được mặt có 7 chấm" là biến cố không thể có.
- Biến cố: "nhận được mặt có số chấm là chẵn" là biến cố ngẫu nhiên.
- Biến cố: "nhận được mặt 1 chấm" là biến cố sơ cấp.

1.1.3. Các phép toán giữa các biến cố

1.1.3.1. Tổng các biến cố

Cho hai biến cố bất kỳ $A, B \subset \Omega$, ta có thể thành lập biến cố:

$A \cup B \equiv A + B$ là chỉ biến cố "A xảy ra hay B xảy ra khi thực hiện phép thử".



Hình 1.1. Hình vẽ minh họa tổng hai biến cố.

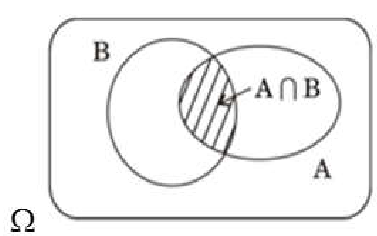
Tổng quát, cho $B_1, B_2, ..., B_n \subset \Omega$, ta có thể thành lập biến cố:

$\bigcup_{i=1}^{n} B_i \equiv \sum_{i=1}^{n} B_i$ là chỉ biến cố "có ít nhất một trong n biến cố đó xảy ra khi thực hiện phép thử".

1.1.3.2. Tích các biến cố

Cho hai biến cố bất kỳ $A, B \subset \Omega$, ta có thể thành lập biến cố:

$A \cap B \equiv A \cdot B$ là chỉ biến cố "A và B cùng xảy ra khi thực hiện phép thử".



Hình 1.2. Hình vẽ minh họa tích hai biến cố.

Tổng quát, cho $B_1, B_2, ..., B_n \subset \Omega$, ta có thể thành lập biến cố:

$\bigcap_{i=1}^{n} B_i \equiv \prod_{i=1}^{n} B_i$ là chỉ biến cố "cả n biến cố đó cùng xảy ra khi thực hiện phép thử".

Ví dụ 1.3. Khảo sát một lớp học về sự yêu thích môn xác suất thống kê và môn kinh tế học. Chọn ngẫu nhiên một sinh viên trong lớp này.

Gọi A là biến cố "nhận được sinh viên thích môn xác suất thống kê" và B là biến cố "nhận được sinh viên thích môn kinh tế học". Suy ra

Biến cố "sinh viên thích ít nhất một môn" là biến cố: $A + B$.

Biến cố "sinh viên thích cả hai môn" là biến cố: AB.

1.1.4. Quan hệ giữa các biến cố

1.1.4.1. Hai biến cố độc lập

Hai biến cố A và B được gọi là độc lập với nhau nếu việc xảy ra hay không xảy ra của biến cố A không phụ thuộc vào việc biến cố B xảy ra hay không xảy ra và ngược lại.

Nếu hai biến cố A và B không độc lập với nhau thì ta gọi là hai biến cố phụ thuộc.

Tổng quát,

- $B_1, B_2, ..., B_n$ là họ các biến cố độc lập với nhau từng đôi nếu hai biến cố bất kỳ trong n biến cố đó độc lập với nhau.
- $B_1, B_2, ..., B_n$ là họ các biến cố độc lập toàn phần nếu mỗi biến cố đó độc lập với một tổ hợp bất kỳ của các biến cố còn lại.

1.1.4.2. Hai biến cố xung khắc

Hai biến cố A và B gọi là xung khắc với nhau nếu chúng không thể đồng thời xảy ra trong cùng một phép thử.

A và B xung khắc khi và chỉ khi $A \cap B = \varnothing$.

Tổng quát, cho $B_1, B_2, ..., B_n$ là họ các biến cố xung khắc từng đôi một nếu bất kỳ 2 biến cố nào trong nhóm này cũng xung khắc với nhau, nghĩa là

$$B_i \cap B_j = \varnothing \text{ với } i \neq j \text{ và } i, j = \overline{1, n}.$$

Ví dụ 1.4. Trong một giỏ hàng có hai loại sản phẩm: Sản phẩm loại 1 và sản phẩm loại 2. Lấy ngẫu nhiên từ giỏ hàng đó ra một sản phẩm.

Gọi A là biến cố "nhận được sản phẩm loại 1".

Gọi B là biến cố "nhận được sản phẩm loại 2".

$\Rightarrow$ A và B là 2 biến cố xung khắc.

Ví dụ 1.5. Gieo đồng thời hai con xúc xắc.

Gọi C là biến cố “Con xúc xắc thứ nhất xuất hiện mặt 6 chấm”.

Gọi D là biến cố “Con xúc xắc thứ hai xuất hiện mặt 6 chấm”.

$\Rightarrow$ C và D không xung khắc.

1.1.4.3. Họ đầy đủ các biến cố

$B_1, B_2, ..., B_n$ là họ đầy đủ các biến cố nếu thỏa mãn hai điều kiện sau:

i) $B_1 \cup B_2 \cup ... \cup B_n = \Omega$ và

ii) $B_i \cap B_j = \varnothing,\ i \neq j$.

Ví dụ 1.6. Gieo một con xúc xắc.

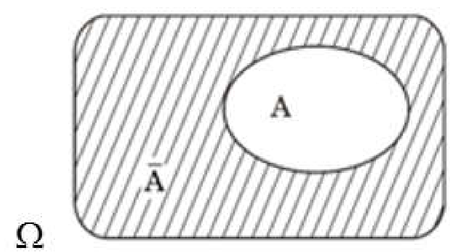
Gọi B_i là biến cố “nhận được mặt có i chấm”, $i = \overline{1,6}$.

Các biến cố $B_1, B_2, ..., B_6$ tạo nên một họ đầy đủ các biến cố.

1.1.4.4. Hai biến cố đối lập

Hai biến cố A và $\overline{A}$ gọi là hai biến cố đối lập với nhau nếu chúng tạo nên một họ đầy đủ các biến cố.

$$A \text{ và } \overline{A} \text{ là hai biến cố đối lập} \Leftrightarrow A \cup \overline{A} = \Omega \text{ và } A \cap \overline{A} = \varnothing.$$



Hình 1.3. Hình vẽ minh họa hai biến cố đối lập.

1.2. Xác suất của biến cố

1.2.1. Khái niệm chung về xác suất

Quan sát các biến cố đối với một phép thử, mặc dù không thể khẳng định một biến cố có xảy ra hay không nhưng người ta có thể phỏng chừng cơ may xảy ra của các biến cố này là ít hay nhiều. Chẳng hạn, với phép thử *"tung xúc xắc"*, biến cố *"nhận được mặt 1"* ít xảy ra hơn biến cố *"nhận được mặt chẵn"*. Do đó, người ta tìm cách định lượng khả năng xuất hiện khách quan của một biến cố mà ta sẽ gọi là *xác suất* của biến cố đó.

Xác suất của một biến cố là một con số đặc trưng cho khả năng xảy ra khách quan của biến cố đó.

Xác suất của biến cố A, ký hiệu là $P(A)$, có thể được định nghĩa bằng nhiều cách.

1.2.2. Định nghĩa cổ điển

Xét một phép thử τ với n kết quả có thể xảy ra, nghĩa là không gian mẫu Ω có n biến cố sơ cấp, và biến cố $A \subset \Omega$ có k phần tử. Nếu *các biến cố sơ cấp có cùng khả năng xảy ra* thì xác suất của A được định nghĩa là

$$P(A) = \frac{|A|}{|\Omega|} = \frac{k}{n}. \tag{1.1}$$

Trong đó $|A|, |\Omega|$ là số khả năng của biến cố A và số khả năng của Ω.

Ví dụ 1.7.

a) Xét phép thử "tung một con xúc xắc" với các biến cố

$A \equiv$ "nhận được mặt 6",

$B \equiv$ "nhận được mặt chẵn".

Theo công thức (1.1), ta có

$$P(A) = \frac{1}{6} \text{ và } P(B) = \frac{3}{6} = 0,5.$$

b) Xét phép thử "lấy ngẫu nhiên một sản phẩm trong một giỏ hàng đựng 4 sản phẩm loại 1 và 6 sản phẩm loại 2" với các biến cố

$C \equiv$ "nhận được sản phẩm loại 1",

$D \equiv$ "nhận được sản phẩm loại 2".

Theo công thức (1.1), ta có

$$P(C) = \frac{4}{10} = 0,4 \text{ và } P(D) = \frac{6}{10} = 0,6.$$

Lưu ý rằng, đối với định nghĩa cổ điển, ta cần hai điều kiện:

Số kết quả của phép thử là hữu hạn,

Các kết quả đồng khả năng xảy ra.

Khi một trong hai điều kiện trên không xảy ra, ta không thể dùng định nghĩa cổ điển để xác định xác suất của một biến cố. Ta có thể định nghĩa xác suất bằng phương pháp thống kê như sau.

1.2.3. Định nghĩa xác suất bằng tần suất

Giả sử phép thử τ có thể lập lại nhiều lần trong điều kiện giống nhau. Nếu trong n lần thực hiện phép thử mà biến cố A xảy ra k lần thì tỷ số $\frac{k}{n}$ được gọi là *tần suất* xảy ra của A trong n phép thử.

Người ta chứng minh được rằng, khi n đủ lớn, tần suất của biến cố A sẽ dao động xung quanh một giá trị cố định nào đó mà ta gọi là xác suất của A, ký hiệu $P(A)$. Ta có

$$P(A) = \lim_{n\to\infty} \frac{k}{n}$$

Trong thực tế, với n đủ lớn, người ta lấy tần suất của A làm giá trị gần đúng cho xác suất của biến cố A, nghĩa là

$$P(A) \approx \frac{k}{n}. \tag{1.2}$$

Ví dụ 1.8.

a) Thống kê trên 10000 người dân thành phố cho thấy có 51 người bị bệnh cao huyết áp, theo công thức (1.2), ta nói xác suất của biến cố "bị bệnh cao huyết áp" là

$$\frac{51}{10000} \approx 0,005.$$

b) Một nhà máy gồm ba phân xưởng A, B, C. Kiểm tra một lô hàng của nhà máy gồm 1000 sản phẩm, người ta thấy có 252 sản phẩm của phân xưởng A, 349 của phân xưởng B và 399 của phân xưởng C. Theo công thức (1.2), ta nói xác suất

nhận được sản phẩm từ phân xưởng A là $P(A) = \dfrac{252}{10000} \approx 0,25,$

nhận được sản phẩm từ phân xưởng B là $P(B) = \dfrac{349}{10000} \approx 0,35,$ và

nhận được sản phẩm từ phân xưởng C là $P(C) = \dfrac{399}{10000} \approx 0,4.$

Ta còn nói, các phân xưởng A, B, C tương ứng làm ra 25%, 35% và 40% tổng sản lượng nhà máy.

Tương tự, để tìm xác suất làm ra sản phẩm hỏng của phân xưởng A, người ta thống kê trên một số sản phẩm của phân xưởng A và quan sát số sản phẩm hỏng. Chẳng hạn, nếu trong 400 sản phẩm của phân xưởng A nêu trên có 4 sản phẩm hỏng, theo công thức (1.2), ta nói xác suất làm ra một sản phẩm hỏng của phân xưởng A là $\dfrac{4}{400} = 0,01.$

Ví dụ 1.9. Xét phép thử τ : "tung đồng xu", một cách trực giác, ta cho rằng các biến cố sơ cấp w_1: "nhận được mặt sấp" và w_2: "nhận được mặt ngửa" là đồng khả năng xảy ra, nên do định nghĩa cổ điển, $P(w_1) = P(w_2) = 0,5.$ Khi đó, người ta nói đồng xu này là "công bằng", "đồng chất đẳng hướng", ... Bằng thực nghiệm, một số nhà khoa học đã tung một đồng xu nhiều lần và nhận được kết quả sau:

Người thực hiện	Số lần thảy	Số lần mặt ngửa	Tần suất
Buffon	4040	2048	0,5069
Pearson	12000	6019	0,5016
Pearson	24000	12012	0,5005

và khi đó, ta nói xác suất nhận được mặt ngửa $\approx 0,5$.

1.2.4. Định nghĩa hình học về xác suất

Định nghĩa hình học về xác suất có thể sử dụng khi xác suất để một điểm ngẫu nhiên rơi vào một phần nào đó của một miền cho trước tỷ lệ với độ đo của miền đó (độ dài, diện tích, thể tích…) và không phụ thuộc vào dạng thức của miền đó.

Nếu độ đo hình học của toàn bộ miền cho trước là S, còn độ đo hình học của một phần A nào đó của nó là S_A thì xác suất để điểm ngẫu nhiên rơi vào A sẽ bằng:

$$P = \frac{S_A}{S} \tag{1.3}$$

trong đó: $0 < S, S_A < +\infty$.

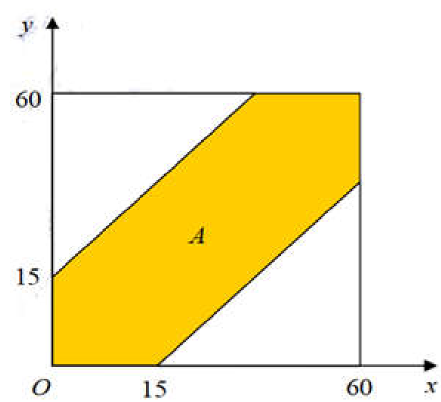
Ví dụ 1.10. Giả sử hai người X và Y hẹn gặp nhau trong khoảng thời gian là 60 phút, với điều kiện người tới trước sẽ đợi người tới sau tối đa 15 phút, sau đó đi khỏi. Tính xác suất để X và Y gặp nhau.

Giải. Gọi x là thời gian đến của X, y là thời gian đến của Y. Khi đó không gian các biến cố sơ cấp sinh ra khi X và Y tới gặp nhau có dạng:

$$\Omega = \left\{ (x, y) \in \mathbb{R}^2 : 0 \le x \le 60, 0 \le y \le 60 \right\}$$

Gọi A là biến cố hai người gặp nhau, khi đó

$$A = \left\{ (x, y) \in \mathbb{R}^2 : |x - y| \le 15 \right\}$$



Hình 1.4. Hình vẽ minh họa biến cố hai người gặp nhau.

Theo công thức xác suất hình học (1.3), ta có

$$P(C)=\frac{60\cdot 60-45\cdot 45}{60\cdot 60}=\frac{7}{16}.$$

1.2.5. Định nghĩa tiên đề về xác suất

Vào những năm 30 của thế kỷ 20, nhà Toán học người Nga là Kolmogorov đã xây dựng hệ tiên đề làm cơ sở cho việc định nghĩa một cách hoàn chỉnh khái niệm xác suất về mặt lý thuyết. Hệ tiên đề được xây dựng trên cơ sở khái niệm về không gian biến cố sơ cấp $w_1, w_2, ..., w_n$, thực tế là tập hợp tất cả các trường hợp có thể xảy ra của một phép thử. Lúc đó mỗi biến cố A có thể quan niệm như một tập hợp của không gian đó.

Tiên đề 1. Với mọi biến cố A đều có $0 \le P(A) \le 1$.

Tiên đề 2. Nếu $w_1, w_2, ..., w_n$ tạo nên không gian các biến cố sơ cấp thì:

$$P(w_1)+P(w_2)+\cdots+P(w_n)=1.$$

Tiên đề 3. Nếu các biến cố $A_1, A_2, ..., A_n, ...$ là các tập hợp con không giao nhau của các biến cố sơ cấp thì:

$$P\left(\sum_{i=1}^{\infty} A_i\right)=\sum_{i=1}^{\infty} P(A_i).$$

Tiên đề 4. Với A, B là hai biến cố bất kỳ, ta có

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

Tiên đề 5. Với hai biến cố A và $\overline{A}$, ta có

$$P(\overline{A})=1-P(A).$$

1.2.6. Nguyên lý xác suất nhỏ và xác suất lớn

Trong nhiều bài toán thực tế, ta thường gặp các biến cố có xác suất rất nhỏ, gần bằng 0. Qua nhiều lần quan sát, người ta thấy rằng: các biến cố có xác suất nhỏ gần như không xảy ra khi thực hiện phép thử. Trên cơ sở đó có thể đưa ra "*Nguyên lý không thực tế không thể có của các biến cố có xác suất nhỏ*" sau đây: *Nếu một biến cố có xác suất rất nhỏ thì thực tế có thể cho rằng trong một phép thử, biến cố đó sẽ không xảy ra.*

Việc quy định một mức xác suất được coi là "rất nhỏ" tuỳ thuộc vào từng bài toán cụ thể. Chẳng hạn: Nếu xác suất để một loại dù không mở khi nhảy dù là 0,01 thì mức xác suất này chưa thể coi là nhỏ và ta không nên sử dụng loại dù đó. Song nếu xác suất để một chuyến tàu đến ga chậm 10 phút là 0,01 thì ta có thể coi mức xác suất đó là nhỏ, tức là có thể cho rằng xe lửa đến ga đúng giờ.

Một mức xác suất nhỏ mà với nó ta có thể cho rằng: biến cố đang xét không xảy ra trong một phép thử được gọi là *mức ý nghĩa*. Tuỳ theo từng bài toán cụ thể, mức ý nghĩa thường được lấy trong khoảng từ 0,01 đến 0,05.

Tương tự như vậy ta có thể nêu ra “Nguyên lý thực tế chắc chắn xảy ra của các biến cố có xác suất lớn” như sau: Nếu một biến cố có xác suất gần bằng 1 thì thực tế có thể cho rằng biến cố đó sẽ xảy ra trong một phép thử.

Cũng như trên, việc quy định mức xác suất được coi là lớn hay nhỏ tuỳ thuộc vào bài toán cụ thể. Thông thường người ta lấy trong khoảng từ 0,95 đến 0,99.

1.3. Xác suất có điều kiện

Xét ví dụ sau: “Tung hai con xúc xắc” với không gian mẫu là

$$\Omega=\{(1,1),(1,2),...,(1,6),(2,1),(2,2),...,(5,6),(6,6)\}$$

(tổng cộng có 36 khả năng (phần tử)) và xét các biến cố

A: “tổng số nút xuất hiện cộng lại bằng 6”,

B: “số nút của xúc xắc thứ nhất là số lẻ”.

Ta có:

$$A=\{(1,5),(2,4),(3,3),(4,2),(5,1)\},$$

$$B=\{(1,1),...,(1,6),(3,1),...,(3,6),(5,1),...,(5,6)\},$$

nên từ định nghĩa cổ điển,

$$P(A)=\frac{|A|}{|\Omega|}=\frac{5}{36} \text{ và } P(B)=\frac{|B|}{|\Omega|}=\frac{18}{36}=0,5.$$

Bây giờ, ta tung hai con xúc xắc và giả sử ta nhận được thông tin thêm là số nút của xúc xắc thứ nhất đã là số lẻ (nghĩa là biến cố B đã xảy ra). Khi đó, phép thử trên trở thành phép thử: “tung hai con xúc xắc khác nhau với số nút của xúc xắc thứ nhất là số lẻ”. Do đó, không gian mẫu Ω bị thu hẹp lại là

$$\Omega^{/}=\{(1,1),...,(1,6),(3,1),...,(3,6),(5,1),...,(5,6)\}$$

và hiện tượng biến cố A xảy ra khi biết biến cố B đã xảy ra trở thành hiện tượng biến cố

$$A^{/}=\{(1,5),(3,3),(5,1)\}=AB$$

xảy ra đối với phép thử τ' và do đó có xác suất là

$$P(A^{/})=\frac{|A^{/}|}{|\Omega^{/}|}=\frac{3}{18}=\frac{1}{6}.$$

Ta ký hiệu $A^{/} = A|B$ và $P(A^{/}) = P(A|B)$ được gọi là xác suất để biến cố A xảy ra khi biết biến cố B xảy ra. Từ nhận xét

$$P(A|B) = \frac{1}{6} = \frac{P(AB)}{P(B)}.$$

1.3.1. Định nghĩa

Xét biến cố B với $P(B) > 0$. Xác suất của biến cố A, khi biết biến cố B xảy ra là

$$P(A|B) = \frac{P(AB)}{P(B)}. \tag{1.4}$$

Ví dụ 1.11. Trong một bình có 5 quả cầu trắng và 3 quả cầu đen. Lấy ngẫu nhiên lần lượt 2 quả cầu theo phương thức không hoàn lại.

Giải

Gọi A_i là biến cố "nhận được quả cầu trắng lần thứ i", i = 1, 2.

Theo định nghĩa xác suất cổ điển, xác suất để lần thứ nhất lấy được cầu trắng là:

$$P(A_1) = \frac{5}{8}$$

Nếu lần thứ nhất lấy được quả cầu trắng (tức là biến cố A_1 đã xảy ra) thì trong bình còn lại 7 quả cầu, trong đó có 4 quả cầu trắng. Vậy xác suất để lần thứ hai lấy được cầu trắng với điều kiện lần thứ nhất đã lấy được cầu trắng là:

$$P(A_2|A_1) = \frac{4}{7}.$$

Nếu lần thứ nhất lấy được quả cầu đen (tức là biến cố $\overline{A_1}$ đã xảy ra) thì trong bình còn lại 7 quả cầu, trong đó có 5 quả cầu trắng. Vậy xác suất để lần thứ hai lấy được cầu trắng với điều kiện lần thứ nhất đã lấy được cầu đen là:

$$P(A_2|\overline{A}_1) = \frac{5}{7}.$$

1.3.2. Công thức nhân xác suất

Với hai biến cố A và B bất kỳ, ta có

$$P(AB) = P(A)P(B|A). \tag{1.5}$$

Tổng quát, với n biến cố bất kỳ $A_1, A_2, ..., A_n$, ta có

$$P(A_1A_2...A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2...A_{n-1}). \tag{1.6}$$

Ví dụ 1.12. Một thủ quỹ có một chùm chìa khóa gồm 9 chiếc chìa giống hệt nhau trong đó chỉ có 2 chìa có thể mở được tủ sắt. Anh ta thử ngẫu nhiên từng chìa (chìa không trúng được bỏ ra trong lần thử kế tiếp). Tìm xác suất để anh ta mở được tủ vào đúng lần thứ ba.

Giải

Đặt A_i là biến cố "lần thứ i, mở được tủ". Với quy ước rằng khi biến cố A_i xảy ra thì các biến cố $A_1, A_2, ..., A_{i-1}$ vẫn có thể đã xảy ra, biến cố "mở được tủ vào đúng lần thứ ba" là $\overline{A}_1\overline{A}_2A_3$ và do quy tắc nhân xác suất, ta có

$$P\left(\overline{A}_1\overline{A}_2A_3\right) = P\left(\overline{A}_1\right)P\left(\overline{A}_2\middle|\overline{A}_1\right)P\left(A_3\middle|\overline{A}_1\overline{A}_2\right).$$

Do

$$P\left(\overline{A}_1\right) = 1 - P\left(A_1\right) = 1 - \frac{2}{9} = \frac{7}{9},$$

$$P\left(\overline{A}_2\middle|\overline{A}_1\right) = 1 - P\left(A_2\middle|\overline{A}_1\right) = 1 - \frac{2}{8} = \frac{6}{8},$$

$$P\left(A_3\middle|\overline{A}_1\overline{A}_2\right) = \frac{2}{7},$$

ta suy ra:

$$P\left(\overline{A}_1\overline{A}_2A_3\right) = \frac{7}{9}\cdot\frac{6}{8}\cdot\frac{2}{7} = \frac{1}{6}.$$

1.3.3. Công thức xác suất đầy đủ (công thức xác suất toàn phần)

Với hai biến cố A, B bất kỳ, ta có

$$P\left(A\right) = P\left(B\right)P\left(A\middle|B\right) + P\left(\overline{B}\right)P\left(A\middle|\overline{B}\right). \tag{1.7}$$

Tổng quát, cho $B_1, B_2, ..., B_n$ là họ đầy đủ các biến cố và với mọi biến cố A, ta có

$$P\left(A\right) = P\left(B_1\right)P\left(A\middle|B_1\right) + P\left(B_2\right)P\left(A\middle|B_2\right) + \cdots + P\left(B_n\right)P\left(A\middle|B_n\right). \tag{1.8}$$

hay

$$P\left(A\right) = \sum_{i=1}^{n} P\left(B_i\right)P\left(A\middle|B_i\right). \tag{1.9}$$

Chứng minh

Do BA và $\overline{B}A$ là hai biến cố xung khắc và $A = BA + \overline{B}A$ nên

$$\begin{aligned} P\left(A\right) &= P\left(BA + \overline{B}A\right) = P\left(BA\right) + P\left(\overline{B}A\right) \\ &= P\left(B\right)P\left(A\middle|B\right) + P\left(\overline{B}\right)P\left(A\middle|\overline{B}\right). \end{aligned}$$

Tổng quát, do các biến cố $B_1A, B_2A, \ldots, B_nA$ xung khắc từng đôi một và $A = B_1A + B_2A + \cdots B_nA$ nên do công thức cộng xác suất:

$$\begin{aligned} P(A) &= P(B_1A + B_2A + \cdots B_nA) = P(B_1A) + P(B_2A) + \cdots P(B_nA) \\ &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_n)P(A|B_n) \\ &= \sum_{i=1}^{n} P(B_i)P(A|B_i). \end{aligned}$$

và do công thức nhân xác suất,

$$P(B_iA) = P(B_i)P(A|B_i)$$

với mọi i, ta suy ra

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

Ví dụ 1.13. Có 2 xạ thủ loại I và 8 xạ thủ loại II, xác suất bắn trúng đích của các loại xạ thủ loại I là 0,9 và loại II là 0,7.

a) Chọn ngẫu nhiên ra một xạ thủ và xạ thủ đó bắn một viên đạn. Tìm xác suất để viên đạn đó trúng đích.

b) Chọn ngẫu nhiên ra hai xạ thủ và mỗi người bắn một viên đạn. Tìm xác suất để cả hai viên đạn đó trúng đích.

Giải

a) Gọi A là biến cố "Viên đạn trúng đích".

B_1 là biến cố "Chọn xạ thủ loại I bắn".

B_2 là biến cố "Chọn xạ thủ loại II bắn".

$$P(B_1) = \frac{2}{10} = 0,2, \quad P(A|B_1) = 0,9$$

$$P(B_2) = \frac{8}{10} = 0,8, \quad P(A|B_2) = 0,7$$

Ta có B_1, B_2 tạo thành họ đầy đủ các biến cố. Áp dụng công thức (1.8), ta có:

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) = 0,2 \cdot 0,9 + 0,8 \cdot 0,7 = 0,74.$$

b) Gọi B là biến cố "Cả 2 viên đạn trúng đích".

$B_i, (i = 1,2)$ là biến cố "Chọn được i xạ thủ loại I ".

$$P(B_0) = \frac{C_8^2}{C_{10}^2} = \frac{28}{45}; \; P(B|B_0) = 0,7.0,7 = 0,49$$

$$P\left(B_1\right)=\frac{C_2^1.C_8^1}{C_{10}^2}=\frac{16}{45};\ P\left(B\middle|B_1\right)=0,9.0,7=0,63$$

$$P\left(B_2\right)=\frac{C_2^2}{C_{10}^2}=\frac{1}{45};\ P\left(B\middle|B_2\right)=0,9.0,9=0,81.$$

Ta có B_1, B_2, B_3 tạo thành họ đầy đủ các biến cố. Áp dụng công thức(1.9), ta có

$$\begin{aligned}P(B) &= P\left(B_0\right)\cdot P\left(B\middle|B_0\right)+P\left(B_1\right)\cdot P\left(B\middle|B_1\right)+P\left(B_2\right)\cdot P\left(B\middle|B_2\right)\\ &=\frac{28}{45}\cdot 0,49+\frac{16}{45}\cdot 0,63+\frac{1}{45}0,81=0,5469.\end{aligned}$$

1.3.4. Công thức Bayes

Cho $B_1,B_2,...,B_n$ là họ đầy đủ các biến cố và xét biến cố A với $P(A)>0$. Với mỗi $k=1,2,...,n,$ ta có

$$P\left(B_k\middle|A\right)=\frac{P\left(B_k\right)\cdot P\left(A\middle|B_k\right)}{\sum_{i=1}^{n}P\left(B_i\right)P\left(A\middle|B_i\right)}. \tag{1.10}$$

Chứng minh

Áp dụng công thức nhân xác suất

$$P(A)P\left(B_k\middle|A\right)=P\left(AB_k\right)=P\left(B_kA\right)=P\left(B_k\right)P\left(A\middle|B_k\right)$$

và công thức xác suất toàn phần

$$P(A)=\sum_{i=1}^{n}P\left(B_i\right)P\left(A\middle|B_i\right),$$

ta suy ra

$$P\left(B_k\middle|A\right)=\frac{P\left(B_k\right)\cdot P\left(A\middle|B_k\right)}{P(A)}=\frac{P\left(B_k\right)\cdot P\left(A\middle|B_k\right)}{\sum_{i=1}^{n}P\left(B_i\right)P\left(A\middle|B_i\right)}.$$

Ví dụ 1.14. Tỷ lệ chính phẩm của máy thứ nhất là 99%, của máy thứ hai là 98%. Một lô sản phẩm gồm 40% sản phẩm của máy thứ nhất và 60% sản phẩm của máy thứ hai. Người ta lấy ngẫu nhiên ra một sản phẩm để kiểm tra thấy là sản phẩm tốt. Tìm xác suất để sản phẩm đó do máy thứ nhất sản xuất.

Giải

Gọi A là biến cố "Sản phẩm kiểm tra là sản phẩm tốt"

B_1 là biến cố "Sản phẩm do máy thứ nhất sản xuất".

B_2 là biến cố “Sản phẩm do máy thứ hai sản xuất”.

$$P(B_1) = 40\% = 0,4;\ P(B_2) = 60\% = 0,6$$

$$P(A|B_1) = 99\% = 0,99;\ P(A|B_2) = 98\% = 0,98$$

Do B_1, B_2 là họ đầy đầy đủ các biến cố. Áp dụng công thức (1.10), ta có

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1)+P(B_2)P(A|B_2)}$$
$$= \frac{0,4\cdot 0,99}{0,4\cdot 0,99 + 0,6\cdot 0,98} = 0,4.$$

1.3.5. Sự độc lập của các biến cố

Hai biến cố A, B được gọi là *độc lập* nếu xác suất để biến cố này xảy ra không phụ thuộc vào việc biến cố kia xảy ra, nghĩa là

$$P(A|B) = P(A)$$

và do đó

$$P(AB) = P(A)P(B). \qquad (1.11)$$

Tổng quát, n biến cố $A_1, A_2, ..., A_n$ được gọi là *độc lập* nếu mỗi biến cố A_i, với $i = 1, 2, ..., n$, độc lập với tích bất kỳ các biến cố còn lại.

Do định nghĩa, nếu ba biến cố A, B, C là độc lập thì A độc lập với B, C và BC nên

$$P(AB) = P(A)P(B),$$

$$P(AC) = P(A)P(C),$$

$$P[A(BC)] = P(A)P(BC),$$

và vì B, C cũng độc lập với nhau, nên

$$P(BC) = P(B)P(C),$$

và do đó

$$P(ABC) = P(A)P(B)P(C). \qquad (1.12)$$

Chú ý. Nếu A và B là biến cố độc lập thì $\overline{A}$ và B; A và $\overline{B}$; $\overline{A}$ và $\overline{B}$ cũng độc lập.

Ví dụ 1.15. Trong một bình có 5 quả cầu trắng và 3 quả cầu đen. Lấy ngẫu nhiên lần lượt 2 quả cầu. Tính xác suất để lấy được 2 quả cầu trắng trong hai trường hợp sau:

a) Lấy hoàn lại.

b) Lấy không hoàn lại.

Giải.

Gọi A là biến cố "Lấy được 2 quả cầu trắng".

A_i là biến cố "Lần thứ i lấy được cầu trắng", i = 1, 2.

Suy ra biến cố lấy được hai của cầu trắng là: $A_1.A_2$

$$P(A) = P(A_1.A_2)$$

a) Nếu lấy 2 quả cầu theo phương thức lần lượt có hoàn lại thì hai biến cố A_1 và A_2 là độc lập với nhau. Theo công thức (1.5), ta có

$$P(A) = P(A_1.A_2) = P(A_1)\cdot P(A_2) = \frac{5}{8}\cdot\frac{5}{8} = \frac{25}{64}.$$

b) Nếu lấy 2 quả cầu theo phương thức lần lượt không hoàn lại thì hai biến cố A_1 và A_2 là phụ thuộc với nhau. Theo công thức (1.5), ta có

$$P(A) = P(A_1.A_2) = P(A_1)\cdot P(A_2|A_1) = \frac{5}{8}\cdot\frac{4}{7} = \frac{5}{14}.$$

Ví dụ 1.16. Tung một đồng xu 3 lần. Tìm xác suất để 3 lần đều được mặt sấp.

Gọi A_i, $(i = 1,2,3)$ là biến cố "nhận được mặt sấp lần tung thứ i",

Ta có

$$P(A_i) = \frac{1}{2}$$

A là biến cố "Tung 3 lần đều được mặt sấp".

$$A = A_1A_2A_3$$

Các biến cố A_1, A_2, A_3 là độc lập toàn phần. Theo công thức (1.12), ta có

$$P(A) = P(A_1A_2A_3) = P(A_1)\cdot P(A_2)\cdot P(A_3) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}.$$

1.4. Công thức Bernoulli

Trong nhiều bài toán thực tế ta thường gặp trường hợp cùng một phép thử được lặp lại nhiều lần. Trong kết quả của mỗi phép thử có thể xảy ra hoặc không xảy ra một biến cố A nào đó và ta không quan tâm đến kết quả của từng phép thử mà quan tâm đến tổng số lần xảy ra của biến cố A trong cả dãy phép thử đó. Chẳng hạn, nếu tiến hành sản xuất hàng loạt một loại chi tiết nào đó thì ta thường quan tâm đến tổng số chi tiết đạt chuẩn của cả quá trình sản xuất. Trong những bài toán như vậy cần phải biết cách xác định xác suất để biến cố A xảy ra một số lần nhất định trong kết quả của cả một dãy phép thử. Bài toán này sẽ được giải quyết khá dễ dàng nếu các phép thử là độc lập với nhau.

Các phép thử được gọi là độc lập với nhau nếu xác suất xảy ra một biến cố nào đó trong từng phép thử sẽ không phụ thuộc vào biến cố đó có xảy ra ở các phép thử khác hay không. Chẳng hạn, tung nhiều lần một đồng xu sẽ tạo nên các phép thử độc lập, lấy nhiều lần sản phẩm từ một lô sản phẩm theo phương thức hoàn lại cũng tạo nên các phép thử độc lập v.v…

Giả sử ta thực hiện n phép thử độc lập. Trong mỗi phép thử chỉ có hai trường hợp: hoặc biến cố A xảy ra, hoặc biến cố A không xảy ra. Xác suất xảy ra của biến cố A trong mỗi phép thử đều bằng p và xác suất không xảy ra của biến cố A trong mỗi phép thử đều bằng $q = 1 - p$. Những bài toán thỏa mãn cả ba giả thiết được gọi là tuân theo lược đồ Bernoulli. Khi đó xác suất để trong n phép thử độc lập biến cố A xuất hiện đúng k lần, ký hiệu $P_n(k)$, được tính bằng công thức Bernoulli:

$$P_n(k) = C_n^k p^k q^{n-k}, \quad k = 0, 1, 2, ..., n. \tag{1.13}$$

Đặt H_k : “biến cố A xảy ra đúng k lần”, với $0 \le k \le n.$ Ta có

$$P(H_k) = P_n(k) = C_n^k p^k (1-p)^{n-k}.$$

Chứng minh

Dùng quy nạp trên n. Hiển nhiên công thức đúng với $n = 1$ vì khi đó $H_0 = \overline{A}$ và $H_1 = A.$ Do đó

$$P(H_0) = C_1^0 p^0 (1-p)^{1-0} = 1 - p$$

và

$$P(H_1) = C_1^1 p^1 (1-p)^{1-1} = p.$$

Giả sử công thức đúng với $n \ge 1,$ nghĩa là khi thực hiện n lần phép thử τ một cách độc lập thì xác suất để biến cố A xảy ra đúng k lần là

$$P(H_k) = C_n^k p^k (1-p)^{n-k}.$$

Bây giờ, thực hiện phép thử τ thêm một lần nữa một cách độc lập và gọi X là biến cố: *“A xảy ra trong lần thử thứ* $n+1$” thì biến cố: *“A xảy ra đúng k lần trong* $n+1$ *phép thử*” là

$$H_k \overline{A} + H_{k-1} A.$$

Do các biến cố $H_k \overline{A}$ và $H_{k-1} A$ là xung khắc, H_k và $\overline{A}$ cũng như H_{k-1} và A là các biến cố độc lập nên

$$\begin{aligned}
P\left(H_k\overline{A}+H_{k-1}A\right) &= P\left(H_k\overline{A}\right)+P\left(H_{k-1}A\right)\\
&= P\left(H_k\right)P\left(\overline{A}\right)+P\left(H_{k-1}\right)P\left(A\right)\\
&= C_n^k p^k\left(1-p\right)^{n-k}\left(1-p\right)+C_n^{k-1}p^{k-1}\left(1-p\right)^{n-k+1}p\\
&= C_n^k p^k\left(1-p\right)^{n-k+1}+C_n^{k-1}p^k\left(1-p\right)^{n-k+1}\\
&= \left(C_n^k+C_n^{k-1}\right)p^k\left(1-p\right)^{n-k+1}\\
&= C_{n+1}^k p^k\left(1-p\right)^{(n+1)-k}.
\end{aligned}$$

Ví dụ 1.17. Xác suất chữa khỏi bệnh A của một phương pháp điều trị là 95%. Với 10 người bị bệnh A được điều trị bằng phương pháp này, tính xác suất để

a) có 8 người khỏi bệnh.

b) có nhiều nhất 9 người khỏi bệnh.

Giải

Do việc khỏi bệnh của người này và người khác là độc lập nhau nên số người khỏi bệnh trong 10 người điều trị thỏa lược đồ Bernoulli với $n=10$ và $p=0,95$. Theo công thức (1.13). Ta có

$$P\left(H_k\right)=C_{10}^k 0,05^k\left(0,95\right)^{10-k}$$

a) Xác suất để có 8 người khỏi bệnh là $P\left(H_8\right)=C_{10}^8\left(0,05\right)^8\left(0,95\right)^{10-8}=0,0746$.

b) Biến cố: *"có nhiều nhất 9 người khỏi bệnh"* là biến cố đối của biến cố : *"có 10 người khỏi bệnh"* nên có xác suất là

$$P\left(H_{k\le 9}\right)=1-C_{10}^{10}\left(0,05\right)^{10}\left(0,95\right)^{10-10}=0,4013.$$

1.5. Tóm tắt chương 1

1. Xác suất của biết cố A:

$$P\left(A\right)=\frac{|A|}{|\Omega|}$$

($|A|$ và $|\Omega|$ lần lượt là số khả năng thuận lợi cho A và Ω).

2. Tính chất:

i) $0\le P\left(A\right)\le 1$.

ii) $P\left(\overline{A}\right)=1-P\left(A\right)$.

3. Công thức cộng:

i) Nếu $A_1, A_2, ..., A_n$ xung khắc với nhau từng đôi một thì

$$P(A_1 + A_2 + \cdots + A_2) = P(A_1) + P(A_2) + \cdots + P(A_2).$$

ii) Với A và B là hai biết cố bất kỳ

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

4. Công thức xác suất có điều kiện:

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

5. **Công thức nhân:**

i) Nếu $A_1, A_2, ..., A_n$ bất kỳ thì

$$P(A_1A_2...A_n) = P(A_1)P(A_2|A_1)\cdots P(A_n|A_1A_2....A_{n-1}).$$

ii) Nếu $A_1, A_2, ..., A_n$ độc lập với nhau từng đôi một thì

$$P(A_1A_2\cdots A_2) = P(A_1)P(A_2)\cdots P(A_n).$$

6. Công thức đầy đủ (toàn phần) và công thức Bayes:

Với $B_1, B_2, ..., B_n$ là họ đầy đủ các biến cố và với mọi biến cố A, ta có

i) Công thức đầy đủ

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

ii) Công thức Bayes

$$P(B_k|A) = \frac{P(B_k)\cdot P(A|B_k)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}, \quad k = 1, 2, ..., n.$$

7. Công thức Bernoulli:

Đặt H_k : "biến cố A xảy ra đúng k lần", với $0 \le k \le n, 0 < p < 1.$ Ta có

$$P(H_k) = P_n(k) = C_n^k p^k (1-p)^{n-k}.$$

1.6. Bài tập

Biểu diễn các biến cố

Bài số 1. Kiểm tra 3 sản phẩm. Gọi A_k là biến cố sản phẩm thứ k tốt. Hãy trình bày các cách biểu diễn qua A_k và qua giản đồ Venn các biến cố sau đây:

A: tất cả đều xấu,

B: có ít nhất một sản phẩm xấu,

C: có ít nhất một sản phẩm tốt,

D: không phải tất cả sản phẩm đều tốt,

E: có đúng một sản phẩm xấu,

F: có ít nhất 2 sản phẩm tốt.

Bài số 2. Ba người, mỗi người bắn một phát. Gọi A_i là biến cố thứ i bắn trúng. Hãy biểu diễn qua A_i các biến cố sau :

A: chỉ có người thứ nhất bắn trúng,

B: người thứ nhất bắn trúng và người thứ hai bắn trật,

C: cả 3 người đều bắn trúng,

D: có ít nhất 2 người bắn trúng,

E: chỉ có 2 người bắn trúng,

F: không ai bắn trúng,

G: không có hơn 2 người bắn trúng,

H: người thứ nhất bắn trúng, hoặc người thứ hai và người thứ ba cùng bắn trúng,

I: người thứ nhất bắn trúng hay người thứ hai bắn trúng,

K: có ít nhất 1 người bắn trúng.

Bài số 3. Ba sinh viên A, B, C cùng thi môn xác suất thống kê. Xét các biến cố:

A: sinh viên A đậu,

B: sinh viên B đậu,

C: sinh viên C đậu.

Hãy biểu diễn qua A, B, C các biến cố sau:

a) chỉ có A đậu,

b) A đậu và B rớt,

c) có ít nhất một người đậu,

d) cả 3 cùng đậu,

e) có ít nhất 2 người đậu,

f) chỉ có 2 người đậu,

g) không ai đậu,

h) không có quá 2 người đậu.

Bài số 4. Quan sát 4 sinh viên làm bài thi. Kí hiệu $B_j (j = 1, 2, 3, 4)$ là biến cố sinh viên j làm bài thi đạt yêu cầu. Hãy viết các biến cố sau đây

a) Có đúng một sinh viên đạt yêu cầu,

b) có đúng 3 sinh viên đạt yêu cầu,

c) có ít nhất 1 sinh viên đạt yêu cầu,

d) không có sinh viên nào đạt yêu cầu.

Định nghĩa xác suất, xác suất có điều kiện, công thức cộng, công thức nhân

Bài số 5. Thống kê 2000 sinh viên một khóa của trường đại học theo giới tính và ngành học thu được các số liệu sau:

	Nam	Nữ
Học tài chính ngân hàng	400	500
Học quản trị kinh doanh	800	300

Lấy ngẫu nhiên một sinh viên khóa đó. Tìm xác suất để nhận được:

a) Sinh viên là Nam.

b) Sinh viên học tài chính ngân hàng.

c) Sinh viên nam và tài chính ngân hàng.

d) Hoặc sinh viên nam, hoặc học tài chính ngân hàng.

e) Nếu đã chọn được một sinh viên nam thì xác suất để người đó học tài chính ngân hàng bằng bao nhiêu?

Đáp số: a) 0,6; b) 0,45; c) 0,2; d) 0,85; e) 1/3.

Bài số 6. Một công ty liên doanh cần tuyển một kế toán trưởng, một trưởng phòng tiếp thị, có 40 người dự tuyển trong đó có 15 nữ. Tính xác suất trong 2 người được tuyển có:

a) kế toán trưởng là nữ,

b) ít nhất 1 nữ.

Đáp số: a) 0,375;b) 0,6154.

Bài số 7. Một lô hàng có 10 sản phẩm, trong đó có 7 sản phẩm tốt, 3 sản phẩm xấu. Lấy ngẫu nhiên từ lô hàng ra 4 sản phẩm. Tính xác suất để 4 sản phẩm lấy ra có 3 sản phẩm tốt.

Đáp số: 0,5.

Bài số 8. Một lớp học có 50 học sinh trong kỳ thi giỏi Toán và Văn, trong đó có 20 người giỏi Toán, 25 người giỏi Văn, 10 người giỏi cả Toán lẫn Văn. Chọn ngẫu nhiên 1 học sinh của lớp này. Tính xác suất để học sinh được chọn giỏi Toán hoặc Văn.

Đáp số: 0,7.

Bài số 9. Trong 1 khu phố, tỷ lệ người mắc bệnh tim là 6%; mắc bệnh phổi là 8% và mắc cả hai bệnh là 5%. Chọn ngẫu nhiên 1 người trong khu phố đó. Tính xác suất để người đó không mắc cả 2 bệnh tim và bệnh phổi.

Đáp số: 0,91.

Bài số 10. Trong 100 người phỏng vấn có 40 người thích dùng nước hoa A, 28 người thích dùng nước hoa B, 10 người thích dùng cả 2 loại A, B. Chọn ngẫu nhiên 1 người trong số 100 người trên. Tính xác suất người này :

a) thích dùng ít nhất 1 loại nước hoa trên,

b) không dùng loại nào cả.

Đáp số: a) 0,58; b) 0,42.

Bài số 11. Một cơ quan có 210 người, trong đó có 100 người ở gần cơ quan, 60 người trong 100 người gần cơ quan là nữ, biết rằng số nữ chiếm gấp đôi số nam trong cơ quan.

Chọn ngẫu nhiên 1 người trong cơ quan. Tính xác suất :

a) người này là nam,

b) người này ở gần cơ quan,

c) người này phải trực đêm (người trực đêm phải ở gần cơ quan hoặc là nam).

Đáp số: a) 1/3; b) 0,4762; c) 0,619.

Bài số 12. Cho A và B là 2 biến cố sao cho $P(A)=\dfrac{1}{2}$, $P(B)=\dfrac{1}{3}$, $P(AB)=\dfrac{1}{6}$. Hãy tính:

1) $P(A+B)$ 2) $P(\overline{A}+\overline{B})$ 3) $P(\overline{A+B})$

4) $P(\overline{AB})$ 5) $P(A\overline{B})$ 6) $P(\overline{A}B)$

7) $P(\overline{A}+B)$ 8) $P(A|B)$ 9) $P(\overline{A}|B)$

10) $P(AB|B)$ 11) $P(A\overline{B}|B)$ 12) $P(A\overline{B}|\overline{B})$

Bài số 13. Đội tuyển cầu lông của Trường Đại học Tài chính - Marketing có 3 vận động viên, mỗi vận động viên thi đấu một trận. Xác suất thắng trận của các vận viên A, B, C lần lượt là: 0,9; 0,7; 0,8. Tính xác suất :

a) Đội tuyển thắng ít nhất 1 trận,

b) Đội tuyển thắng 2 trận,

c) C thua, biết rằng đội tuyển thắng 2 trận.

Đáp số: a) 0,994; b) 0,398; c) 0,3166.

Bài số 14. Cho 3 biến cố A, B, C sao cho

P(A) = 0,5; P(B) = 0,7; P(C) = 0,6; P(AB) = 0,3; P(BC) = 0,4; P(AC) = 0,2 và P(ABC) = 0,1.

a) Tìm xác suất để cả 3 biến cố A, B, C đều không xảy ra.

b) Tìm xác suất để có đúng 2 trong 3 biến cố đó xảy ra.

c) Tìm xác suất để chỉ có đúng 1 biến cố trong 3 biến cố đó xảy ra.

Đáp số: a) 0; b) 0,6; c) 0,3.

Bài số 15. Một người có 5 con gà mái, 2 con gà trống nhốt chung trong một cái lồng. Một người đến mua, người bán gà bắt ngẫu nhiên 1 con. Người mua chấp nhận con đó.

a) Tính xác suất để người đó mua được con gà mái.

Người thứ hai lại đến mua, người bán gà lại bắt ngẫu nhiên ra 1 con.

b) Tìm xác suất để người thứ hai mua được con gà trống.

c) Xác suất này sẽ bằng bao nhiêu nếu người bán gà quên mất rằng con gà bán cho người thứ nhất là gà trống hay gà mái.

Đáp số: a) $\frac{5}{7}$; *b)* $\frac{1}{3}$; *c)* $\frac{2}{7}$.

Bài số 16. Hai công ty A, B cùng kinh doanh một mặt hàng. Xác suất để công ty A thua lỗ là 0,2; xác suất để công ty B thua lỗ là 0,4. Tuy nhiên trên thực tế, khả năng cả 2 công ty cùng thua lỗ là 0,1. Tìm xác suất để

a) chỉ có một công ty thua lỗ,

b) có ít nhất một công ty làm ăn không thua lỗ.

Đáp số: a) 0,4; b) 0,9.

Bài số 17. Một thủ quỹ có một chùm chìa khóa gồm 12 chiếc bề ngoài giống hệt nhau, trong đó có 4 chìa mở được cửa chính của thư viện. Cô ta thử từng chìa một một cách ngẫu nhiên, chìa nào không trúng thì bỏ ra. Tìm xác suất để cô ta mở được cửa chính của thư viện ở lần mở thứ 5.

Đáp số : 0,071.

Bài số 18. Một lô hàng có 6 sản phẩm tốt, 4 sản phẩm xấu. Lấy ngẫu nhiên không hoàn lại từng sản phẩm cho đến khi lấy được 2 sản phẩm tốt thì ngừng,

a) Tính xác suất để ngừng lại ở lần lấy sản phẩm thứ 2,

b) Biết đã ngừng lại ở lần lấy sản phẩm thứ 4. Tính xác suất để lần lấy thứ nhất lấy được sản phẩm tốt.

Đáp số a) $\frac{1}{3}$; *b)* $\frac{1}{3}$.

Bài số 19. Một chàng trai viết 4 lá thư cho 4 cô gái; nhưng vì đãng trí nên anh ta bỏ 4 lá thư vào 4 phong bì một cách ngẫu nhiên, dán kín rồi mới ghi địa chỉ gửi,

a) Tính xác suất để không có cô nào nhận đúng thư viết cho mình,

b) Tính xác suất để có ít nhất 1 cô nhận đúng thư của mình,

c) Tổng quát hóa với n cô gái. Tính xác suất có ít nhất 1 cô nhận đúng thư. Xấp xỉ giá trị xác suất này khi cho $n \to \infty$.

Đáp số a) 0,625; b) 0,375; c) e^{-1}.

Bài số 20. Trong 1 lô hàng 10 sản phẩm có 2 sản phẩm xấu, chọn không hoàn lại để phát hiện ra 2 sản phẩm xấu, khi nào chọn được sản phẩm xấu thứ 2 thì dừng lại.

a) Tính xác suất dừng lại ở lần chọn thứ 4.

b) Biết rằng đã chọn được sản phẩm xấu ở lần chọn thứ nhất, tính xác suất dừng lại ở lần chọn thứ 4.

c) Nếu việc kiểm tra dừng lại ở lần chọn thứ 3, tính xác suất lần chọn đầu được sản phẩm xấu.

Đáp số : a) $\frac{1}{15}$; *b)* $\frac{1}{9}$; *c)* $\frac{1}{2}$.

Bài số 21. Đội tuyển bóng bàn Thành phố có 4 vận động viên A, B, C, D . Mỗi vận động viên thi đấu 1 trận, với xác suất thắng trận lần lượt la : 0,6, 0,7, 0,8, 0,9. Tính

a) xác suất đội tuyển thắng ít nhất 1 trận,

b) xác suất đội tuyển thắng 2 trận,

c) xác suất đội tuyển thắng 3 trận,

d) xác suất D thua, trong trường hợp đội tuyển thắng 3 trận.

Đáp số: a) 0,9976; b) 0,2144; c) 0,4404; d) 0,763.

Bài số 22. Ở một cơ quan nọ có 3 chiếc ôtô. Khả năng có sự cố của mỗi xe ôtô lần lượt là 0,15 ; 0,20 ; 0,10.

a) Tìm khả năng 3 ôtô cùng bị hỏng.

b) Tìm khả năng có ít nhất 1 ôtô hoạt động tốt.

c) Tìm khả năng cả 3 ôtô cùng hoạt động được.

d) Tìm xác suất có không quá 2 ôtô bị hỏng.

Công thức xác suất đầy đủ và công thức Bayes

Đáp số: a) 0,003; b) 0,997; c) 0,612; d) 0,997.

Bài số 23. Một nhà máy sản xuất bóng đèn, máy A sản xuất 25%, máy B: 35%, máy C: 40% số bóng đèn. Tỉ lệ sản phẩm hỏng của mỗi máy trên số sản phẩm do máy đó sản xuất lần lượt là 3%, 2%, 1%. Một người mua 1 bóng đèn do nhà máy sản xuất.

a) Tính xác suất để sản phẩm này tốt.

b) Biết rằng sản phẩm này là xấu. Tính xác suất để sản phẩm do máy C sản xuất.

Đáp số: a) 0,9815;b) 0,2162.

Bài số 24. Trong một trạm cấp cứu bỏng : 80% bệnh nhân bỏng do nóng, 20% bỏng do hóa chất. Loại bỏng do nóng có 30% bi biến chứng, loại bỏng do hóa chất có 50% bị biến chứng.

a) Chọn ngẫu nhiên một bệnh án. Tính xác suất để gặp một bệnh án của bệnh nhân bị biến chứng.

b) Rút ngẫu nhiên được một bệnh án của một bệnh nhân bị biến chứng. Tính xác suất để bệnh án đó là của bệnh nhân bị biến chứng do nóng gây ra? do hóa chất gây ra?

Đáp số: a) 0,34; b) 0,7059; 0,2941.

Bài số 25. Một lô hạt giống được phân thành ba loại. Loại 1 chiếm 2/3 số hạt cả lô, loại 2 chiếm 1/4, còn lại là loại 3. Loại 1 có tỉ lệ nẩy mầm 80%, loại 2 có tỉ lệ nẩy mầm 60% và loại 3 có tỉ lệ nẩy mầm 40%. Hỏi tỉ lệ nẩy mầm chung của lô hạt giống là bao nhiêu?

Đáp số: 0,72.

Bài số 26. Hai nhà máy cùng sản xuất 1 loại linh kiện điện tử. Năng suất nhà máy hai gấp 3 lần năng suất nhà máy một. Tỷ lệ hỏng của nhà máy một và hai lần lượt là 0,1% và 0,2%. Giả sử linh kiện bán ở Trung tâm chỉ do hai nhà máy này sản xuất. Mua 1 linh kiện ở Trung tâm.

a) Tính xác suất để linh kiện ấy hỏng.

b) Giả sử mua linh kiện và thấy linh kiện bị hỏng. Theo ý bạn thì linh kiện đó do nhà máy nào sản xuất.

Đáp số: a) 0,175%; b) nhà máy 2.

Bài số 27. Có 8 bình đựng bi, trong đó có :

2 bình loại 1: mỗi bình đựng 6 bi trắng 3 bi đỏ,

3 bình loại 2: mỗi bình đựng 5 bi trắng 4 bi đỏ,

3 bình loại 3: mỗi bình đựng 2 bi trắng 7 bi đỏ.

Lấy ngẫu nhiên một bình và từ bình đó lấy ngẫu nhiên 1 bi.

a) Tính xác suất để bi lấy ra là bi trắng.

b) Biết rằng bi lấy ra là bi trắng. Tính xác suất để bình lấy ra là bình loại 3.

Đáp số: a) 0,4583; b) 0,182.

Bài số 28. Một chuồng gà có 9 con gà mái và 1 con gà trống. Chuồng gà kia có 1 con mái và 5 con trống. Từ mỗi chuồng lấy ngẫu nhiên 1 con đem bán. Các con gà còn lại được dồn vào chuồng thứ ba. Nếu ta lại bắt ngẫu nhiên 1 con gà nữa từ chuồng này ra thì xác suất để bắt được con gà trống là bao nhiêu?

Đáp số: 0,3619.

Bài số 29. Có 2 hộp áo; hộp một có 10 áo trong đó có 1 phế phẩm; hộp hai có 8 áo trong đó có 2 phế phẩm. Lấy ngẫu nhiên 1 áo từ hộp một bỏ sang hộp hai; sau đó từ hộp này chọn ngẫu nhiên ra 2 áo. Tìm xác suất để cả 2 áo này đều là phế phẩm.

Đáp số: 1/30.

Bài số 30. Có 3 xạ thủ cùng bắn vào một con thú, mỗi người bắn 1 viên đạn, với xác suất bắn trúng lần lượt là 0,6; 0,7; 0,8. Biết rằng nếu trúng 1 phát đạn thì xác suất để con thú bị tiêu diệt là 0,5; trúng 2 phát thì xác suất để con thú bị tiêu diệt là 0,8; còn nếu trúng 3 phát đạn thì chắc chắn con thú bị tiêu diệt.

a) Tính xác suất con thú bị tiêu diệt.

b) Giả sử con thú bị tiêu diệt. Tính xác suất nó bị trúng 2 phát đạn.

Đáp số: a) 0,7916; b) 0,4568.

Bài số 31. Có 3 hộp bi; hộp một có 10 bi trong đó có 3 bi đỏ; hộp hai có 15 bi trong đó có 4 bi đỏ; hộp ba có 12 bi trong đó có 5 bi đỏ. Gieo một con xúc xắc. Nếu xuất hiện mặt 1 thì chọn hộp một, xuất hiện mặt hai thì chọn hộp 2, xuất hiện các mặt còn lại thì chọn hộp ba. Từ hộp được chọn, lấy ngẫu nhiên 1 bi

a) Tính xác suất để được bi đỏ,

b) Giả sử lấy được bi đỏ. Tính xác suất để bi đỏ này thuộc hộp hai.

Đáp số: a) 0,3722; b) 0,1194.

Bài số 32. Một hộp có 15 quả bóng bàn, trong đó có 9 mới 6 cũ, lần đầu chọn ra 3 quả để sử dụng, sau đó bỏ vào lại, lần hai chọn ra 3 quả.

a) Tính xác suất 3 quả bóng chọn lần hai là 3 bóng mới.

b) Biết rằng lần hai chọn được 3 bóng mới, tính xác suất lần đầu chọn được 2 bóng mới.

Đáp số: a) 0,0893; b) 0,4089.

Bài số 33. Có 3 cái thùng. Thùng 1 có 6 bi trắng, 4 bi đỏ; thùng 2 có 5 bi trắng, 5 bi đỏ và thùng 3 có 10 bi trắng. Giả sử người ta lấy ngẫu nhiên 2 bi từ thùng 1 bỏ vào thùng 2. Sau đó, lại lấy ngẫu nhiên 1 bi từ thùng 2 bỏ vào thùng 3 rồi từ thùng 3 lấy ngẫu nhiên ra 1 bi. Tìm xác suất để bi lấy ra là đỏ.

Đáp số: 0,0439.

Công thức Bernoulli

Bài số 34. Một bác sĩ chữa khỏi bệnh A cho một người với xác suất là 95%. Giả sử có 10 người bị bệnh A đến chữa một cách độc lập nhau. Tính xác suất để

a) Có 8 người khỏi bệnh,

b) Có nhiều nhất 9 người khỏi bệnh.

Đáp số: a) 0,0746; b) 0,4013.

Bài số 35. Một thiết bị có 10 chi tiết với độ tin cậy của mỗi chi tiết là 0,9. (Xác suất làm việc tốt trong khoảng thời gian nào đó).

Tính xác suất để trong khoảng thời gian ấy :

a) Có đúng một chi tiết làm việc tốt,

b) Có ít nhất 2 chi tiết làm việc tốt.

Đáp số: a) $9 \cdot 10^{-9} \approx 0$; *b)* 1.

Bài số 36. Một cầu thủ đá thành công quả phạt 11m với xác suất 80%.

- Đá 4 thành công 2.

- Đá 6 thành công 3.

Công việc nào dễ thực hiện ?

Đáp số: Đá 4 thành công 2 dễ hơn.

Bài số 37. Trong một thành phố có 70% dân cư thích xem bóng đá. Chọn ngẫu nhiên 10 người, tính xác suất có :

a) 5 người thích xem bóng đá,

b) ít nhất 2 người thích xem bóng đá.

Đáp số: a) 0,1029; b) 0,9999.

Bài số 38. Một nhà toán học có xác suất giải được một bài toán khó là 0,9. Cho nhà toán học này 5 bài toán khó được chọn một cách ngẫu nhiên.

a) Tính xác suất để nhà toán học này giải được 3 bài.

b) Tính xác suất để nhà toán học này giải được ít nhất 1 bài.

c) Tính số bài có khả năng nhất mà nhà toán học này giải được.

Đáp số: a) 0,0729; b) 0,99999; c) 5 bài.

Bài số 39. Tỷ lệ mắc bệnh Basedow ở một vùng rừng núi nào đó là 70%. Trong đợt khám tuyển sức khoẻ để xuất cảnh, người ta khám cho 100 người. Tìm xác suất để

a) Trong 100 người có 60 người bị Basedow,

b) Trong 100 người có 75 người bị Basedow,

c) Trong 100 người có ít nhất một người bị Basedow.

Đáp số: a) 0,0085; b) 0,0496; c) 1.

Bài số 40. Một lô hàng với tỷ lệ phế phẩm là 5%. Cần phải lấy mẫu cỡ bao nhiêu sao cho xác suất để bị ít nhất một phế phẩm không bé hơn 0,95.

Đáp số: $n \geq 59$.

Bài số 41. Hai đấu thủ A, B thi đấu cờ. Xác suất thắng của người A trong một ván là 0,6 (không có hòa). Trận đấu bao gồm 5 ván, người nào thắng một số ván lớn hơn là người thắng cuộc. Tính xác suất để người B thắng cuộc.

Đáp số: 0,31744.

Bài số 42. Một máy sản xuất lần lượt từng sản phẩm. Xác suất sản xuất ra một phế phẩm của máy là 0,01.

a) Cho máy sản xuất 10 sản phẩm. Tính xác suất để có 2 phế phẩm.

b) Máy cần sản xuất ít nhất bao nhiêu sản phẩm để xác suất có ít nhất một chính phẩm trên 0,99.

Đáp số: a) 0,0042; b) 2.

Bài số 43. Một xí nghiệp có hai phân xưởng A và B cùng sản xuất một loại sản phẩm với tỉ lệ phế phẩm tương ứng là 2% và 3%. Cho mỗi phân xưởng sản xuất ra 5 sản phẩm. Tính xác suất để số phế phẩm do hai phân xưởng sản xuất là bằng nhau.

Đáp số:0,7885.

1.7. Tài liệu tham khảo

[1] Nguyễn Cao Văn, Trần Thái Ninh, Ngô Văn Thứ, Lý thuyết xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[2] Nguyễn Cao Văn, Trần Thái Ninh, Nguyễn Thế Hệ, Bài tập xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[3] Phạm Hoàng Uyên, Lê Thị Thiên Hương, Huỳnh Văn Sáu, Nguyễn Phúc Sơn, Huỳnh Tố Uyên, Lý thuyết xác suất, NXB Đại học Quốc gia Thành Phố Hồ Chí Minh, 2015.

[4] Anderson, Sweeney, and William [2010], Statistics for Business and Economics, South-Western Cengage Learning (11^{th} Edition).

[5] Michael Barrow, Statistics for Economics, Accounting and Business Studies-Prentice Hall, 2006.

[6] Newbold Paul - Statistics for Bussiness and Economics, 5th edition - Prentice Hall, 2005.

Thuật ngữ chính chương 1	
Tiếng Anh	**Tiếng Việt**
Axiom	Tiên đề
Addition rule	Công thức cộng
Bayes’ formula	Công thức Bayes
Countable additivity	Công thức cộng
Classical probability	Xác suất cổ điển
Conditional probability	Xác suất có điều kiện
Define	Định nghĩa
Event	Biến cố
Element	Phần tử
Experiment	Phép thử
Experimental probability	Xác suất thực nghiệm
Event A occurs	Biến cố A xảy ra
Independent	Độc lập
Infinite sequence of outcomes	Dãy vô hạn kết cục
Monotonicity	Tính đơn điệu
Mutually exclusive events	Họ biến cố xung khắc
Multiplication rule	Công thức nhân
Outcome	Kết cục
Odds in favor	Tỷ lệ thuận lợi
Relative frequency	Tần số tương đối
Probability	Xác suất
Posterior probability	Xác suất hậu nghiệm
Prior probability	Xác suất tiên nghiệm
Sample space	Không gian mẫu
Subset	Tập con
Sequence of mutually exclusive events	Một dãy các biến cố xung khắc từng đôi một
The finite additivity of the probability	Tính cộng hữu hạn của xác suất
The law of large numbers	Luật số lớn
The number of elements	Số lượng các phần tử

Chương 2 **ĐẠI LƯỢNG NGẪU NHIÊN VÀ PHÂN PHỐI XÁC SUẤT**

Mục tiêu chương 2

Chương này giúp sinh viên:

- Phân biệt được thế nào là đại lượng ngẫu nhiên liên tục và đại lượng ngẫu nhiên rời rạc.
- Thành lập được bảng phân phối xác suất và tính được các hàm xác suất, hàm phân phối.
- Tính và nêu được ý nghĩa các tham số đặc trưng như trung bình, phương sai, mốt, trung vị,...
- Biết áp dụng các luật phân phối như: Nhị thức, Siêu bội, Poisson, Chuẩn, ….

2.1. Đại lượng ngẫu nhiên

2.1.1. Khái niệm

Xét phép thử τ với không gian mẫu Ω. Giả sử, ứng với mỗi biến số sơ cấp $w \in \Omega$, ta liên kết với một số thực $X(w) \in \mathbb{R}$, thì X được gọi là một *biến số ngẫu nhiên*.

Ví dụ 2.1. Với trò chơi sấp ngửa bằng cách tung đồng xu, giả sử nếu xuất hiện mặt sấp, ta được 1 đồng; nếu xuất hiện mặt ngửa, ta mất 1 đồng. Khi đó, ta có

Phép thử τ : "*tung đồng xu*",

Không gian mẫu $\Omega = \{S, N\}$,

Biến số ngẫu nhiên X với $X(S) = 1$ và $X(N) = -1$.

Tổng quát, biến số ngẫu nhiên X của một phép thử τ với không gian mẫu Ω là một ánh xạ

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ w &\mapsto X(w) \end{aligned}$$

2.1.2. Phân loại

- Khi $X(\Omega)$ là một tập hợp hữu hạn $\{x_1, x_2, ..., x_k\}$ hay là một dãy $\{x_1, x_2, ..., x_n, ...\}$, ta nói X là một *biến số ngẫu nhiên rời rạc.*

Ví dụ 2.2. Số chấm xuất hiện ở mặt trên của xúc xắc; số sinh viên vắng mặt trong một buổi học; số máy hỏng trong từng ca sản xuất,… là các biến ngẫu nhiên rời rạc.

- Khi $X(\Omega)$ là một khoảng của $\mathbb{R}$ (hay cả $\mathbb{R}$), ta nói X là một *biến số ngẫu nhiên liên tục*.

Ví dụ 2.3. Gọi X là kích thước của chi tiết do một máy sản xuất ra, X là biến ngẫu nhiên liên tục.

Do các biến số ngẫu nhiên X là các ánh xạ có giá trị trong $\mathbb{R}$ nên với một hàm số $u:\mathbb{R}\to\mathbb{R}$, ta có thể thành lập biến số ngẫu nhiên $u(X)$, với

$$\begin{aligned} u(X):\Omega &\to \mathbb{R} \\ w &\mapsto u[X(w)] \end{aligned}$$

Chẳng hạn, với $u(x)=x-\mu$, $\mu\in\mathbb{R}$ ta có biến số ngẫu nhiên $u(X)=X-\mu$, với $(X-\mu)(w)=X(w)-\mu$. với mọi $w\in\Omega$, và với $u(x)=(x-\mu)^2$, $\mu\in\mathbb{R}$ ta có biến số ngẫu nhiên $u(X)=(X-\mu)^2$, với $(X-\mu)^2(w)=(X(w)-\mu)^2$ với mọi $w\in\Omega$.

2.2. Phân phối xác suất của đại lượng ngẫu nhiên

2.2.1. Đại lượng ngẫu nhiên rời rạc

Để xác định một biến số ngẫu nhiên rời rạc, người ta cần xác định các giá trị x_i, $i=1,2,...$có thể nhận được bởi biến ngẫu nhiên này và đồng thời cũng cần xác định xác suất để X nhận giá trị này là bao nhiêu, nghĩa là, cần xác định

$$P\{w\in\Omega: X(w)=x_i\}\equiv P(X=x_i),\ \text{với } i=1,2,...$$

2.2.1.1. Bảng phân phối xác suất

Xét biến số ngẫu nhiên rời rạc $X:\Omega\to\mathbb{R}$, với $X(\Omega)=\{x_1,x_2,...,x_n,...\}$. Giả sử $x_1<x_2<\cdots<x_n<\cdots$. Ta lập bảng các giá trị tương ứng

X	x_1	x_2	...	x_n	...
P	p_1	p_2	...	p_n	...

với $p_i=P(X=x_i)$, được gọi là *bảng phân phối xác suất* của X.

Ví dụ 2.4. Trong hộp có 10 sản phẩm trong đó có 6 chính phẩm. Lấy ngẫu nhiên 2 sản phẩm. Lập bảng phân phối xác suất cho số chính phẩm được lấy ra.

Giải

Gọi X là số chính phẩm được lấy ra trong 3 sản phẩm, $X\in\{0,1,2\}$.

$$P(X=0)=\frac{C_4^2}{C_{10}^2}=\frac{6}{45}=\frac{2}{15},$$

$$P(X=1)=\frac{C_6^1.C_4^1}{C_{10}^2}=\frac{24}{45}=\frac{8}{15},$$

$$P(X=2)=\frac{C_6^2}{C_{10}^2}=\frac{15}{45}=\frac{1}{3}.$$

Vậy bảng phân phối xác suất của X là:

X	0	1	2
P	$\frac{2}{15}$	$\frac{8}{15}$	$\frac{1}{3}$

Ví dụ 2.5. Một cơ quan có 3 xe ôtô : 1 xe 4 chỗ; 1 xe 50 chỗ và 1 xe tải. Xác suất để trong một ngày làm việc, các xe được xử dụng là 0,8; 0,4 và 0,9. Hãy lập bảng phân phối xác suất cho số xe được xử dụng trong một ngày của cơ quan.

Giải

Gọi X là số xe được xử dụng trong một ngày của cơ quan. Ta có $X \in \{0,1,2,3\}$.

Gọi A_1, A_2, A_3 lần lượt là biến cố *"xe 4 chỗ"; "xe 50 chỗ"; "xe tải"* được xử dụng trong ngày của cơ quan. Khi đó, A_1, A_2, A_3 là các biến cố độc lập, $P(A_1)=0,8; P(A_2)=0,4; P(A_3)=0,9$ và

$$\begin{aligned}P(X=0)&=P(\overline{A}_1\overline{A}_2\overline{A}_3)=P(\overline{A}_1)P(\overline{A}_2)P(\overline{A}_3)\\&=0,2\cdot 0,6\cdot 0,1=0,012\end{aligned}$$

$$\begin{aligned}P(X=1)&=P(A_1\overline{A}_2\overline{A}_3+\overline{A}_1A_2\overline{A}_3+\overline{A}_1\overline{A}_2A_3)\\&=P(A_1\overline{A}_2\overline{A}_3)+P(\overline{A}_1A_2\overline{A}_3)+P(\overline{A}_1\overline{A}_2A_3)\\&=P(A_1)P(\overline{A}_2)P(\overline{A}_3)+P(\overline{A}_1)P(A_2)P(\overline{A}_3)+P(\overline{A}_1)P(\overline{A}_2)P(A_3)\\&=0,8\cdot 0,6\cdot 0,1+0,2\cdot 0,4\cdot 0,1+0,2\cdot 0,6\cdot 0,9=0,164\end{aligned}$$

$$\begin{aligned}P(X=2)&=P(A_1A_2\overline{A}_3+A_1\overline{A}_2A_3+\overline{A}_1A_2A_3)\\&=P(A_1A_2\overline{A}_3)+P(A_1\overline{A}_2A_3)+P(\overline{A}_1A_2A_3)\\&=P(A_1)P(A_2)P(\overline{A}_3)+P(A_1)P(\overline{A}_2)P(A_3)+P(\overline{A}_1)P(A_2)P(A_3)\\&=0,8\cdot 0,4\cdot 0,1+0,8\cdot 0,6\cdot 0,9+0,2\cdot 0,4\cdot 0,9=0,536\end{aligned}$$

$$P(X=3)=P(A_1A_2A_3)=P(A_1)P(A_2)P(A_3)=0,8\cdot 0,4\cdot 0,9=0,288$$

Do đó, bảng phân phối xác suất của cho X là

X	0	1	2	3
P	0,012	0,164	0,536	0,288

2.2.1.2. Hàm xác suất

Hàm số $f:\mathbb{R}\to\mathbb{R}$ xác định bởi

$$f(x)=\begin{cases} p_i & \text{khi} \quad x=x_i, \\ 0 & \text{khi} \quad x\neq x_i, \forall i. \end{cases}$$

được gọi là *hàm xác suất* của X. Từ tính chất của bảng phân phối xác suất, ta có

(i) $\forall x\in\mathbb{R}, f(x)\geq 0,$ và

(ii) $\sum_x f(x)=1.$

Ví dụ 2.6. Từ bảng phân phối xác suất của ví dụ 2.5. Ta có hàm xác suất của X như sau:

$$f(x)=\begin{cases} 0{,}012 & \text{khi} \quad x=0, \\ 0{,}164 & \text{khi} \quad x=1, \\ 0{,}536 & \text{khi} \quad x=2, \\ 0{,}288 & \text{khi} \quad x=3, \\ 0 & \text{khi} \quad x\neq 0,1,2,3. \end{cases}$$

2.2.1.3. Hàm phân phối (tích lũy)

Cho $f:\mathbb{R}\to\mathbb{R}$ là hàm xác suất của biến ngẫu nhiên rời rạc X, hàm số $F:\mathbb{R}\to\mathbb{R}$, được xác định bởi

$$F(x)=P(X\leq x)=\sum_{x\geq x_i} f(x_i),$$

được gọi là *hàm phân phối tích lũy*, hay vắn tắt là *hàm phân phối*, của X.

Bằng cách liệt kê các giá trị của $X(\Omega)$ theo thứ tự tăng dần, khi X lấy giá trị tạo thành một dãy $x_1<x_2<\cdots<x_n<\cdots$ ta có hàm phân phối

$$F(x)=\begin{cases} 0 & \text{khi} \quad x<x_1, \\ f(x_1)+f(x_2)+\cdots+f(x_{n-1}) & \text{khi} \quad x_{n-1}\leq x<x_n, \\ 1 & \text{khi} \quad x\geq x_n. \end{cases}$$

Từ tính chất của hàm xác suất và định nghĩa của hàm phân phối, ta có

(i) $0\leq F(x)\leq 1, \forall x\in\mathbb{R},$

(ii) $\lim\limits_{x\to-\infty} F(x)=0$ và $\lim\limits_{x\to+\infty} F(x)=1,$

(iii) F là hàm tăng, và F liên tục bên phải tại mọi $x\in\mathbb{R}.$

Ví dụ 2.7. Với biến số ngẫu nhiên X cho bởi ví dụ 2.6, ta có hàm phân phối

$$F(x)=\begin{cases}0,012 & \text{khi} \quad x<0,\\ 0,164 & \text{khi} \quad 0\le x<1,\\ 0,536 & \text{khi} \quad 1\le x<2,\\ 0,288 & \text{khi} \quad 2\le x<3,\\ 0 & \text{khi} \quad 3\le x.\end{cases}$$

2.2.2. Đại lượng ngẫu nhiên liên tục

2.2.2.1. Hàm mật độ (xác suất)

Hàm số $f:\mathbb{R}\to\mathbb{R}$ được gọi là *hàm mật độ xác suất*, hay vắn tắt là *hàm mật độ*, của biến số ngẫu nhiên liên tục X nếu

$$P(a\le X\le b)=\int_a^b f(x)dx$$

với mọi $a,b\in\mathbb{R}$, $a\le b$.

Từ định nghĩa, dễ dàng suy ra

(i) $\forall x\in\mathbb{R}, f(x)\ge 0,$ và

(ii) $\int\limits_{-\infty}^{+\infty} f(x)dx=1.$

Ví dụ 2.8. Cho biến số ngẫu nhiên liên tục X với hàm mật độ xác suất

$$f(x)=\begin{cases}Ax(3-x) & \text{khi} \quad x\in[0,3],\\ 0 & \text{khi} \quad x\notin[0,3].\end{cases}$$

Xác định hằng số A.

Giải

+) Do $f(x)\ge 0, \forall x\in\mathbb{R}$ nên $A>0$

+) Ta có

$$\int_{-\infty}^{+\infty} f(x)dx=1 \Leftrightarrow \int_{-\infty}^{0} f(x)dx+\int_0^3 f(x)dx+\int_3^{+\infty} f(x)dx=1$$

$$\Leftrightarrow A\int_0^3\left(3x-x^2\right)dx=1\Leftrightarrow \frac{9}{2}A=1\Leftrightarrow A=\frac{2}{9}.$$

2.2.2.2. Hàm phân phối (tích lũy)

Cho $f:\mathbb{R}\to\mathbb{R}$ là hàm mật độ của biến ngẫu nhiên liên tục X, hàm số $F:\mathbb{R}\to\mathbb{R}$, được gọi là *hàm phân phối tích lũy*, hay vắn tắt là *hàm phân phối*, của biến số ngẫu nhiên liên tục X nếu

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

với mọi $x \in \mathbb{R}$.

Trực tiếp từ định nghĩa, ta được

(i) $0 \le F(x) \le 1, \forall x \in \mathbb{R}$,

(ii) $\lim_{x \to -\infty} F(x) = 0$ và $\lim_{x \to +\infty} F(x) = 1$,

(iii) F là hàm tăng, và F liên tục bên phải tại mọi $x \in \mathbb{R}$.

Ví dụ 2.9. Cho biến số ngẫu nhiên liên tục X với hàm mật độ xác suất như ví dụ 2.8. Tìm hàm phân phối của X.

Giải

+) Trường hợp 1. Nếu $x < 0$ thì

$$F(x) = \int_{-\infty}^{x} f(t)dt = 0$$

+) Trường hợp 2. Nếu $0 \le x < 3$ thì

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_{0}^{x} f(t)dt = \int_{0}^{x} \left(3t - t^2\right)dt = \frac{3}{2}x^2 - \frac{1}{3}x^3.$$

+) Trường hợp 3. Nếu $3 \le x$ thì

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} f(t)dt + \int_{0}^{3} f(t)dt + \int_{3}^{+\infty} f(t)dt = \int_{0}^{3} \left(3t - t^2\right)dt = 1.$$

Vậy hàm phân phối của X là

$$F(x) = \begin{cases} 0 & \text{khi } x < 0, \\ \frac{3}{2}x^2 - \frac{1}{3}x^3 & \text{khi } 0 \le x < 3, \\ 1 & \text{khi } x \ge 3. \end{cases}$$

Ví dụ 2.10. Cho X là biến số ngẫu nhiên liên tục có hàm phân phối xác suất

$$F(x) = \begin{cases} 0 & \text{khi } x \le -1, \\ a\left(b + x - \frac{1}{3}x^3\right) & \text{khi } -1 < x \le 1, \\ 1 & \text{khi } x > 1. \end{cases}$$

Tìm các hằng số a, b.

Do X là biến số ngẫu nhiên liên tục nên hàm phân phối xác suất liên tục bên phải tại mọi $x \in \mathbb{R}$. Đặc biệt, tại $x = \pm 1$, $\lim_{x \to -1^+} F(x) = F(-1)$ cho

$$a\left(b - \frac{2}{3}\right) = 0 \qquad (*)$$

và $\lim_{x \to 1^+} F(x) = F(1)$ cho

$$a\left(b + \frac{2}{3}\right) = 1 \qquad (**)$$

Từ (*) và (**), ta suy ra $a = \frac{3}{4}$ và $b = \frac{2}{3}$.

2.3. Các số đặc trưng của đại lượng ngẫu nhiên

2.3.1. Kỳ vọng

Cho X là đại lượng ngẫu nhiên với hàm xác suất (hàm mật độ xác suất) $f(x)$ và $u(X)$ là một hàm theo biến số ngẫu nhiên X. *Kỳ vọng* của $u(X)$ được xác định là

$E[u(X)] = \sum_i u(x_i) f(x_i)$ khi X là biến ngẫu nhiên rời rạc.

$E[u(X)] = \int_{-\infty}^{+\infty} u(x) f(x) dx$ khi X là biến ngẫu nhiên liên tục.

2.3.2. Trung bình

Cho X là biến ngẫu nhiên với hàm xác suất (hàm mật độ xác suất) $f(x)$, khi $u(X) = X$, thì $E(X)$ được gọi là *trung bình* của X, ký hiệu μ_X, nghĩa là

$E(X) = \sum_i x_i f(x_i)$ khi X là biến số ngẫu nhiên rời rạc, và

$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$ khi X là biến ngẫu nhiên liên tục.

Tính chất:

(i) $E(C) = C$ với C là hằng số.

(ii) $E(aX + bY) = aE(X) + bE(Y)$ (với $a, b \in \mathbb{R}$ và X, Y là hai đại lượng ngẫu nhiên).

2.3.3. Phương sai

Cho X là biến ngẫu nhiên với hàm xác suất (hàm mật độ xác suất) $f(x)$, khi đó với $u(X) = (X - \mu_X)^2$, thì $E(X - \mu_X)^2$, được gọi là *phương sai* của X, ký hiệu σ_X^2 hay $\mathrm{var}(X)$, nghĩa là

$$\sigma_X^2 = \sum_i (x_i - \mu_X)^2 f(x_i) \text{ khi } X \text{ là biến số ngẫu nhiên rời rạc, và}$$

$$\sigma_X^2 = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f(x) dx \text{ khi } X \text{ là biến ngẫu nhiên liên tục.}$$

Tính chất:

(i) $\mathrm{var}(C) = 0$ với C là hằng số.

(ii) $\mathrm{var}(aX + bY) = a^2 \mathrm{var}(X) + b^2 \mathrm{var}(Y)$ (với $a, b \in \mathbb{R}$ và X, Y là hai đại lượng ngẫu nhiên độc lập).

2.3.4. Mệnh đề. Cho X là biến số ngẫu nhiên với trung bình $E(X)$. Ta có

$$\mathrm{var}(X) = E(X^2) - [E(X)]^2$$

Chứng minh. Do E tuyến tính, nghĩa là

$$E(aX + bY) = aE(X) + bE(Y)$$

ta suy ra

$$\mathrm{var}(X) = E\left[(X - \mu_X)^2\right] = E\left[X^2 - 2XE(X) + [E(X)]^2\right]$$
$$= E(X^2) - 2E(X)E(X) + [E(X)]^2 = E(X^2) - [E(X)]^2.$$

2.3.5. Độ lệch chuẩn

Độ lệch chuẩn của biến ngẫu nhiên X, ký hiệu σ_X là căn bậc hai của phương sai.

$$\sigma_X = \sqrt{\sigma_X^2}$$

Chú ý: X, $E(X)$, σ_X có cùng đơn vị đo.

Ví dụ 2.11. Biến ngẫu nhiên rời rạc X có bảng phân phối xác suất:

X	1	3	4
P	0,1	0,5	0,4

Tính trung bình, phương sai và độ lệch chuẩn của X.

Giải

Trung bình của X:

$$E(X) = 1 \cdot 0,1 + 3 \cdot 0,5 + 4 \cdot 0,4 = 3,2.$$

Phương sai của X:

$$\operatorname{var}(X) = (1-3,2)^2 \cdot 0,1 + (3-3,2)^2 \cdot 0,5 + (4-3,2)^2 \cdot 0,4 = 0,76.$$

Độ lệch chuẩn của X:

$$\sigma_X = \sqrt{0,76} \approx 0,872.$$

Ví dụ 2.12. Biến ngẫu nhiên liên tục X có hàm mật độ xác suất:

$$f(x) = \begin{cases} 2x & \text{khi} \quad x \in [0;1], \\ 0 & \text{khi} \quad x \notin [0;1]. \end{cases}$$

Tính trung bình, phương sai và độ lệch chuẩn của X.

Giải

Trung bình của X:

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^1 2x^2 dx = \frac{2}{3}x^3 \Big|_0^1 = \frac{2}{3}.$$

Phương sai của X:

$$\operatorname{var}(X) = \int_{-\infty}^{+\infty} (x-\mu_X)^2 f(x) dx = \int_0^1 \left(x - \frac{2}{3}\right)^2 2x dx$$

$$= \int_0^1 \left(2x^3 - \frac{8}{3}x^2 + \frac{8}{9}x\right) dx = \frac{1}{2}x^4 - \frac{8}{9}x^3 + \frac{4}{9}x^2 \Big|_0^1 = \frac{1}{18}.$$

Độ lệch chuẩn của X:

$$\sigma_X = \sqrt{\frac{1}{18}} \approx 0,2357.$$

2.3.6. Ý nghĩa của kỳ vọng và phương sai

Ý nghĩa kỳ vọng :

- Kỳ vọng toán phản ánh giá trị trung tâm của phân phối xác suất của biến ngẫu nhiên.

- Trong kinh tế, kỳ vọng toán đồng thời mang 2 ý nghĩa:

+ Nếu xét trong 1 số lớn phép thử tương tự thì nó phản ánh giá trị trung bình

+ Nếu xét trong 1 phép thử đơn lẻ thì nó phản ánh giá trị mong đợi.

Ý nghĩa phương sai :

- Phương sai phản ánh mức độ phân tán của các giá trị của biến ngẫu nhiên so với giá trị trung bình. Phương sai càng lớn: phân tán càng nhiều quanh giá trị trung bình còn phương sai càng nhỏ: giá trị càng tập trung quanh giá trị trung bình.

- Trong kinh tế, phương sai phản ánh mức độ rủi ro hay độ biến động (kém ổn định).

Ví dụ 2.13. Nhu cầu hàng ngày về rau sạch ở một khu dân cư có bảng phân phối xác suất.

X	20	21	22	23	24	25	26
P	0,05	0,1	0,2	0,3	0,15	0,12	0,08

Mỗi kg rau mua vào giá 2 ngàn đồng, bán ra 2 ngàn rưỡi. Song nếu bị ế phải bán 1 ngàn rưỡi mới hết. Hàng ngày nên đặt mua 22 kg hay 24 kg rau để bán thì tốt hơn.

Giải

Trường hợp 1. Nếu mua 22 kg thì gọi X_1 là số tiền lãi. Ta có

$$P(X_1 = 11) = P(X \geq 22) = 0,85$$

$$P(X_1 = 10) = P(X = 21) = 0,1$$

$$P(X_1 = 9) = P(X = 20) = 0,05$$

Ta có bảng phân phối xác suất

X_1	9	10	11
P	0,05	0,1	0,85

$E(X_1) = 9 \cdot 0,05 + 10 \cdot 0,1 + 11 \cdot 0,85 = 10,8$ (ngàn đồng)

$$\mathrm{var}(X_1) = (9-10,8)^2 \cdot 0,05 + (10-10,8)^2 \cdot 0,1 + (11-10,8)^2 \cdot 0,85 = 0,26$$

Trường hợp 2. Nếu mua 24 kg thì gọi X_2 là số tiền lãi. Ta có

$$P(X_2 = 12) = P(X \geq 24) = 0,35$$

$$P(X_2 = 11) = P(X = 23) = 0,3$$

$$P(X_2 = 10) = P(X = 22) = 0,2$$

$$P(X_2 = 9) = P(X = 21) = 0,1$$

$$P(X_2 = 8) = P(X = 20) = 0,05$$

Ta có bảng phân phối xác suất

X_2	8	9	10	11	12
P	0,05	0,1	0,2	0,3	0,35

$E(X_2) = 8 \cdot 0,05 + 9 \cdot 0,1 + 10 \cdot 0,2 + 11 \cdot 0,3 + 12 \cdot 0,35 = 10,8$ (ngàn đồng)

$$\mathrm{var}(X_2) = (8-10,8)^2 0,05 + (9-10,8)^2 \cdot 0,1 + (10-10,8)^2 \cdot 0,2 + \\ + (11-10,8)^2 \cdot 0,3 + (12-10,8)^2 \cdot 0,35 = 1,36$$

Vậy đặt mua 22 kg hay 24 kg đều có tiền lãi trung bình 10,8 ngàn.

Vì $\operatorname{var}(X_1) = 0,26 < \operatorname{var}(X_2) = 1,36$ nên đặt mua 22 kg thì độ rủi ro thấp hơn đặt mua 24 kg.

Ví dụ 2.14. Khi đầu tư vào 2 thị trường A và B, lãi suất thu được là biến ngẫu nhiên có bảng phân phối xác suất tương ứng:

X_A	-1	5	8
P	0,2	0,5	0,3

X_B	-2	6	9
P	0,2	0,4	0,4

a) Muốn có lãi trung bình cao nên đầu tư vào đâu?

b) Muốn kinh doanh ổn định thì đầu tư vào đâu?

c) Người ta muốn giảm thiểu độ rủi ro bằng cách đầu tư vào cả hai, nên chia tỉ lệ đầu tư như thế nào biết rằng 2 thị trường A và B là độc lập.

Giải

a) Trung bình lãi suất của hai thị trường

$$E(X_A) = -1 \cdot 0,2 + 5 \cdot 0,5 + 8 \cdot 0,3 = 4,7$$

$$E(X_B) = -2 \cdot 0,2 + 6 \cdot 0,5 + 9 \cdot 0,3 = 5,6$$

Vậy muốn có lãi trung bình cao nên đầu tư vào thị trường B.

b) Phương sai lãi suất của hai thị trường

$$\operatorname{var}(X_A) = (-1)^2 \cdot 0,2 + 5^2 \cdot 0,5 + 8^2 \cdot 0,3 - (4,7)^2 = 9,81$$

$$\operatorname{var}(X_B) = (-2)^2 \cdot 0,2 + 6^2 \cdot 0,4 + 9^2 \cdot 0,4 - (5,6)^2 = 16,24.$$

Vậy muốn kinh doanh ổn định nên đầu tư vào thị trường A.

c) Gọi a là tỷ lệ tiền lãi đầu tư vào thị trường A và $(1-a)$ là tỷ lệ tiền lãi đầu tư vào thị trường B. Khi đó, tiền lãi: $Z = aX_A + (1-a)X_B$.

Ta có

$$\operatorname{var}(Z) = a^2 \operatorname{var}(X_A) + (1-a)^2 \operatorname{var}(X_B) = 9,81a^2 + 16,24(1-a)^2$$

Bài toán tìm a sao cho $\operatorname{var}(Z) \to \min$

Đặt $f(a) = 9,81a^2 + 16,24(1-a)^2$

Đạo hàm cấp 1: $f'(a) = 52,1a - 32,48$

Cho $f'(a) = 0 \Leftrightarrow a = 0,6234$

Đạo hàm cấp 2: $f''(a) = 52,1 > 0$

Với $a = 0,6234$ thì $f(a)$ đạt giá trị nhỏ nhất

Vậy đầu tư 62,34% vốn vào thị trường A và 37,66% vốn vào thị trường B sẽ giảm thiểu được rủi ro.

2.3.7. Mốt và trung vị

Mốt của đại lượng ngẫu nhiên rời rạc X, ký hiệu $M_0(X)$, là giá trị x_0 của X sao cho $P(X = x_0)$ là lớn nhất. Người ta còn nói rằng $M_0(X)$, là *giá trị tin chắc nhất* của X. Trong trường hợp X là biến số ngẫu nhiên liên tục với hàm mật độ $f_X(x)$ thì $M_0(X)$, là giá trị x_0 của X sao cho $f_X(x_0)$ là lớn nhất.

Trung vị của đại lượng ngẫu nhiên (rời rạc hay liên tục), ký hiệu $Me(X)$ là giá trị x_0 của X sao cho $P(X \le x_0) = P(X \ge x_0) = 0,5.$

Chú ý rằng Mốt cũng như trung vị của một đại lượng ngẫu nhiên thì không duy nhất.

Ví dụ 2.15. Xét biến số ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	0	1	2	3
P	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Ta có $M_0(X) = 1$ hay $M_0(X) = 2$ và $1 < Me(X) < 2$.

Ví dụ 2.16. Cho biến số ngẫu nhiên liên tục X với hàm mật độ xác suất

$$f(x) = \begin{cases} \frac{2}{9}(3x - x^2) & \text{khi } x \in [0,3], \\ 0 & \text{khi } x \notin [0,3]. \end{cases}$$

Tính mốt và trung vị của X.

Giải

+) Tìm mốt của X

$$f(x) = \frac{2}{9}(3x - x^2)$$

Tập xác định: $D = [0,3]$

Ta có: $f'(x) = \frac{2}{9}(3 - 2x)$

Cho $f'(x) = 0 \Leftrightarrow x = \frac{3}{2} \in D$

Ta lại có: $f(0)=f(3)=0;\ f\left(\frac{3}{2}\right)=\frac{1}{2}$

Hàm mật độ đạt giá trị lớn nhất tại $x=\frac{3}{2}$. Vậy $M_0(X)=\frac{3}{2}$.

+) Tìm trung vị của X

$$f(x)=\frac{2}{9}\left(3x-x^2\right)$$

Tập xác định: $D=[0,3]$

Gọi $Me(X)=x_0\in[0,3]$

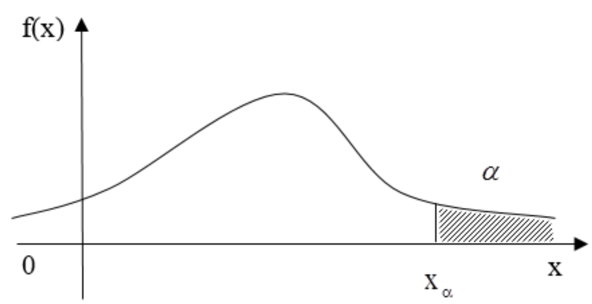
$$P(X\le x_0)=0,5\Leftrightarrow\int_{-\infty}^{x_0}f(x)dx=0,5\Leftrightarrow\frac{2}{9}\int_{-\infty}^{x_0}\left(3x-x^2\right)dx=0,5$$

$$\Leftrightarrow-\frac{2}{27}x_0^3+\frac{1}{3}x_0^2=\frac{1}{2}\Leftrightarrow\left[\begin{array}{l}x_0=\dfrac{3+3\sqrt{3}}{2}\notin[0,3]\\ x_0=\dfrac{3}{2}\in[0,3]\\ x_0=\dfrac{3-3\sqrt{3}}{2}\notin[0,3]\end{array}\right.$$

Vậy $Me(X)=\frac{3}{2}$.

2.3.8. Giá trị tới hạn

Giá trị tới hạn mức α của một biến ngẫu nhiên liên tục X, ký hiệu x_α là giá trị của X thỏa mãn: $P(X>x_\alpha)=\alpha$.



2.3.9. Hệ số đối xứng và hệ số nhọn

Người ta còn có một số tham số liên quan đến hình dáng của hàm mật độ như sau: Với X là biến số ngẫu nhiên với trung bình μ_X và phương sai σ_X^2, giá trị

$$\gamma_1(X) = \frac{E\left[(x-\mu_X)^3\right]}{\sigma_X^3}$$

được gọi là *hệ số đối xứng* của X và giá trị

$$\gamma_2(X) = \frac{E\left[(x-\mu_X)^4\right]}{\sigma_X^4}$$

được gọi là *hệ số nhọn* của X.

Nếu $\gamma_1(X)=0$ thì phân phối của X là đối xứng; lệch phải khi $\gamma_1(X)>0$ và lệch trái khi $\gamma_1(X)<0$. Ngoài ra giá trị $\gamma_2(X)$ càng lớn thì phân phối của X càng nhọn.

Ví dụ 2.17. Đo đường kính (X) một chi tiết máy (đơn vị mm). Ta có các số liệu : 201; 203; 209; 204; 202; 206; 200; 207; 207. Tính hệ số đối xứng và hệ số nhọn.

$$\mu_X = \frac{1}{9}(201+203+209+\cdots+207+207) = 204{,}3333,$$

$$\sigma_X^2 = \frac{1}{9}\left((201-204)^2+(203-204)^2+\cdots+(207-204)^2\right) = 8{,}4444,$$

Hệ số đối xứng

$$\gamma_1(X) = \frac{\frac{1}{9}\left((201-204)^3+(203-204)^3+\cdots+(207-204)^3\right)}{(9{,}5)^{3/2}} = 0{,}3939.$$

Hệ số nhọn

$$\gamma_2(X) = \frac{\frac{1}{9}\left((201-204)^4+(203-204)^4+\cdots+(207-204)^4\right)}{(9{,}5)^2} = 1{,}8029.$$

2.4. Một số quy luật phân phối xác suất quan trọng

2.4.1. Phân phối nhị thức $B(n;p)$

2.4.1.1. Định nghĩa. Biến ngẫu nhiên rời rạc X được gọi là có phân phối nhị thức, ký hiệu $X \sim B(n;p)$ nếu hàm xác suất của X có dạng sau

$$f(x) = \begin{cases} C_n^x p^x (1-p)^{n-x} & \text{khi } x = 0,1,2,...,n \\ 0 & \text{khi } x \neq 0,1,2,...,n \end{cases}$$

Công thức xác suất

$$P(X=k) = C_n^k p^k (1-p)^{n-k}, \text{ với } k = 0,1,2,...,n.$$

Ví dụ 2.18. Trong một vùng dân cư có 70% gia đình có máy giặt, chọn ngẫu nhiên 12 gia đình. Tính xác suất

a) có đúng 5 gia đình có máy giặt.

b) có ít nhất 2 gia đình có máy giặt.

Giải

Gọi X là số gia đình có máy giặt trong số 12 gia đình này, $X \sim B(12;0,7)$. Ta có

$$P(X=k)=C_{12}^{k}(0,7)^{k}(1-0,7)^{12-k}=C_{12}^{k}(0,7)^{k}(0,3)^{12-k}$$

a) Xác suất để nhận được đúng 5 gia đình có máy giặt là

$$P(X=5)=C_{12}^{5}(0,7)^{5}(0,3)^{12-5}=0,0291.$$

b) Xác suất để có ít nhất 2 gia đình có máy giặt là

$$P(X\geq 2)=1-P(X<2)=1-P(X=0)-P(X=1)$$
$$=1-C_{12}^{0}(0,7)^{0}(0,3)^{12}-C_{12}^{1}(0,7)^{1}(0,3)^{11}\approx 1.$$

2.4.1.2. Mệnh đề. Cho $X \sim B(n;p)$, ta có

i) Trung bình: $\mu_X = np.$

ii) Phương sai: $\sigma_X^2 = npq,$ với $q=1-p.$

iii) Giá trị tin chắc nhất : $M_0(X)=k_0,$ với k_0 là số nguyên thỏa bất phương trình $np-q\leq k_0\leq np-q+1.$

Ví dụ 2.19. Một nhân viên tiếp thị bán hàng ở 5 chỗ khác nhau trong ngày. Xác suất bán được hàng ở mỗi nơi đều là 0,4.

a) Tìm xác suất để nhân viên này bán được hàng trong ngày.

b) Mỗi năm nhân viên này đi bán hàng 330 ngày. Gọi Y là số ngày bán được hàng trong năm. Tìm giá trị tin chắc nhất của Y.

Giải

Gọi X là tổng số nơi bán được hàng trong ngày, $X \sim B(5;0,4)$. Ta có

$$P(X=k)=C_{5}^{k}(0,4)^{k}(1-0,4)^{5-k}=C_{5}^{k}(0,4)^{k}(0,6)^{5-k}$$

a) Xác suất để bán được hàng trong ngày là

$$P(X\geq 1)=1-P(X<1)=1-P(X=0)$$
$$=1-C_{5}^{0}(0,4)^{0}(0,6)^{5}=0,92224.$$

b) Tìm số ngày bán được hàng nhiều khả năng nhất trong một năm.

Ta có $Y \sim B(330;0,92224)$ và do đó, số ngày bán được hàng nhiều khả năng nhất là $k_0 = M_0(Y),$ với $k_0 \in \mathbb{N}$ thỏa

$$330 \cdot 0,92224-(1-0,92224) \le k_0 \le 330 \cdot 0,92224-(1-0,92224)+1$$

$$\Leftrightarrow 304,26 \le k_0 \le 305,26 \Leftrightarrow k_0 = 305.$$

Vậy số ngày bán được hàng có nhiều khả năng trong một năm là 305.

2.4.2. Phân phối siêu bội $H(N,K,n)$

2.4.2.1. Định nghĩa. Biến ngẫu nhiên rời rạc X được gọi là có phân phối siêu bội, ký hiệu $X \sim H(N,K,n)$ nếu hàm xác suất của X có dạng sau

$$f(x)=\begin{cases} \dfrac{C_K^x C_{N-K}^{n-x}}{C_N^n} & \text{khi } x \in [\max\{0,n-N+K\},\min\{n,K\}] \\ 0 & \text{khi } x \notin [\max\{0,n-N+K\},\min\{n,K\}] \end{cases}$$

Công thức xác suất

$$P(X=k)=\frac{C_K^k C_{N-K}^{n-k}}{C_N^n}, \text{ với } \max\{0,n-N+K\} \le k \le \min\{n,K\}.$$

2.4.2.2. Mệnh đề. Cho $H(N,K,n)$, ta có

i) Trung bình: $\mu_X = np$,

ii) Phương sai: $\sigma_X^2 = np(1-p)\left(\dfrac{N-n}{N-1}\right)$.

2.4.2.3. Lưu ý : Nếu $X \sim H(N,K,n)$, trong đó $n << N$ thì X được xem như có phân phối nhị thức $X \sim B(n;p)$, với $p=\dfrac{K}{N}$.

Ví dụ 2.20. Từ một hộp đựng 15 quả cam trong đó có 5 quả hư, lấy ra 3 quả. Gọi X là số quả hư trong 3 quả lấy ra.

a) Tính xác suất để cả 3 quả đều hư.

b) Tính trung bình và phương sai của X.

Giải

Ta có $X \sim H(15,5,3)$. Công thức tính xác suất

$$P(X=k)=\frac{C_5^k C_{15-5}^{3-k}}{C_{15}^3}$$

a) Xác suất để cả 3 quả đều hư

$$P(X=3)=\frac{C_5^3 C_{10}^3}{C_{15}^3}=0,021978.$$

b) Trung bình (kỳ vọng) và phương sai của X

$$\mu_X = np = 3\frac{5}{15} = 1;\ \sigma_X^2 = 3\frac{5}{15}\left(1-\frac{5}{15}\right)\left(\frac{15-3}{15-1}\right) = \frac{4}{7}.$$

Ví dụ 2.21. Có 8000 sản phẩm trong đó có 2000 sản phẩm không đạt tiêu chuẩn kỹ thuật. Lấy ngẫu nhiên (không hoàn lại) 10 sản phẩm. Tính xác suất để trong 10 sản phẩm lấy ra có 2 sản phẩm không đạt tiêu chuẩn.

Giải

Cách 1. Tính trực tiếp theo phân phối của nó

Ta có $N = 8000,\ K = 2000,\ n = 10$

Gọi X số sản phẩm trong đạt tiêu chuẩn trong 10 sản phẩm lấy ra, $X \sim H(8000, 2000, 10)$

Công thức tính xác suất:

$$P(X = k) = \frac{C_{2000}^{k} C_{6000}^{10-k}}{C_{8000}^{10}}$$

Xác suất lấy hai sản phẩm không đạt tiêu chuẩn

$$P(X = 2) = \frac{C_{2000}^{2} C_{6000}^{8}}{C_{8000}^{10}} = 0,281697.$$

Cách 2. Tính xấp xỉ (tính gần đúng)

Do $n = 10 << N = 8000$; $p = \frac{K}{N} = 0,25$

$$X \sim H(8000, 2000, 10) \equiv B(10; 0,25)$$

Xác suất lấy hai sản phẩm không đạt tiêu chuẩn

$$P(X = 2) = C_{10}^{2}(0,25)^2(0,75)^8 = 0,28157.$$

2.4.3. Phân phối Poisson $P(\mu)$

2.4.3.1. Định nghĩa. Biến ngẫu nhiên rời rạc X được gọi là có phân phối poisson, ký hiệu $X \sim P(\mu)$ nếu hàm xác suất của X có dạng sau

$$f(x) = \begin{cases} e^{-\mu}\dfrac{\mu^x}{x!} & \text{khi } x = 0,1,2,...,n,... \\ 0 & \text{khi } x \neq 0,1,2,...,n,... \end{cases}$$

Công thức xác suất

$$P(X = k) = e^{-\mu}\frac{\mu^k}{k!},\ \text{với } k = 0,1,2,...,n.$$

2.4.3.2. Mệnh đề. Cho $X \sim P(\mu)$, ta có

i) Trung bình: $\mu_X = \mu$,

ii) Phương sai: $\sigma_X^2 = \mu$,

iii) Độ lệch chuẩn: $\sigma_X = \sqrt{\mu}$.

2.4.3.3. Chú ý: Nếu $X \sim B(n,p)$, trong đó p đủ nhỏ và n đủ lớn thì X được xem như có phân phối Poisson $X \sim P(\mu)$, với $\mu = np$.

Bằng cách viết

$$P(X = k) = C_n^k p^k (1-p)^{n-k} = \frac{n(n-1)...(n-k+1)}{k!} p^k (1-p)^{n-k}$$
$$= \frac{1}{k!} \frac{n(n-1)...(n-k+1)}{n^k} (np)^k (1-p)^{n-k}$$

và với $np = \mu$ không đổi, khi $n \to \infty$ ta có $p \to 0$ và

$$\lim_{n\to\infty} \frac{n(n-1)...(n-k+1)}{n^k} = 1,$$

$$\lim_{n\to\infty} (1-p)^{n-k} = \lim_{p\to 0} (1-p)^{\frac{\mu}{p}-k} = \lim_{p\to 0} \left[(1-p)^{\frac{1}{-p}} \right]^{-\mu+kp} = e^{-\mu}.$$

Từ đó, suy ra $C_n^k p^k (1-p)^{n-k} \approx \frac{e^{-\mu}\mu^k}{k!}$ khi n khá lớn.

Trong ứng dụng, khi $X \sim B(n;p)$, trong đó $n > 50$, $p < 0{,}01$ và $np < 5$ thì ta có thể dùng xấp xỉ $X \sim P(np)$.

Ví dụ 2.22. Một máy dệt có 4000 ống sợi. Xác suất để mỗi ống sợi bị đứt trong 1 phút là 0,0005. Tính xác suất để trong 1 phút

a) có 3 ống sợi bị đứt,

b) có ít nhất 2 ống sợi bị đứt.3

Giải

Cách 1. Tính trực tiếp theo luật phân phối của nó

Ta có $n = 4000$; $p = 0{,}0005$

Gọi X là số ống sợi bị đứt trong 1 phút (trong 4000 ống sợi), $X \sim B(4000; 0{,}0005)$

Công thức tính xác suất:

$$P(X = k) = C_{4000}^k (0{,}0005)^k (0{,}9995)^{4000-k}$$

a) có 3 ống sợi bị đứt,

$$P(X=3)=C_{4000}^{3}(0,0005)^{3}(0,9995)^{3997}=0,1804822$$

b) có ít nhất 2 ống sợi bị đứt.

$$\begin{aligned}P(X\ge 2)&=1-P(X<2)=1-P(X=0)-P(X=1)\\&=1-C_{4000}^{0}(0,0005)^{0}(0,9995)^{4000}-C_{4000}^{1}(0,0005)^{1}(0,9995)^{3999}\\&=0,594062.\end{aligned}$$

Cách 2. Tính xấp xỉ (tính gần đúng)

Ta có trung bình : $\mu=np=4000\cdot 0,0005=2<5$ nên phân phối nhị thức được xấp xỉ bằng phân phối Poisson như sau:

$$X\sim B(4000;0,0005)\equiv P(2)$$

Công thức tính xác suất:

$$P(X=k)=e^{-2}\frac{2^{k}}{k!}$$

a) có 3 ống sợi bị đứt,

$$P(X=3)=e^{-2}\frac{2^{3}}{3!}=0,180447.$$

b) có ít nhất 2 ống sợi bị đứt.

$$\begin{aligned}P(X\ge 2)&=1-P(X<2)=1-P(X=0)-P(X=1)\\&=1-e^{-2}\frac{2^{0}}{0!}-e^{-2}\frac{2^{1}}{1!}=0,59399.\end{aligned}$$

2.4.4. Phân phối đều $U[a,b]$

2.4.4.1. Định nghĩa. Biến ngẫu nhiên liên tục X được gọi là có phân phối đều, ký hiệu $X\sim U[a,b]$ nếu hàm mật độ của X có dạng sau

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

Công thức tính xác suất

$$P(\alpha\le X\le\beta)=\frac{\beta-\alpha}{b-a}$$

2.4.4.2. Mệnh đề. Cho $X\sim U[a,b]$, ta có

i) Trung bình: $\mu_X = \dfrac{a+b}{2}$,

ii) Phương sai: $\sigma_X^2 = \dfrac{(b-a)^2}{12}$.

Ví dụ 2.23. Một loại sản phẩm do một nhà máy sản xuất được đóng thành từng hộp. Trọng lượng của hộp là biến ngẫu nhiên phân phối đều trong khoảng $\left[1,9;2,1\right]$ (kg). Tính trọng lượng trung bình của một hộp và tỷ lệ hộp có trọng lượng từ 1,95 kg trở lên.

Giải

Gọi X là trọng lượng của mỗi hộp sản phẩm, $X \sim U\left[1,9;2,1\right]$

Trọng lượng trung bình của một hộp chính là

$$E(X) = \frac{a+b}{2} = \frac{1,9+2,1}{2} = 2\,(kg)$$

Tỷ lệ hộp có trọng lượng từ 1,95 kg trở lên là:

$$P(X \ge 1,95) = P(1,95 \le X \le 2,1) = \frac{2,1-1,95}{2,1-1,9} = 0,75 = 75\%.$$

2.4.5. Phân phối mũ

2.4.5.1. Định nghĩa. Biến ngẫu nhiên liên tục X được gọi là có phân phối mũ, ký hiệu $X \sim \text{Exp}(\lambda)$ nếu hàm mật độ của X có dạng sau

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{khi} \quad x > 0, \\ 0, & \text{khi} \quad x \le 0. \end{cases}$$

2.4.5.2. Mệnh đề. Cho $X \sim \text{Exp}(\lambda)$, ta có

i) Trung bình: $\mu_X = \dfrac{1}{\lambda}$,

ii) Phương sai: $\sigma_X^2 = \dfrac{1}{\lambda^2}$.

Ví dụ 2.24. Tuổi thọ (tính theo giờ) của một trò chơi điện tử bấm tay là một biến số ngẫu nhiên liên tục có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} ke^{-\frac{x}{100}} & \text{khi } x \ge 0 \\ 0 & \text{khi } x < 0 \end{cases}$$

a) Tìm hằng số k.

b) Tính xác suất tuổi thọ của trò chơi này nằm trong khoảng từ 50 đến 150 giờ.

c) Tính xác suất tuổi thọ của trò chơi này ít hơn 100 giờ.

Giải

a) Tìm hằng số k.

Ta có

+) Vì $f(x)$ là hàm mật độ nên $f(x) \ge 0 \Rightarrow k > 0$

+) $\int_{-\infty}^{+\infty} f(x)dx = 1 \Leftrightarrow \int_{0}^{+\infty} ke^{-\frac{x}{100}}dx = 1$

$$\Leftrightarrow k\int_{0}^{+\infty} e^{-\frac{x}{100}}dx = 1 \Leftrightarrow \left(-100k\,e^{-\frac{x}{100}}\right)\Bigg|_{0}^{+\infty} = 1 \Leftrightarrow 100k = 1 \Leftrightarrow k = \frac{1}{100}$$

b) Tính xác suất tuổi thọ của trò chơi này nằm trong khoảng từ 50 đến 150 giờ.

Ta có

$$P(50 \le X \le 150) = \int_{50}^{150} f(x)dx = \frac{1}{100}\int_{50}^{150} e^{-\frac{x}{100}}dx$$

$$= -e^{-\frac{x}{100}}\Bigg|_{50}^{150} = e^{-\frac{1}{2}} - e^{-\frac{3}{2}} \approx 0{,}3834.$$

c) Tính xác suất tuổi thọ của trò chơi này ít hơn 100 giờ.

Ta có

$$P(X < 100) = \int_{-\infty}^{100} f(x)dx = \frac{1}{100}\int_{0}^{100} e^{-\frac{x}{100}}dx$$

$$= -e^{-\frac{x}{100}}\Bigg|_{0}^{100} = 1 - e^{-1} \approx 0{,}632.$$

2.4.6. Phân phối chuẩn tắc $N(0,1)$

2.4.6.1. Định nghĩa. Biến ngẫu nhiên liên tục X được gọi là có phân phối chuẩn tắc, ký hiệu $X \sim N(0,1)$ nếu hàm mật độ của X có dạng sau

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \quad -\infty < x < +\infty,$$

Với $a, b \in \mathbb{R}$, $a \le b$, ta có

$$P(a \le X \le b) = \frac{1}{\sqrt{2\pi}}\int_{a}^{b} e^{-\frac{x^2}{2}}dx = \phi_0(b) - \phi_0(a).$$

Trong đó $\phi_0(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt.$

2.4.6.2. Mệnh đề. Cho $X \sim N(0,1)$, ta có

i) Trung bình: $\mu_X = 0$,

ii) Phương sai: $\sigma_X^2 = 1$,

iii) Độ lệch chuẩn: $\sigma_X = 1$.

2.4.7. Phân phối chuẩn $N(\mu,\sigma^2)$

2.4.7.1. Định nghĩa. Biến ngẫu nhiên liên tục X được gọi là có phân phối chuẩn, ký hiệu $X \sim N(\mu,\sigma^2)$ nếu hàm mật độ của X có dạng sau

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty,$$

Với $a, b \in \mathbb{R}$, $a \le b$, ta có

$$P(a \le X \le b) = \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

2.4.7.2. Mệnh đề. Cho $X \sim N(\mu,\sigma^2)$, ta có

i) Trung bình: $\mu_X = \mu$,

ii) Phương sai: $\sigma_X^2 = \sigma^2$,

iii) Độ lệch chuẩn: $\sigma_X = \sigma$.

2.4.7.3. Chú ý

i) Nếu $X \sim N(\mu,\sigma^2)$ thì đặt $Y = \frac{X-\mu}{\sigma}$, ta có $Y \sim N(0,1)$. Do đó

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Y \le \frac{b-\mu}{\sigma}\right) = \phi_0\left(\frac{b-\mu}{\sigma}\right) - \phi_0\left(\frac{a-\mu}{\sigma}\right)$$

ii) Phân phối chuẩn dùng để khảo sát các hiện tượng bình thường. Cụ thể, nếu $X \sim B(n;p)$ với tích np lớn thì ta xấp xỉ phân phối nhị thức $B(n;p)$ bằng phân phối chuẩn $N(\mu;\sigma^2)$, với $\mu = np$, $\sigma^2 = npq$. Ta có $X \sim B(n;p) \equiv N(np, npq)$

+) $P(X=k) \approx \dfrac{1}{\sqrt{npq}\sqrt{2\pi}} e^{-\frac{(k-np)^2}{2npq}}$,

+) $P(a \le X \le b) \approx \phi_0\left(\dfrac{b-np}{\sqrt{npq}}\right) - \phi_0\left(\dfrac{a-np}{\sqrt{npq}}\right)$.

Ví dụ 2.25. Trọng lượng X (tính bằng gam) một loại trái cây có phân phối chuẩn $N(\mu;\sigma^2)$, với $\mu = 500(\text{gam})$ và $\sigma^2 = 16(\text{gam}^2)$. Trái cây thu hoạch được phân loại theo trọng lượng như sau :

a) loại 1 : trên 505 gam,

b) loại 2 : từ 495 đến 505 gam,

c) loại 3 : dưới 495 gam.

Tính tỷ lệ mỗi loại.

Giải

Gọi X là trọng lượng trái cây thì $X \sim N(\mu;\sigma^2) = N(500;4^2)$.

a) Tỷ lệ trái cây loại 1 là

$$P(X > 505) = \phi_0(+\infty) - \phi_0\left(\frac{505-500}{4}\right) = 0,5 - \phi_0(1,25) = 0,10565.$$

b) Tỷ lệ trái cây loại 2 là

$$P(495 \le X \le 505) = \phi_0\left(\frac{505-500}{4}\right) - \phi_0\left(\frac{495-500}{4}\right) = 0,7887.$$

c) Và tỷ lệ của loại 3 là

$$P(X < 495) = P\left(\frac{X-500}{4} < \frac{495-500}{4}\right) = \phi_0\left(\frac{495-500}{4}\right) - \phi_0(-\infty)$$
$$= \phi_0(-1,25) + \phi_0(+\infty) = -\phi_0(1,25) + 0,5 = 0,10565.$$

Vậy, trái cây thu hoạch được có khoảng 11% loại 1, 78% loại 2 và 11% loại 3.

Ví dụ 2.26. Thời gian chạy 1000m của mỗi sinh viên là một biến ngẫu nhiên phân phối chuẩn có trung bình là 80 giây, độ lệch chuẩn là 10 giây. Một sinh viên phải chạy tối đa là bao nhiêu thời gian để đứng trong 10% số người đứng đầu.

Giải

Gọi X là thời gian chạy hết 1000m của mỗi sinh viên, $X \sim N(\mu = 80, \sigma^2 = 10^2)$.

t là thời gian chạy tối đa để đứng trong 10% số người đứng đầu, ta có:

$$P(X \le t) = 0,1 \Leftrightarrow P(0 \le X \le t) = 0,1$$

$$\Leftrightarrow \phi_0\left(\frac{t-80}{10}\right) - \phi_0\left(\frac{0-80}{10}\right) = 0,1 \Leftrightarrow \phi_0\left(\frac{t-80}{10}\right) + 0,5 = 0,1$$

$$\Leftrightarrow \phi_0\left(\frac{80-t}{10}\right) = 0,4 = \phi_0(1,28) \Leftrightarrow t = 67,2.$$

Vậy thời gian tối đa của sinh viên hoàn thành 1000m để đứng trong 10% số người đứng đầu.

Ví dụ 2.27. Xác suất để một sản phẩm không được kiểm tra chất lượng sau khi sản xuất là 0,2. Tìm xác suất để trong 400 sản phẩm sản xuất ra có:

a) 80 sản phẩm không được kiểm tra chất lượng.

b) Có từ 70 đến 100 sản phẩm không được kiểm tra chất lượng.

Giải

Gọi X là số sản phẩm không được kiểm tra chất lượng, $X \sim B(400;0,2)$

Ta có: $\mu = np = 400 \cdot 0,2 = 80; \sigma^2 = npq = 400 \cdot 0,2 \cdot 0,8 = 8^2$

nên có thể coi $X \sim N(80;8^2)$

a) $P(X = 80) \approx \dfrac{1}{\sqrt{400.0,2.0,8}} \dfrac{1}{\sqrt{2\pi}} e^{-\frac{(80-80)^2}{2\cdot 64}} = \dfrac{1}{8\sqrt{2\pi}} = 0,049868.$

b) $P(70 \le X \le 100) \approx \phi_0\left(\dfrac{100-80}{8}\right) - \phi_0\left(\dfrac{70-80}{8}\right)$

$$= \phi_0(2,5) - \phi_0(-1,25) = \phi_0(2,5) + \phi_0(1,25) = 0,8882.$$

2.4.8. Phân phối Gamma và phân phối Chi bình phương

Định nghĩa: Hàm Gamma $\Gamma : (0,\infty) \to \mathbb{R}$

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

2.4.8.1. Định nghĩa. Biến ngẫu nhiên liên tục X được gọi là có phân phối Gamma, ký hiệu $X \sim \Gamma(\alpha,\beta)$, với $\alpha,\beta > 0$, nếu hàm mật độ của X có dạng sau

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} & \text{khi } x > 0 \\ 0 & \text{khi } x \le 0 \end{cases},$$

2.4.8.2. Mệnh đề. Cho $X \sim \Gamma(\alpha,\beta)$, ta có

i) Trung bình: $\mu_X = \alpha\beta$,

ii) Phương sai: $\sigma_X^2 = \alpha\beta^2$.

2.4.8.3. Định nghĩa. Biến ngẫu nhiên liên tục X được gọi là có phân phối Chi bình phương, ký hiệu $X \sim \chi^2(r)$, nếu hàm mật độ của X có dạng sau

$$f(x) = \begin{cases} \dfrac{1}{\Gamma\left(\dfrac{r}{2}\right)2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}} & \text{khi } x > 0 \\ 0 & \text{khi } x \le 0 \end{cases},$$

nghĩa là $X \sim \Gamma\left(\dfrac{r}{2}, 2\right)$

2.4.8.4. Mệnh đề. Cho $X \sim \chi^2(r)$, ta có

i) Trung bình: $\mu_X = r$,

ii) Phương sai: $\sigma_X^2 = 2r$.

2.4.9. Phân phối Student: $St(n)$

2.4.9.1. Định nghĩa: Biến ngẫu nhiên liên tục T được gọi là có phân phối Student, ký hiệu $T \sim St(n)$, nếu hàm mật độ của T có dạng sau

$$f(t) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{n\pi} \cdot \Gamma\left(\dfrac{n}{2}\right) \cdot \left(1 + \dfrac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

2.4.9.2. Mệnh đề. Cho X là biến ngẫu nhiên liên tục có phân phối Gauss, $X \sim N(0,1)$; Y là biến ngẫu nhiên liên tục có phân phối Chi bình phương với n bậc tự do, $Y \sim \chi^2(n)$ và X, Y là hai biến độc lập.

Đặt $T = \dfrac{X}{\sqrt{Y/n}}$ thì T có phân phối Student với n bậc tự do, $T \sim St(n)$.

2.4.9.3. Mệnh đề. Cho $T \sim St(n)$, ta có

i) Trung bình: $\mu_T = 0$,

ii) Phương sai: $\sigma_T^2 = \dfrac{n}{n-2}$.

2.4.9.4. Chú ý: Nếu $X \sim St(n)$, với $n \ge 30$, thì $X \sim N(0,1)$.

2.4.10. Phân phối Fisher: $F(n,m)$

2.4.10.1. Định nghĩa. Biến ngẫu nhiên liên tục F có phân phối Fisher, ký hiệu $F \sim F(n,m)$ nếu hàm mật độ của F có dạng như sau

$$f(x) = \begin{cases} \dfrac{\sqrt{n^n \cdot m^m} \cdot \Gamma\left(\dfrac{n+m}{2}\right) x^{\frac{n-m}{2}}}{\Gamma\left(\dfrac{n}{2}\right)(m+nx)^{\frac{n+m}{2}}} & \text{khi } x > 0, \\ 0 & \text{khi } x \le 0. \end{cases}$$

với n, m là hai bậc tự do.

2.4.10.2. Mệnh đề. Cho X, Y là hai biến ngẫu nhiên liên tục có phân phối Chi bình phương, $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ và X, Y là hai biến độc lập.

Đặt $F = \dfrac{X/n}{Y/m}$ thì F có phân phối Fisher với n, m bậc tự do, $F \sim F(n,m)$.

2.4.10.3. Mệnh đề. Cho $F \sim F(n,m)$, ta có

i) Trung bình: $\mu_F = \dfrac{n}{n-2}$, $n > 2$,

ii) Phương sai: $\sigma_F^2 = \dfrac{2m^2(n+m^2-2)}{n(m-2)^2(m-4)}$, $m > 4$.

2.5. Tóm tắt chương 2

A. Đại lượng ngẫu nhiên rời rạc

1. Bảng phân phối xác suất

X	x_1	x_2	...	x_n	...
P	p_1	p_2	...	p_n	...

với $p_i = P(X = x_i)$, được gọi là *bảng phân phối xác suất* của X.

Tính chất: $\sum_i p_i = 1$.

2. Hàm xác suất của biến ngẫu nhiên X: Hàm số $f : \mathbb{R} \to \mathbb{R}$ xác định bởi

$$f(x) = \begin{cases} p_i & \text{khi} \quad x = x_i, \\ 0 & \text{khi} \quad x \ne x_i, \forall i. \end{cases}$$

Tính chất : $\forall x \in \mathbb{R}, f(x) \ge 0$, và $\sum_x f(x) = 1$.

3. Hàm phân phối xác suất: Cho $f:\mathbb{R}\to\mathbb{R}$ là hàm xác suất của biến ngẫu nhiên rời rạc X, hàm số $F:\mathbb{R}\to\mathbb{R}$, được xác định bởi

$$F(x)=P(X\le x)=\sum_{x\ge x_i} f(x_i),$$

B. Đại lượng ngẫu nhiên liên tục

1. Hàm mật độ (xác suất): Hàm số $f:\mathbb{R}\to\mathbb{R}$ được gọi là *hàm mật* của biến số ngẫu nhiên liên tục X nếu $P(a\le X\le b)=\int_a^b f(x)dx$

với mọi $a,b\in\mathbb{R}$, $a\le b$. Ta có : $\forall x\in\mathbb{R}$, $f(x)\ge 0$, và $\int_{-\infty}^{+\infty} f(x)dx=1.$

2. Hàm phân phối (tích lũy): Cho $f:\mathbb{R}\to\mathbb{R}$ là hàm mật độ của biến ngẫu nhiên liên tục X, hàm số $F:\mathbb{R}\to\mathbb{R}$, được gọi là *hàm phân phối* của biến số ngẫu nhiên liên tục X nếu $F(x)=P(X\le x)=\int_{-\infty}^{x} f(t)dt$ với mọi $x\in\mathbb{R}$.

C. Trung bình và phương sai: Cho X là biến ngẫu nhiên với hàm xác suất (hàm mật độ xác suất) $f(x)$,

1. Trung bình:

$E(X)=\sum_i x_i f(x_i)$ khi X là biến số ngẫu nhiên rời rạc, và

$E(X)=\int_{-\infty}^{+\infty} xf(x)dx$ khi X là biến ngẫu nhiên liên tục.

2. Phương sai

$\sigma_X^2=\sum_i (x_i-\mu_X)^2 f(x_i)$ khi X là biến số ngẫu nhiên rời rạc, và

$\sigma_X^2=\int_{-\infty}^{+\infty} (x-\mu_X)^2 f(x)dx$ khi X là biến ngẫu nhiên liên tục.

3. Độ lệch chuẩn:

$\sigma_X=Se(X)$ gọi là độ lệch chuẩn của X.

4. Mệnh đề. Cho X là biến số ngẫu nhiên với trung bình $E(X)$. Ta có

$$\mathrm{var}(X)=E(X^2)-[E(X)]^2$$

D. Các quy luật thường gặp

1. Phân phối nhị thức: $X \sim B(n;p)$

i) Công thức xác suất

$$P\left(X=k\right)=C_n^k p^k (1-p)^{n-k}, \text{ với } k=0,1,2,...,n.$$

ii) Trung bình: $\mu_X = np$,

iii) Phương sai: $\sigma_X^2 = np(1-p)$,

iv) Giá trị tin chắc nhất : $M_0\left(X\right)=k_0$, với k_0 là số nguyên thỏa bất phương trình $np-q \le k_0 \le np-q+1$.

2. Phân phối siêu bội: $X \sim H(N,K,n)$

i) $P\left(X=k\right)=\dfrac{C_K^k C_{N-K}^{n-k}}{C_N^n}$, với $\max\{0,n-N+K\} \le k \le \min\{n,K\}$.

ii) Trung bình: $\mu_X = np$,

iii) Phương sai: $\sigma_X^2 = np(1-p)\left(\dfrac{N-n}{N-1}\right)$.

3. Phân phối Poisson : $X \sim P(\mu)$

i) Công thức xác suất

$$P\left(X=k\right)=e^{-\mu}\frac{\mu^k}{k!}, \text{ với } k=0,1,2,...,n.$$

ii) Trung bình: $\mu_X = \mu$,

iii) Phương sai: $\sigma_X^2 = \mu$.

4. Phân phối chuẩn tắc: $X \sim N\left(0,1\right)$

i) $P\left(a \le X \le b\right)=\phi_0\left(b\right)-\phi_0\left(a\right)$, với $\phi_0(x)=\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_0^x e^{-\frac{t^2}{2}}dt.$

ii) Trung bình: $\mu_X = 0$,

iii) Phương sai: $\sigma_X^2 = 1$,

5. Phân phối chuẩn: $X \sim N\left(\mu,\sigma^2\right)$

i) $P\left(a \le X \le b\right)=\phi_0\left(\dfrac{b-\mu}{\sigma}\right)-\phi_0\left(\dfrac{a-\mu}{\sigma}\right)$,

i) Trung bình: $\mu_X = \mu$,

ii) Phương sai: $\sigma_X^2 = \sigma^2$.

2.6. Bài tập

Bài số 1. Cho X là một biến số ngẫu nhiên có phân phối xác suất như sau

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

a) Xác định a.

b) Tính $P[X \geq 5]$, $P[X < 3]$.

c) Tính k nhỏ nhất sao cho $P[X \leq k] \geq 0,5$.

Đáp số: a) $a = 0,1$*; b)* $P[X \geq 5] = 0,2$; $P[X < 3] = 0,3$*; c)* $k = 3$.

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

a) A bao nhiêu thì người chơi về lâu về dài huề vốn (gọi là trò chơi công bằng),

b) A bao nhiêu thì trung bình mỗi lần người chơi mất 1 ngàn đồng.

Đáp số: a) $A = 1$; *b)* $A = 2$.

Bài số 3. Một nhà đầu tư có 3 dự án. Gọi X_i $(i = 1,2,3)$ là số tiền thu được khi thực hiện dự án thứ i (giá trị âm chỉ số tiền bị thua lỗ). X_i là biến số ngẫu nhiên. Qua nghiên cứu, giả sử có số liệu như sau : (Đơn vị tính : 100 triệu đồng)

X_1	-20	30	60
P	0,3	0,2	0,5

X_2	-20	-10	100
P	0,4	0,2	0,4

X_3	-25	-30	80
P	0,2	0,3	0,5

Theo anh (chị), ta nên chọn dự án nào ?

Đáp số: Nên chọn dự án 1.

Bài số 4. Có 3 xạ thủ cùng bắn vào một mục tiêu, mỗi người bắn 1 viên, trong cùng một số điều kiện nhất định. Xác suất để mỗi xạ thủ bắn trúng mục tiêu lần lượt là 0,6; 0,7; 0,9. Gọi X là số viên đạn trúng mục tiêu. Tính $E(X)$; $Var(X)$ và $Mod(X)$.

Đáp số: $EX = 2,2$; $Var(X) = 0,54$; $Mod(X) = 2$.

Bài số 5. Một phân xưởng có ba máy M_1, M_2, M_3. Trong một giờ, mỗi máy sản xuất được 10 sản phẩm. Số sản phẩm không đạt tiêu chuẩn trong 10 sản phẩm của M_1, M_2, M_3 lần lượt là 1, 2, 1. Lấy ngẫu nhiên 1 sản phẩm từ 10 sản phẩm do mỗi máy sản xuất. Gọi X là số sản phẩm không đạt tiêu chuẩn trong 3 sản phẩm được lấy ra.

a) Lập bảng phân phối xác suất của X.

b) Tìm E(X), Var(X), Mod(X).

c) Tính $P(X \le 1)$.

Đáp số: a)

X	0	1	2	3
P	0,648	0,306	0,044	0,002

b) $EX = 0,4$; $Var(X) = 0,34$; $Mod(X) = 0$; c) 0,954.

Bài số 6. Tỷ lệ khách hàng phản ứng tích cực đối với một chiến dịch quảng cáo là biến số ngẫu nhiên có bảng phân phối xác suất như sau :

X (%)	0	10	20	30	40	50
P	0,1	0,2	0,35	0,2	0,1	0,05

a) Tính tỷ lệ khách hàng bình quân phản ứng tích cực đối với chiến dịch quảng cáo đó.

b) Tìm xác suất để có trên 20% khách hàng phản ứng tích cực đối với chiến dịch quảng cáo.

Đáp số: a) 21,5%; b) 0,35.

Bài số 7. Qua theo dõi trong nhiều năm kết hợp với sự đánh giá của các chuyên gia tài chính thì lãi suất đầu tư vào một công ty là biến số ngẫu nhiên có bảng phân phối xác suất như sau:

X (%)	9	10	11	12	13	14	15
P	0,05	0,15	0,3	0,2	0,15	0,1	0,05

a) Tính xác suất để khi đầu tư vào công ty đó thì sẽ đạt được lãi suất ít nhất là 12%.

b) Tính lãi suất kỳ vọng khi đầu tư vào công ty đó.

c) Mức độ rủi ro khi đầu tư vào công ty đó có thể đánh giá bằng cách nào?

Đáp số a) 0,5; b) $EX = 11,75$ *; c)* $\sigma_X^2 = 2,2875$.

Bài số 8. Cho biến số ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

X	1	4	8
P	0,3	0,1	0,6

Tính $P\left(\left|X - E(X)\right| < 4\right)$.

Đáp số: 0,7.

Bài số 9. Lợi nhuận X thu được khi đầu tư vào một dự án có bảng phân phối xác suất như sau (đơn vị : tỷ đồng).

X	-2	-1	0	1	2	3
P	0,1	0,1	0,2	0,2	0,3	0,1

a) Tìm mức lợi nhuận có khả năng nhiều nhất khi đầu tư vào dự án đó.

b) Việc đầu vào dự án này có hiệu quả hay không? Tại sao?

c) Làm thế nào để đo được mức độ rủi ro của vụ đầu tư này? Hãy tìm mức độ rủi ro đó.

Đáp số: a) $\mathrm{Mod}(X) = 2$*; b)* $EX = 0{,}8$*;* $\sigma_X^2 = 2{,}16$.

Bài số 10. Tại một cửa hàng bán xe máy Honda người ta thống kê được số xe máy bán ra hàng tuần (X) với bảng phân phối xác suất như sau :

X	0	1	2	3	4	5	6	7	8	9	10	11
P	0,05	0,12	0,17	0,08	0,12	0,2	0,07	0,02	0,07	0,02	0,03	0,05

a) Tìm số xe trung bình bán được mỗi tuần.

b) Tìm phương sai và độ lệch chuẩn của số xe bán được mỗi tuần và giải thích ý nghĩa của kết quả nhận được.

Đáp số: a) 4,33; b) $\sigma_X^2 = 8{,}3411; \sigma_X = 2{,}89$.

Bài số 11. Sản phẩm nhà máy được đóng thành từng hộp, mỗi hộp có 10 sản phẩm. Gọi X là số sản phẩm loại một có trong hộp. Cho biết X có bảng phân phối xác suất như sau:

X	6	7
P	0,7	0,3

Khách hàng chọn cách kiểm tra để mua hàng như sau : Từ mỗi hộp lấy ngẫu nhiên ra 3 sản phẩm để kiểm tra, nếu thấy có ít nhất 2 sản phẩm loại một thì mua hộp đó. Lấy ngẫu nhiên 3 hộp để kiểm tra. Tính xác suất để có 2 hộp được mua.

Đáp số: 0,438.

Bài số 12. Thống kê số khách hàng trên một xe buýt tại một tuyến giao thông ta thu được các số liệu sau:

Số khách trên một chuyến	30	40	45	50	60
Tần suất tương ứng	0,15	0,2	0,3	0,25	0,1

a) Tìm kỳ vọng và phương sai của số khách hàng đi mỗi chuyến và giải thích ý nghĩa của kết quả nhận được.

b) Chi phí cho mỗi chuyến xe là 400 ngàn đồng không phụ thuộc vào số khách đi trên xe thì công ty xe buýt có thể thu lãi bình quân cho mỗi chuyến xe là 312 ngàn đồng. Công ty phải quy định giá vé là bao nhiêu?

Đáp số : a) $E(X) = 44,5$; $Var(X) = 67,25$. *b) 16 ngàn đồng.*

Bài số 13. Tuổi thọ của một loại bóng đèn nào đó là một biến số ngẫu nhiên liên tục X (đơn vị năm) với hàm mật độ như sau

$$f(x) = \begin{cases} kx^2(4-x) & \text{khi} \quad 0 \le x \le 4 \\ 0 & \text{khi} \quad x \notin [0,4] \end{cases}$$

a) Tìm k và vẽ đồ thị $f(x)$.

b) Tìm xác suất để bóng đèn hỏng trước khi nó được 1 năm tuổi.

Đáp số: a) $k = \frac{3}{64}$; *b) 0,0508.*

Bài số 14. Khối lượng của một con vịt 6 tháng tuổi là một biến số ngẫu nhiên X (đơn vị tính là Kg) có hàm mật độ

$$f(x) = \begin{cases} k(x^2-1) & \text{khi} \quad 1 \le x \le 3 \\ 0 & \text{khi} \quad x \notin [1,3] \end{cases}$$

a) Tìm k.

b) Với k tìm được, tính

(i) khối lượng trung bình của vịt 6 tháng tuổi,

(ii) tỷ lệ vịt chậm lớn, biết vịt 6 tháng tuổi chậm lớn là vịt có khối lượng nhỏ hơn 2Kg,

(iii) hàm phân phối xác suất của X.

Đáp số: a) $k = \frac{3}{20}$; *b)* $EX = 2,4$; $P(X < 2) = 0,2$

$$F(x) = \begin{cases} 0 & \text{khi } x \le 1 \\ \dfrac{x^3 - 3x + 2}{20} & \text{khi } 1 < x \le 3 \\ 1 & \text{khi } x > 3 \end{cases}$$

Bài số 15. Cho hàm mật độ của biến ngẫu nhiên X có dạng

$$f(x) = \begin{cases} a\cos x & \text{khi} \quad x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right], \\ 0 & \text{khi} \quad x \notin \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \end{cases}$$

a) Tìm a và xác định hàm phân phối xác suất của X.

b) Tính xác suất để X nhận giá trị trong khoảng $\left(\frac{\pi}{4}, \pi\right)$.

Đáp số: a) $a = 0,5;\ F(x) = \begin{cases} 0 & \text{khi } x \le -\frac{\pi}{2} \\ 0,5(\sin x + 1) & \text{khi } -\frac{\pi}{2} < x \le \frac{\pi}{2} \\ 1 & \text{khi } x > \frac{\pi}{2} \end{cases}$

b) 0,1465.

Bài số 16. Cho X là biến số ngẫu nhiên liên tục có hàm phân phối xác suất sau:

$$F(x) = \frac{1}{4} + \frac{1}{\pi}\arctan x$$

a) Tính $P(0 < X < 1)$.

b) Tìm hàm mật độ xác suất của X.

Đáp số: a) 0,25; b) $f(x) = \dfrac{1}{\pi\left(1 + x^2\right)}$.

Bài số 17. Cho X là biến số ngẫu nhiên liên tục có hàm phân phối xác suất sau:

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\frac{x}{2}$$

Tìm giá trị x_1 thỏa mãn điều kiện: $P(X > x_1) = \frac{1}{4}$.

Đáp số: $x_1 = 2$.

Bài số 18. Thời gian xếp hàng chờ mua hàng của khách là biến số ngẫu nhiên liên tục X với hàm phân phối xác suất như sau:

$$F(x) = \begin{cases} 0 & \text{khi } x \le 0 \\ ax^3 - 3x^2 + 2x & \text{khi } 0 < x \le 1 \\ 1 & \text{khi } x > 1 \end{cases}$$

a) Tìm hệ số a.

b) Tìm thời gian trung bình.

c)Tìm xác suất để trong 3 người xếp hàng thì có không quá 2 người phải chờ quá 0,5 phút.

Đáp số: a) $a = 2$ *; b)* $EX = 0,5$ *; c) 0,875.*

Bài số 19. Tỷ lệ mắc một loại bệnh trong một vùng dân cư là biến số ngẫu nhiên liên tục X có hàm mật độ như sau:

$$f(x) = \begin{cases} \dfrac{1}{20} & \text{khi } x \in [5,25] \\ 0 & \text{khi } x \notin [5,25] \end{cases}$$

a) Tính $P(|X-10| > 2,5)$.

b) Tính tỷ lệ mắc bệnh trung bình và phương sai của X.

Đáp số: a) 0,75; b) $EX = 15; \sigma_X^2 = 33,3$.

Bài số 20. Cho biến số ngẫu nhiên liên tục X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} x^m \dfrac{e^{-x}}{m!} & \text{khi } x \geq 0 \\ 0 & \text{khi } x < 0 \end{cases}$$

Tính trung bình và phương sai của X.

Đáp số : $EX = \sigma_X^2 = m + 1$.

Bài số 21. Xác suất để một con gà đẻ trong ngày là 0,6. Nuôi 5 con.

1) Tính xác suất để trong một ngày :

a) không con nào đẻ,

b) cả 5 con đẻ,

c) có ít nhất 1 con đẻ,

d) có ít nhất 2 con đẻ.

2) Nếu muốn mỗi ngày có trung bình 100 trứng thì phải nuôi bao nhiêu con gà.

Đáp số: 1) a) 0,01024; b) 0,07776; c) 0,98976; d) 0,91296; 2) 167 con.

Bài số 22. Một sọt cam có 10 trái trong đó có 4 trái hư. Lấy ngẫu nhiên ra 3 trái.

a) Tính xác suất lấy được 3 trái hư.

b) Tính xác suất lấy được 1 trái hư

c) Tính xác suất lấy được ít nhất 1 trái hư.

d) Tính xác suất lấy được nhiều nhất 2 trái hư.

Đáp số: a) 0,033; b) 0,5; c) 0,83; d) 0,967.

Bài số 23. Một tổng đài bưu điện có các cuộc điện thoại gọi đến xuất hiện ngẫu nhiên, độc lập với nhau và có tốc độ trung bình 2 cuộc gọi trong 1 phút. Biết rằng số cuộc gọi trong một khoảng thời gian cố định có phân phối Poisson. Tìm xác suất để

a) có đúng 5 cuộc điện thoại trong 2 phút,

b) không có cuộc điện thoại nào trong khoảng thời gian 30 giây,

c) có ít nhất 1 cuộc điện thoại trong khoảng thời gian 10 giây.

Đáp số: a) 0,1563; b) 0,3679; c) 0,2835.

Bài số 24. Xác suất để một máy sản xuất ra phế phẩm là 0,02.

a) Tính xác suất để trong 10 sản phẩm do máy sản xuất có không quá 1 phế phẩm.

b) Một ngày máy sản xuất được 250 sản phẩm. Tìm số phế phẩm trung bình và số phế phẩm tin chắc nhất của máy đó trong một ngày.

Đáp số: a) 0,9838; b) $\mathrm{E}(\mathrm{X}) = 5$; $\mathrm{Mod}(\mathrm{X}) = 5$.

Bài số 25. Xác suất để một máy sản xuất ra sản phẩm loại A là 0,25. Tính xác suất để trong 80 sản phẩm do máy sản xuất ra có từ 25 đến 30 sản phẩm loại A.

Đáp số: 0,11927.

Bài số 26. Gieo 100 hạt giống của một loại nông sản. Xác suất nảy mầm của mỗi hạt là 0,8. Tính xác suất để có ít nhất 90 hạt nảy mầm.

Đáp số: 0,0057.

Bài số 27. Giả sử xác suất trúng số là 1%. Mỗi tuần mua một vé số. Hỏi phải mua vé số liên tiếp trong tối thiểu bao nhiêu tuần để có không ít hơn 95% hy vọng trúng số ít nhất 1 lần.

Đáp số: 299.

Bài số 28. Bưu điện dùng một máy tự động đọc địa chỉ trên bì thư để phân loại từng khu vực gởi đi, máy có khả năng đọc được 5000 bì thư trong 1 phút. Khả năng đọc sai 1 địa chỉ trên bì thư là 0,04% (xem như việc đọc 5000 bì thư này là 5000 phép thử độc lập).

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

b) Tính số bì thư tin chắc nhất trong mỗi phút máy đọc sai.

c) Tính xác suất để trong một phút máy đọc sai ít nhất 3 bì thư.

Đáp số: a) 2; b) 2; c) 0,3233.

Bài số 29. Giả sử tỷ lệ dân cư mắc bệnh A trong vùng là 10%. Chọn ngẫu nhiên 1 nhóm 400 người.

a) Viết công thức tính xác suất để trong nhóm có nhiều nhất 50 người mắc bệnh A.

b) Tính xấp xỉ xác suất đó bằng phân phối chuẩn.

Đáp số: b) 0,9599.

Bài số 30. Sản phẩm sau khi hoàn tất được đóng thành kiện, mỗi kiện gồm 10 sản phẩm với tỷ lệ thứ phẩm là 20%. Trước khi mua hàng, khách hàng muốn kiểm tra bằng cách từ mỗi kiện chọn ngẫu nhiên 3 sản phẩm.

1) Lập bảng phân phối xác suất của số sản phẩm tốt trong 3 sản phẩm lấy ra.

2) Nếu cả 3 sản phẩm được lấy ra đều là sản phẩm tốt thì khách hàng sẽ đồng ý mua kiện hàng đó. Tính xác suất để khi kiểm tra 100 kiện

a) có ít nhất 80 kiện hàng được mua,

b) có ít nhất 60 kiện được mua.

Đáp số: 1)

X	0	1	2	3
P	0	0,066	0,467	0,467

2a) $8,2\cdot 10^{-12}$*; 2b) 0,0052.*

Bài số 31. Một trạm cho thuê xe Taxi có 3 chiếc xe. Hàng ngày trạm phải nộp thuế 8USD cho 1 chiếc xe (bất kể xe đó có được thuê hay không). Mỗi chiếc được cho thuê với giá 20USD. Giả sử số xe được yêu cầu cho thuê của trạm trong 1 ngày là đại lượng ngẫu nhiên có phân phối Poisson với $\mu = 2,8$.

a) Tính số tiền trung bình trạm thu được trong một ngày.

b) Giải bài toán trên trong trường hợp trạm có 4 chiếc xe.

c) Theo bạn, trạm nên có 3 hay 4 chiếc xe ?

Đáp số: a) 32; b) 24; c) Thuê 3 xe .

Bài số 32. Đường kính của một loại chi tiết do một máy sản xuất có phân phối chuẩn, kỳ vọng 20mm, phương sai $(0,2\text{mm})^2$. Lấy ngẫu nhiên 1 chi tiết máy. Tính xác suất để

a) có đường kính trong khoảng 19,9mm đến 20,3mm,

b) có đường kính sai khác với kỳ vọng không quá 0,3mm.

Đáp số: a) 0,6247; b) 0,8664.

Bài số 33. Trong hệ thống tỷ giá hối đoái thả nổi, sự biến động của tỷ giá hối đoái chịu sự tác động của nhiều nhân tố và có thể xem như là biến ngẫu nhiên có phân phối chuẩn. Giả sử ở giai đoạn nào đó tỷ giá của USD với VND có trung bình là 18000đ và độ lệch chuẩn là 800đ. Tìm xác suất để trong một ngày nào đó.

a) Tỷ giá sẽ cao hơn 19000đ,

b) Tỷ giá sẽ thấp hơn 17500đ,

c) Tỷ giá nằm trong khoảng từ 17500đ đến 19500.

Đáp số: a) 0,1057; b) 0,266; c) 0,7036.

Bài số 34. Khối lượng của một gói đường (đóng bằng máy tự động) có phân phối chuẩn. Trong 1000 gói đường có 70 gói có khối lượng lớn hơn 1015. Hãy ước lượng xem có bao nhiêu gói đường có khối lượng ít hơn 1008g. Biết rằng khối lượng trung bình của 1000 gói đường là 1012g.

Đáp số: 24,4 gói.

Bài số 35. Lãi suất (%) đầu tư vào một dự án năm 2000 được coi như 1 đại lượng ngẫu nhiên có phân phối theo quy luật chuẩn. Theo đánh giá của uỷ ban đầu tư thì lãi suất cao hơn 20% có xác suất 0,1587, và lãi suất cao hơn 25% có xác suất là 0,0228. Vậy khả năng đầu tư mà không bị thua lỗ là bao nhiêu?

Đáp số: 0,9987.

Bài số 36. Một công ty kinh doanh mặt hàng A dự định sẽ áp dụng một trong 2 phương án kinh doanh. Ký hiệu X_1 là lợi nhuận thu được khi áp dụng phương án thứ 1, X_2 là lợi nhuận thu được khi áp dụng phương án thứ 2. X_1, X_2 đều được tính theo đơn vị triệu đồng/ tháng) và $X_1 \sim N(140, 2500)$, $X_2 \sim N(200, 3600)$. Nếu biết rằng, để công ty tồn tại và phát triển thì lợi nhuận thu được từ mặt hàng kinh doanh A phải đạt ít nhất 80 triệu đồng/tháng. Hãy cho biết công ty nên áp dụng phương án nào để kinh doanh mặt hàng A? Vì sao?.

Đáp số: $P(X_1 \geq 80) = 0,8849 < P(X_2 \geq 80) = 0,9772$, *chọn phương án thứ 2.*

Bài số 37. Độ dài của 1 chi tiết máy được tiện ra có phân phối chuẩn $N(\mu cm; (0,2cm)^2)$. Sản phẩm coi là đạt nếu độ dài sai lệch so với độ dài trung bình không quá 0,3cm.

a) Tính xác suất chọn ngẫu nhiên 1 sản phẩm thì được sản phẩm đạt yêu cầu.

b) Chọn ngẫu nhiên 3 sản phẩm. Tính xác suất có ít nhất 2 sản phẩm đạt yêu cầu .

c) Nếu sản phẩm tốt mà bị loại trong kiểm tra thì mắc phải sai lầm loại 1, nếu sản phẩm không đạt mà được nhận thì mắc phải sai lầm loại 2. giả sử khả năng mắc phải sai lầm loại 1, loại 2 lần lượt là 0,1 và 0,2. Tính xác suất để trong 3 lần kiểm tra hoàn toàn không nhầm lẫn.

Đáp số a) 0,8664; b) 0,9512; c) 0,697.

Bài số 38. Khối lượng của 1 loại trái cây có quy luật phân phối chuẩn với khối lượng trung bình là 250g, độ lệch chuẩn về khối lượng là 5g.

a) Một người lấy 1 trái từ trong sọt trái cây ra. Tính xác suất người này lấy được trái loại 1 (trái loại 1 là trái có khối lượng > 260g).

b) Nếu lấy được trái loại 1 thì người này sẽ mua sọt đó. Người này kiểm tra 100 sọt, tính xác suất mua được 6 sọt.

Đáp số: a) 0,0228; b) 0,019.

Bài số 39. Có hai thị trường A và B, lãi suất của cổ phiếu trên hai thị trường này là các đại lượng ngẫu nhiên có phân phối chuẩn, độc lập với nhau, có kỳ vọng và phương sai được cho trong bảng dưới đây:

	Trung bình	Phương sai
Thị trường A	19%	36
Thị trường B	22%	100

a) Nếu mục đích là đạt lãi suất tối thiểu bằng 10% thì nên đầu tư vào loại cổ phiếu nào?

b) Để giảm rủi ro đến mức thấp nhất thì nên đầu tư vào cổ phiếu trên cả hai thị trường theo tỷ lệ như thế nào?

Đáp số: a) nên đầu tư vào cổ phiếu trên thị trường loại A.

b) 74% vào thị trường A còn lại là thị trường B.

Bài số 40. Nghiên cứu chiều cao của những người trưởng thành, người ta nhận thấy rằng chiều cao đó tuân theo quy luật phân bố chuẩn với trung bình là 175cm và độ lệch tiêu chuẩn 4cm. Hãy xác định :

a) Tỷ lệ người trưởng thành có chiều cao trên 180cm.

b) Tỷ lệ người trưởng thành có chiều cao từ 166cm đến 177cm.

c) Giá trị h_0, nếu biết rằng 33% người trưởng thành có chiều cao dưới mức h_0.

d) Giới hạn biến động chiều cao của 90% người trưởng thành xung quanh giá trị trung bình của nó.

Đáp số: a) 0,1056; b) 0,6793; c) 173,24; d) 6,56.

Bài số 41. Chiều dài của chi tiết được gia công trên máy tự động là biến ngẫu nhiên tuân theo quy luật phân phối chuẩn với độ lệch tiêu chuẩn là 0,01mm. Chi tiết được coi là đạt tiêu chuẩn nếu kích thước thực tế của nó sai lệch so với kích thước trung bình không vượt quá 0,02mm.

a) Tìm tỷ lệ chi tiết không đạt tiêu chuẩn.

b) Xác định độ đồng đều (phương sai) cần thiết của sản phẩm để tỷ lệ chi tiết không đạt tiêu chuẩn chỉ còn 1%.

Đáp số: a) 0,9545; b) $\left(7,752.10^{-3}\right)^2$.

Bài số 42. Khối lượng X của một loại trái cây ở nông trường được biết có kỳ vọng 250gr và phương sai $81\left(\text{gr}\right)^2$. Trái cây được đóng thành sọt, mỗi sọt 100 trái. Mỗi sọt được gọi là loại A nếu khối lượng không dưới 25kg. Kiểm tra ngẫu nhiên 100 sọt. Tính xác suất :

a) có nhiều nhất 60 sọt loại A,

b) ít nhất 45 sọt loại A.

Đáp số: a) 0,9824; b) 0,8644.

Bài số 43. Việc kiểm tra các viên bi được tiến hành như sau: nếu viên bi không lọt qua lỗ có đường kính d_1 song lọt qua lỗ có đường kính d_2 thì viên bi được coi là đạt tiêu chuẩn, nếu không thì viên bi bị loại. Biết đường kính các viên bi sản xuất ra là biến ngẫu nhiên có phân phối chuẩn với trung bình là $\frac{d_1 + d_2}{2}$ và độ lệch chuẩn là $\frac{d_2 - d_1}{4}$. Tìm xác suất để viên bi bị loại.

Đáp số: 0,0456.

Bài số 44. Một đề thi trắc nghiệm có 4 câu hỏi lý thuyết và 3 bài tập độc lập nhau. Khả năng để một sinh viên trả lời đúng một câu hỏi lý thuyết là 0,7 và đúng một bài tập là 0,5. Trả lời đúng một câu hỏi lý thuyết được 1 điểm, sai được 0 điểm. Trả lời đúng mỗi bài tập được 2 điểm, sai được 0 điểm. Tìm số điểm trung bình mà sinh viên đó đạt được.

Đáp số: 5,8.

Bài số 45. Trọng lượng sản phẩm do một máy sản xuất là ĐLNN tuân theo luật phân phối chuẩn với độ lệch chuẩn là 1,6gam. Sản phẩm được coi là đạt tiêu chuẩn nếu trọng lượng của nó sai lệch so với trung bình không quá 2gam.

a) Tính tỉ lệ sản phẩm đạt tiêu chuẩn do máy đó sản xuất.

b) Cần sản xuất tối thiểu bao nhiêu sản phẩm để xác suất "có ít nhất 100 sản phẩm đạt tiêu chuẩn" không bé hơn 90%.

Đáp số : a)0,7888; b) 134.

2.7. Tài liệu tham khảo

[1] Nguyễn Cao Văn, Trần Thái Ninh, Ngô Văn Thứ, Lý thuyết xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[2] Nguyễn Cao Văn, Trần Thái Ninh, Nguyễn Thế Hệ, Bài tập xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[3] Phạm Hoàng Uyên, Lê Thị Thiên Hương, Huỳnh Văn Sáu, Nguyễn Phúc Sơn, Huỳnh Tố Uyên, Lý thuyết xác suất, NXB Đại học Quốc gia Thành Phố Hồ Chí Minh, 2015.

[4] Anderson, Sweeney, and William [2010], Statistics for Business and Economics, South-Western Cengage Learning (11th Edition).

[5] Michael Barrow, Statistics for Economics, Accounting and Business Studies-Prentice Hall, 2006.

[6] Newbold Paul - Statistics for Bussiness and Economics, 5th edition - Prentice Hall, 2005.

Thuật ngữ chính chương 2

Tiếng Anh	Tiếng Việt
Binomial random variables	Biến ngẫu nhiên có phân phối nhị thức
continuous random variable	Biến ngẫu nhiên liên tục
Corollary	Hệ quả
Compute	Tính toán
Chi-squared distribution	Phân phối Chi bình phương
Cumulative distribution function	Hàm phân phối
continuous random variable	Biến ngẫu nhiên liên tục
Discrete random variable	Biến ngẫu nhiên rời rạc
Density function	Hàm mật độ
Gamma and Beta distributions	Phân phối Gamma và Beta
Hypergeometric Random Variable	Biến ngẫu nhiên có phân phối siêu bội
Poisson random variable	Biến ngẫu nhiên có phân phối Poisson
Random variable	Biến ngẫu nhiên
Integration by parts	Tích phân từng phần
Mean	Trung bình
Median	Trung vị
Measurable	Đo được
Normal distribution	Phân phối chuẩn
Normal random variables	Biến ngẫu nhiên có phân phối chuẩn
Expected value	Giá trị kỳ vọng
Exponential random variable	Biến ngẫu nhiên có phân phối mũ
Standard deviation	Độ lệch chuẩn
Parameter	Tham số
Probability density function	Hàm mật độ xác suất
Probability function	Hàm xác suất
Pareto distribution	Phân phối Pareto
Theorem	Định lý
Uniform distribution function	Hàm phân phối đều
Variance	Phương sai

Chương 3

MẪU NGẪU NHIÊN VÀ BÀI TOÁN ƯỚC LƯỢNG

Mục tiêu chương 3

Chương này giúp sinh viên:

- Phân biệt được thế nào là tổng thể và mẫu ngẫu nhiên.
- Tính được các tham số đặc trưng mẫu như trung bình, phương sai, tỷ lệ...
- Nắm được bài toán ước lượng điểm và ước lượng khoảng.
- Hiểu được bài toán ước lượng khoảng và tìm được lượng khoảng như ước lượng trung bình, phương sai, tỷ lệ.

3.1. Mẫu ngẫu nhiên

3.1.1. Tổng thể nghiên cứu

Trong thực tế thường phải nghiên cứu một tập hợp các phần tử đồng nhất theo một hay nhiều dấu hiệu định tính hoặc định lượng đặc trưng cho các phần tử đó. Chẳng hạn một doanh nghiệp phải nghiên cứu tập hợp các khách hàng thì dấu hiệu định tính có thể là mức độ hài lòng của khách hàng đối với sản phẩm hoặc dịch vụ của doanh nghiệp, còn dấu hiệu định lượng là nhu cầu của khách hàng về số lượng sản phẩm của doanh nghiệp.

3.1.1.1. Định nghĩa

Toàn bộ tập hợp các phần tử đồng nhất theo một dấu hiệu nghiên cứu định tính hoặc định lượng nào đó được gọi là tổng thể nghiên cứu hay tổng thể.

Ví dụ 3.1.

a) Nghiên cứu tập hợp các khách hàng của một doanh nghiệp theo dấu hiệu định tính - mức độ hài lòng về sản phẩm, hay định lượng - nhu cầu về số lượng sản phẩm.

b) Nghiên cứu tập hợp học sinh của một lớp: định tính: học lực; định lượng: chiều cao/ cân nặng.

3.1.1.2. Các phương pháp mô tả tổng thể

Giả sử trong tổng thể, dấu hiệu nghiên cứu X nhận các giá trị $X_1, X_2, ..., X_k$ với các tần số tương ứng $N_1, N_2, ..., N_k$; $\sum_{i=1}^{k} N_i = N$ (N còn gọi là kích thước tổng thể).

a. Bảng phân phối tần số của tổng thể

X	X_1	X_2	…	X_k
Tần số	N_1	N_2	…	N_k

b. Bảng phân phối tần suất

Đặt $p_i = \dfrac{N_i}{N}$ $(i = 1,2,...,k)$, gọi là tần suất xuất hiện giá trị X_i

X	X_1	X_2	…	X_k
Tần suất	p_1	p_2	…	p_k

Trong đó, $\sum\limits_{i=1}^{k} p_i = 1; \quad 0 \le p_i \le 1$

Nhận xét : Việc mô tả dấu hiệu X trên một tổng thể bằng các phương pháp trên cho phép chúng ta có thể coi dấu hiệu X như một biến ngẫu nhiên.

Ví dụ 3.2. Điểm thi của học sinh một trường:

X	0	1	…
Tần suất	p_0	p_1	…

Mặc dù kết quả thi đã có (tất nhiên) nhưng dựa trên số liệu thống kê, điểm của một thí sinh A nào đó được coi như 1 biến ngẫu nhiên.

3.1.1.3. Các tham số đặc trưng của tổng thể:

a. Trung bình tổng thể: Là trung bình số học của các giá trị của dấu hiệu trong tổng thể, ký hiệu là μ và được tính bởi công thức:

$$\mu = \frac{1}{N}\sum_{i=1}^{k} N_i X_i = \sum_{i=1}^{k} p_i X_i \tag{3.1}$$

Nếu xem dấu hiệu nghiên cứu như biến ngẫu nhiên X thì trung bình tổng thể chính là kỳ vọng toán của biến ngẫu nhiên đó.

b. Phương sai tổng thể: Là trung bình số học của bình phương các sai lệch giữa các giá trị của các dấu hiệu trong tổng thể và trung bình tổng thể, ký hiệu σ^2 được tính bởi công thức:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} N_i (X_i - \mu)^2 = \sum_{i=1}^{k} p_i (X_i - \mu)^2 \tag{3.2}$$

c. Độ lệch chuẩn tổng thể: ký hiệu là σ và được tính bởi công thức:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{k} p_i (X_i - \mu)^2} \tag{3.3}$$

d. Tỷ lệ tổng thể: là tỷ số giữa số phần tử mang dấu hiệu nghiên cứu, M và kích thước của tổng thể N, ký hiệu p và được xác định

$$p = \frac{M}{N}. \tag{3.4}$$

3.1.2. Mẫu ngẫu nhiên

Các tham số đặc trưng của tổng thể có thể xác định được một cách trực tiếp nếu áp dụng phương pháp nghiên cứu toàn bộ tổng thể. Tuy nhiên trong thực tế việc áp dụng phương pháp này gặp phải những khó khăn chủ yếu sau:

- Phải chịu chi phí rất lớn về thời gian, nhân lực, tiền bạc và phương tiện.
- Nếu quy mô của tập hợp quá lớn có thể xảy ra trường hợp trùng hoặc bỏ sót các phần tử, sai sót trong quá trình thu thập thông tin ban đầu, hạn chế độ chính xác của kết quả phân tích.
- Nếu các phần tử của tập hợp bị phá huỷ trong quá trình nghiên cứu thì phương pháp nghiên cứu toàn bộ trở thành vô nghĩa.
- Chưa thể xác định được toàn bộ N phần tử của tổng thể.

Vì thế trong thực tế phương pháp nghiên cứu toàn bộ thường chỉ được áp dụng đối với các tập hợp có quy mô nhỏ, còn chủ yếu người ta áp dụng phương pháp nghiên cứu không toàn bộ, đặc biệt là phương pháp mẫu bằng cách chọn ra từ tổng thể n phần tử và chỉ tập trung nghiên cứu các phần tử đó. Tập hợp n phần tử này được gọi là mẫu kích thước n.

Phương pháp chọn mẫu:

- Mỗi lần lấy vào mẫu chỉ một phần tử.
- Lấy phần tử nào đưa vào mẫu là hoàn toàn ngẫu nhiên.
- Các phần tử được lấy vào mẫu theo phương thức hoàn lại.

Định nghĩa: Mẫu ngẫu nhiên kích thước n là tập hợp của n biến ngẫu nhiên độc lập $X_1, X_2, ..., X_n$ được thành lập từ biến ngẫu nhiên X trong tổng thể và có cùng quy luật phân phối xác suất với X, ký hiệu: $T = T(X_1, X_2, ..., X_n)$. Ta có

$$E(X_i) = \mu,\ i = 1,2,...,n \text{ và } var(X_i) = \sigma^2,\ i = 1,2,...,n.$$

Ví dụ 3.3. Gọi X là số chấm thu được khi gieo một xúc xắc, X là biến ngẫu nhiên với bảng phân phối xác suất :

X	1	2	3	4	5	6

P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Nếu gieo con xúc xắc 3 lần và gọi X_i là số chấm xuất hiện ở lần gieo thứ i $(i=\overline{1,3})$ thì ta có 3 biến ngẫu nhiên X_1, X_2, X_3 độc lập, cùng phân phối xác suất với X tạo nên một mẫu ngẫu nhiên kích thước $n=3$.

$$T = T(X_1, X_2, X_3).$$

Hơn nữa, mỗi biến ngẫu nhiên X_i trong mẫu đều có bảng phân phối xác suất giống như bảng phân phối xác suất của biến ngẫu nhiên X, do đó:

$$E(X_i) = \frac{1}{6}(1+2+3+4+5+)6 = \frac{21}{6} = E(X), \text{ i=1,2,3.}$$

$$\operatorname{var}(X_i) = \frac{1}{6}\left(1^2+2^2+3^2+4^2+5^2+6^2\right) - \left(\frac{21}{6}\right)^2 = \frac{35}{12} = \operatorname{var}(X), \text{ i=1,2,3.}$$

3.1.3. Các đặc trưng quan trọng của mẫu

3.1.3.1. Thống kê

Thống kê là một biểu thức theo mẫu $X_1, X_2, ..., X_n$ và không phụ thuộc vào các tham số chưa biết. Ký hiệu $T = T(X_1, X_2, ..., X_n)$.

Khi đã quan sát được mẫu, ta có thể tính ra giá trị của một thống kê. Tùy theo từng vấn đề nghiên cứu, ta có thể đặt ra một hay nhiều thống kê khác nhau. Các thống kê thường dùng là

a. Trung bình mẫu :

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.5}$$

b. Phương sai mẫu có hiệu chỉnh:

$$S_X^2 = \frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{3.6}$$

c. Giá trị nhỏ nhất của mẫu : $Y_1 = \min\{X_1, X_2, ..., X_n\}$,

d. Giá trị lớn nhất của mẫu : $Y_n = \max\{X_1, X_2, ..., X_n\}$,

e. Khoảng biến thiên của mẫu : $Y_n - Y_1$.

Ví dụ 3.4. Quan sát chiều cao X (cm) của 10 người, ta ghi được 158, 163, 157, 162, 154, 152, 160, 159, 165, 156.

Với mẫu trên, ta tính được

- Chiều cao trung bình mẫu : $\overline{X} = 158,6\text{cm}$,
- Phương sai của mẫu : $S_X^2 = 16,49\text{cm}^2$,
- Người thấp nhất : $Y_1 = 152\text{cm}$,
- Người cao nhất : $Y_2 = 165\text{cm}$,
- Khoảng biến thiên của mẫu : $R = 13\text{cm}$.

Tương tự như các tham số đặc trưng cho một biến ngẫu nhiên, trung bình mẫu là giá trị mà ta hy vọng nhận được khi xem xét một phần tử của mẫu, phương sai mẫu cho ta biết mức độ phân tán của số liệu mẫu. Phương sai càng nhỏ, số liệu càng ít phân tán.

Để có thể tính toán phương sai một cách nhanh chóng, ta có thể dùng kết quả sau

3.1.3.2. Mệnh đề. *Xét mẫu* $X_1, X_2, ..., X_n$ *với trung bình* $\overline{X}$ *và phương sai* S_X^2, *ta có*

$$S_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right). \tag{3.7}$$

Lưu ý rằng trong trường hợp số liệu mẫu có lặp lại và ta ghi nhận các tần số xuất hiện số liệu như sau

X	X_1	X_2	...	X_k
Tần số	n_1	n_2	...	n_k

thì cỡ mẫu là $n = n_1 + n_2 + \cdots + n_k$ và các công thức tính trung bình cũng như phương sai trở thành

$$\overline{X} = \frac{n_1X_1 + n_2X_2 + \cdots + n_kX_k}{n} = \frac{1}{n}\sum_{i=1}^{k} n_iX_i,$$

$$S_X^2 = \frac{n_1\left(X_1 - \overline{X}\right)^2 + n_2\left(X_2 - \overline{X}\right)^2 + \cdots + n_k\left(X_k - \overline{X}\right)^2}{n-1}$$

$$= \frac{1}{n-1}\sum_{i=1}^{k} n_i\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_iX_i^2 - n\overline{X}^2\right).$$

Ví dụ 3.5. Một máy tự động đóng bột vào bao. Cân ngẫu nhiên 15 bao được các trọng lượng sau:

39,75	40,25	39,50	40,25	40,50
40,00	39,75	40,00	40,00	39,25
39,25	39,50	40,00	39,50	39,50

a) Lập bảng phân phối tần số thực nghiệm của trọng lượng các bao bột.

b) Tính giá trị trung bình và phương sai mẫu hiệu chỉnh.

Giải

a) Bảng phân phối tần số thực nghiệm:

Trọng lượng (kg)	39,25	39,50	39,75	40,00	40,25	40,50
Số bao	2	4	2	4	2	1

b) Gọi X là trọng lượng các bao bột.

Ta có:

Cỡ mẫu: $n = 2 + 4 + 2 + 4 + 2 + 1 = 15$

Trung bình mẫu:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i X_i = \frac{2\cdot 39,25 + 4\cdot 39,5 + 2\cdot 39,75 + 4\cdot 40 + 2\cdot 40,25 + 1\cdot 40,50}{15} = 39,8$$

Phương sai mẫu có hiệu chỉnh

$$\sum_{i=1}^{k} n_i X_i^2 = 2\cdot 39,25^2 + 4\cdot 39,5^2 + 2\cdot 39,75^2 + 4\cdot 40^2 + 2\cdot 40,25^2 + 1\cdot 40,50^2$$

$$= 23762,625$$

Áp dụng (3.7), ta có

$$S_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i X_i^2 - n\overline{X}^2\right) = \frac{1}{14}\left(23762,625 - 15\cdot 39,8^2\right) = 0,1446.$$

Ví dụ 3.6. Gặt ngẫu nhiên 100 điểm trồng lúa của một vùng nông thôn ta thu được bảng số liệu như sau:

Năng suất	30	33	34	36	40
Số điểm	15	20	41	18	6

Xác định các thống kê đặc trưng mẫu.

Giải

Gọi X là năng suất lúa (tạ/ha).Ta có mẫu cụ thể kích thước $n = 100$.

Ta có thể tính toán dựa vào bảng sau:

x_i	n_i	$n_i X_i$	$n_i X_i^2$
30	15	450	13500
33	20	660	21780
34	41	1394	47396
36	18	648	23328
40	6	240	9600

	$\sum n_i = n = 100$	$\sum n_i x_i = 3392$	$\sum n_i x_i^2 = 115604$

Từ đó ta có:

Năng suất lúa trung bình: $\overline{X} = \dfrac{3392}{100} = 33,92$

Phương sai mẫu có hiệu chỉnh của năng suất lúa:

$$S_X^2 = \frac{1}{99}\left(115604 - 100 \cdot 33,92^2\right) = 5,5289.$$

3.1.3.3. Định lý Lindeberg-Lévy. Nếu mẫu $X_1, X_2, ..., X_n$ lấy từ tổng thể có phân phối chuẩn $N\left(\mu, \sigma^2\right)$ nghĩa là $X_i \sim N\left(\mu, \sigma^2\right)$, với mọi i, thì

i) $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

ii) $\dfrac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1)$.

3.2. Trình bày kết quả điều tra

Đối với dữ liệu định tính hoặc dữ liệu định lượng ít biểu hiện, ta sử dụng:

- Bảng tần số, tần suất, tần số tích lũy, tần suất tích lũy;
- Biểu đồ hình tròn, biểu đồ hình cột, biểu đồ hình thanh.

3.2.1. Trình bày kết quả điều tra dưới dạng bảng

Bảng tần số, tần suất

Trị số của biến (X_i)	**Tần số (f_i)**	**Tần suất (%)**
X_1	f_1	$\dfrac{f_1}{n}$
X_2	f_2	$\dfrac{f_2}{n}$
…	…	…
X_k	f_k	$\dfrac{f_k}{n}$
Tổng số	$\sum_{i=1}^{k} f_i = n$	100%

Bảng tần số gồm hai phần

- Trị số của biến nghiên cứu, ký hiệu X_i.

- Số lần xuất hiện của trị số gọi là tần số, ký hiệu f_i.

Cũng có thể thể hiện tần số bằng hình thức phần trăm (%)

Ví dụ 3.7. Kết quả khảo sát 200 người về lựa chọn màu của xe máy mà họ yêu thích như sau:

Màu xe	Tần số	Tần suất (%)
Đỏ	30	15
Đen	60	30
Xanh	50	25
Trắng	20	10
Khác	40	20
Tổng số	200	100

Việc cộng dồn các giá trị tần số và tần suất cho ta các giá trị tần số tích lũy và tần suất tích lũy cho dữ liệu định tính dạng thứ bậc nhằm cung cấp thông tin thêm cho người đọc.

Trị số của biến (X_i)	Tần số tích lũy (f_i)	Tần suất tích lũy (%)
X_1	f_1	$\frac{f_1}{n}$
X_2	f_1+f_2	$\frac{f_1+f_2}{n}$
…	…	…
X_k	$f_1+f_2+\cdots+f_k$	$\frac{f_1+f_2+\cdots+f_k}{n}$

Ví dụ 3.8. Khảo sát 50 người về mức lương hằng tháng mà họ nhận được trong vòng 5 năm trở lại đây như sau:

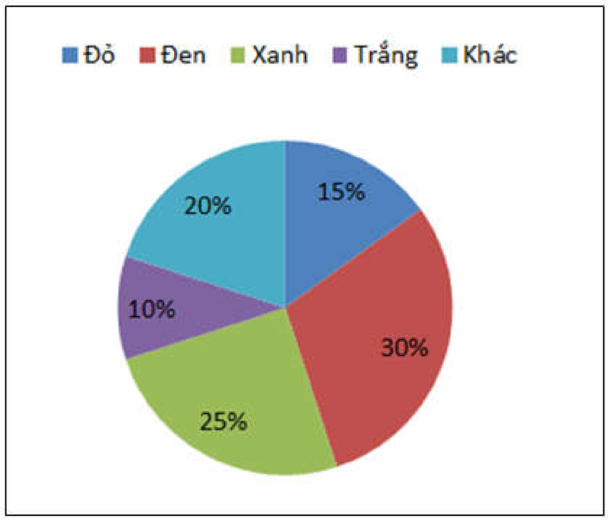
Mức lương (triệu đồng/tháng)	Tần số	Tần số tích lũy	Tần suất (%)	Tần suất tích lũy (%)
<3	5	5	10	10
[3;5)	15	20	30	40
[5;10)	20	40	40	80
>10	10	50	20	100
Tổng số	50		100	

Ý nghĩa: Giá trị 40 của tần số tích lũy cho biết có 40 người trong tổng số những người được khảo sát có thu nhập nhiều nhất là 10 triệu đồng/tháng.

3.2.2. Trình bày kết quả điều tra bằng biểu đồ

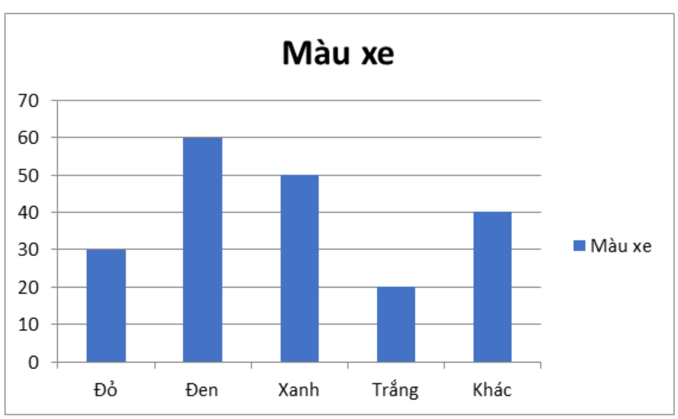
3.3.2.1. Biểu đồ hình tròn: là biểu đồ mà dữ liệu thể hiện các nhóm giá trị khác nhau được phân biệt dựa vào mà sắc, nhóm giá trị nào có tần số càng lớn thì phần màu sắc tượng trưng cho nhóm giá trị đó sẽ càng to hơn các nhóm giá trị khác. Mỗi một màu sắc được thể hiện là một hình quạt

Ví dụ 3.9. Với dữ liệu ở ví dụ 3.7, ta có dạng biểu diễn số lượng màu xe như hình vẽ sau: trong đó mỗi màu xe là hình quạt.



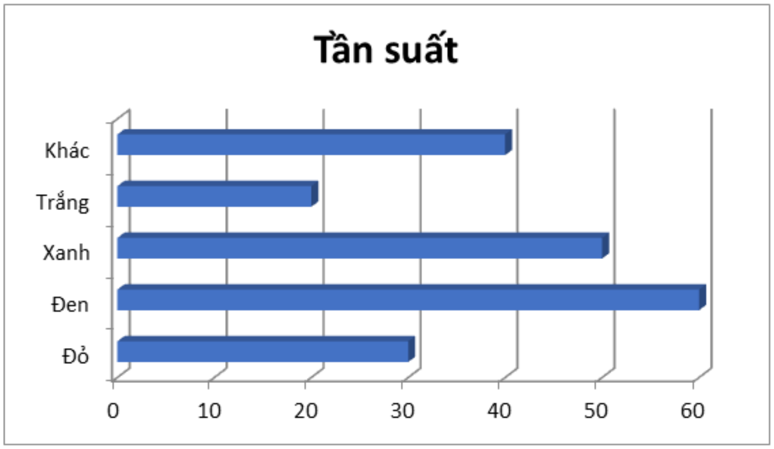
3.3.2.2. Biểu đồ hình cột: là dạng biểu đồ được sử dụng cho các dữ liệu định tính hoặc định lượng nhưng có ít biểu hiện hoặc định lượng đã được phân khoảng. Trục hoành thể hiện giá trị của vấn đề nghiên cứu; trục tung thể hiện tần số. Mỗi cột thể hiện một giá trị; độ cao của cột là tần số.

Ví dụ 3.10. Với dữ liệu ở ví dụ 3.7, ta có dạng biểu diễn số lượng màu xe được thống kê khi khảo sát như hình vẽ sau:



3.3.2.3. Biểu đồ hình thanh: là dạng biểu đồ hình cột mà khi quay ngang, hai trục đổi vị trí cho nhau, nó hay sử dụng khi giá trị của vấn đề nghiên cứu dài.

Ví dụ 3.11. Với dữ liệu ở ví dụ 3.7, ta có dạng biểu diễn số lượng màu xe như hình vẽ sau:



3.2.3. Tính giá trị của các đặc trưng mẫu qua số liệu điều tra

3.2.3.1. Các đặc trưng đo lường khuynh hướng tập trung

a. Số trung bình: là mức độ đại diện điển hình cho 1 tiêu thức nào đó của tổng thể mà các đơn vị của tổng thể biểu hiện nhiều mức độ khác nhau.

Số trung bình cộng đơn giản: với mẫu khảo sát : $X_1, X_2, ..., X_n$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.8}$$

Số trung bình cộng có trọng số : với số liệu mẫu có lặp lại và ta ghi nhận các tần số xuất hiện số liệu như sau

X	X_1	X_2	...	X_k
Tần số	f_1	f_2	...	f_k

thì cỡ mẫu là $n = f_1 + f_2 + \cdots + f_k$ và công thức tính trung bình trở thành

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} f_i X_i. \tag{3.9}$$

Lưu ý: Nếu số liệu được cho dưới dạng có khoảng cách tổ thì ta lấy điểm giữa đại diện.

Ví dụ 3.12. Năng suất lao động của công nhân trong một phân xưởng sau

Mức năng suất lao động (tạ/người) (X_i)	21	23	25	27	29
Số công nhân (f_i)	5	10	30	15	5

Tính Năng suất lao động trung bình. Áp dụng công thức (3.9), ta có:

$$\overline{X}=\frac{5\times21+10\times23+30\times25+15\times27+5\times29}{5+10+30+15+5}=25{,}154.$$

b. Mốt (Mode): là biểu hiện của một tiêu thức được gặp nhiều nhất trong tổng thể. Đối với một dãy số lượng biến, mốt là lượng biến có tần số lớn nhất. Ký hiệu M_0

Cách xác định mốt

+) Đối với đại lượng biến rời rạc: Mốt là lượng biến có tần số lớn nhất.

Ví dụ 3.13. Với dữ liệu ví dụ 3.12. Ta có tần số lớn nhất là 30 nên $M_0 = 25$.

+) Đối với đại lượng biến có khoảng cách tổ

Bước 1: Xác định tổ chứa mốt, là tổ có mật độ phân phối lớn nhất.

Mật độ phân phối là tỷ số giữa các tần số và trị số khoảng cách tổ tương ứng

$$\text{Mật độ phân phối}=\frac{\text{Tần số}}{\text{Trị số khoảng cách tổ}}=\frac{f_i}{d_i}(i=\overline{1,n})$$

Bước 2. Tính trị số Mốt gần đúng theo CT

$$M_0=x_{M_{0(min)}}+d_{M_0}\frac{F_{M_0}-F_{M_{0-1}}}{\left(F_{M_0}-F_{M_{0-1}}\right)+\left(F_{M_0}-F_{M_{0+1}}\right)} \quad (3.10)$$

- $x_{M_0(min)}$: là giới hạn dưới của khoảng cách tổ có Mốt
- d_{M_0} : là trị số khoảng cách tổ có Mốt
- F_{M_0} : là mật độ phân phối của tổ có Mốt
- $F_{M_{0-1}}$: là mật độ phân phối của tổ đứng trước tổ có Mốt
- $F_{M_{0+1}}$: là mật độ phân phối của tổ đứng sau tổ có Mốt

Lưu ý: Nếu khoảng cách tổ bằng nhau thì công thức xác định mốt là

$$M_0=x_{M_{0(min)}}+d_{M_0}\frac{f_{M_0}-f_{M_{0-1}}}{\left(f_{M_0}-f_{M_{0-1}}\right)+\left(f_{M_0}-f_{M_{0+1}}\right)} \quad (3.11)$$

Ví dụ 3.14. Năng suất lao động của công nhân trong một phân xưởng sau

Phân tổ công nhân theo mức năng suất lao động (tạ/người)	Số công nhân (f_i)
20-22	5
22-24	10
24-26	20
26-28	15

28-30	5

Do khoảng cách tổ bằng nhau nên áp dụng công thức (3.11), ta có

$$M_0 = 24 + (26-24)\frac{10-5}{(10-5)+(10-7,5)} = 25,33.$$

c. Số trung vị : là lượng biến của đơn vị tổng thể ở vị trí giữa trong dãy số lượng biến. Số trung vị chia dãy số lượng biến thành hai phần, mỗi phần có số đơn vị tổng thể bằng nhau, ký hiệu Me.

Đối với dãy số lượng biến rời rạc:

+) Trường hợp số đơn vị tổng thể lẻ ($n = 2m+1$). Số trung vị sẽ là đại lượng ở vị trí thứ ($m+1$): $Me = X_{m+1}$.

Ví dụ 3.15. Ta có mức năng suất lao động của năm công nhân trong một tổ là 20, 22, 25, 27, 29 tạ/người. Vậy số trung vị: Me = 25 tạ/ người.

+) Trường hợp số đơn vị tổng thể chẵn ($n = 2m$). Số trung vị sẽ là đại lượng giữa vị trí thứ (m) và ($m+1$): $Me = \dfrac{X_m + X_{m+1}}{2}$.

Ví dụ 3.16. Ta có mức năng suất lao động của sáu công nhân trong một tổ là: 20, 22, 25, 27, 29, 30 tạ/người. vậy số trung vị: Me = 26 tạ/ người.

Đối với dãy số có lượng biến là khoảng cách tổ

Bước 1. Xác định số tổ chứa trung vị là số tổ có tần số tích lũy bằng hoặc lớn hơn hay bằng một nửa tổng các tần số cộng 1.

Bước 2. Xác định giá trị gần đúng của trung vị

$$M_e = X_{M_{e(min)}} + d_{M_e}\frac{\frac{1}{2}\sum f_i - S_{M_{e-1}}}{f_{M_e}} \tag{3.12}$$

Trong đó

- $X_{Me(min)}$: là giá trị giới hạn dưới của tổ chứa trung vị
- d_{Me} : là trị số khoảng cách tổ chứa trung vị
- $\sum f_i$:là tổng tần số
- S_{Me-1} : là tần số tích lũy của tổ đứng trước tổ chứa trung vị.
- f_{Me} : Tần số của tổ chứa trung vị.

Ví dụ 3.17. Về mức lương của công nhân trong một phân xưởng của xí nghiệp X như sau:

Mức lương 1 công nhân (triệu đồng / người)	Số công nhân (người)	Tần số Tích lũy
2,7 – 2,9	5	5
2,9 – 3,1	10	15
3,1 – 3,3	20	35
3,3 – 3,5	15	50
3,5 – 3,7	5	55

Ta có $35 \geq 55:2$ nên tổ chứa trung vị là tổ thứ 3. Áp dụng công thức (3.12), ta có

$$M_e = 3,1 + 0,2\frac{55:2-15}{20} = 3,225.$$

3.2.3.2. Các đặc trưng đo lường khuynh hướng phân tán

a. Khoảng biến thiên: là hiệu số giữa lượng biến lớn nhất và lượng biến nhỏ nhất của tiêu thức nghiên cứu.

$$R = X_{max} - X_{min} \tag{3.13}$$

b. Phương sai hiệu chỉnh: là số bình quân số học của bình phương các độ chênh lệch giữa các lượng biến với số bình quân số học của các biến đó

Phương sai có hiệu chỉnh đơn giản: với mẫu khảo sát : $X_1, X_2, ..., X_n$

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{3.14}$$

Phương sai có hiệu chỉnh có trọng số : với số liệu mẫu có lặp lại và ta ghi nhận các tần số xuất hiện số liệu như sau

X	X_1	X_2	...	X_k
Tần số	f_1	f_2	...	f_k

thì cỡ mẫu là $n = f_1 + f_2 + \cdots + f_k$ và công thức tính phương sai có hiệu chỉnh trở thành

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i\left(X_i - \overline{X}\right)^2. \tag{3.15}$$

Lưu ý: Nếu số liệu được cho dưới dạng có khoảng cách tổ thì ta lấy điểm giữa đại diện.

c. Độ lệch chuẩn có hiệu chỉnh: là căn bậc 2 của phương sai, hay nói cách khác là số bình quân toàn phương của các độ lệch giữa các lượng biến với số bình quân số học của chúng.

$$S_X = \sqrt{S_X^2}. \tag{3.16}$$

Ví dụ 3.18. Với dữ liệu ví dụ 3.12:

Phương sai có hiệu chỉnh: $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i\left(X_i - \overline{X}\right)^2 = 4{,}0385.$

Độ lệch chuẩn có hiệu chỉnh: $S_X = 2{,}01.$

d. Tứ phân vị (Quartiles): là chia dãy số thành 4 phần, mỗi phần có số đơn vị bằng nhau.

Cách xác định tứ phân vị

+) Tài liệu phân tổ không có khoảng cách tổ

Q_1 : Tứ phân vị thứ 1 là giá trị đứng ở vị trí (n+1)/4, là phân vị thứ 25.

Q_2 : Tứ phân vị thứ 2 chính là trung vị Me đứng ở vị trí (n+1)/2, là phân vị thứ 50.

Q_3 : Tứ phân vị thứ 3 là giá trị đứng ở vị trí 3(n+1)/4, là phân vị thứ 75.

Ví dụ 3.19. Giả sử ta có dãy số liệu như sau

5	5	6	7	8	8	9

Ta có $Q_1 = 5$; $Q_2 = Me = 7$; $Q_3 = 8$.

Nếu (n+1) không chia hết cho 4 thì tứ phân vị được xác định bằng cách cộng thêm vào như ví dụ sau:

Ví dụ 3.20. Xét tiền lương của 8 công nhân như sau

3600	3800	4000	4200	4400	5000	5400	5600

Ta có $\frac{n+1}{4} = \frac{9}{4} = 2{,}25$ nên các xác định tứ phân vị như sau:

Phân vị thứ 1: là giá trị nằm giữa quan sát thứ 2 và thứ 3 theo tọa độ lệch 0,25 gần phía quan sát thứ 2 nên cách xác định Q_1 như sau:

$$Q_1 = 3800 + 0{,}25\left(4000 - 3800\right) = 3850$$

Phân vị thứ 2: là giá trị nằm giữa quan sát thứ 4 và thứ 5 nên cách xác định Q_2 như sau

$$Q_2 = Me = 0{,}5\left(4200 + 4400\right) = 4300$$

Phân vị thứ 3: là giá trị nằm giữa quan sát thứ 6 và thứ 7 theo tọa độ lệch 0,75 gần phía quan sát thứ 6 nên cách xác định Q_3 như sau

$$Q_3 = 5000 + 0{,}75\left(5400 - 5000\right) = 5300$$

+) Tài liệu phân tổ có khoảng cách tổ

Tổ chứa phân vị thứ i có tần số tích lũy $\geq \frac{n+1}{4}i$

Tứ phân vị thứ 1:

$$Q_1 = X_{Q_{1(min)}} + d_{Q_1}\frac{\frac{1}{4}\sum f_i - S_{Q_{1-1}}}{f_{Q_1}} \tag{3.17}$$

Tứ phân vị thứ 3 :

$$Q_3 = X_{Q_{3(min)}} + d_{Q_3}\frac{\frac{3}{4}\sum f_i - S_{Q_{3-1}}}{f_{Q_3}} \tag{3.18}$$

Ví dụ 3.21. Khảo sát doanh thu của các cửa hàng ta có bảng số liệu sau:

Doanh thu (tr.đ)	Cửa hàng (f_i)	Tần số tích lũy (S_i)
200-400	8	8
400-500	12	20
500-600	25	45
600-800	25	70
800-1000	9	79
Tổng	79	

Tính phân vị thứ 1 và phân vị thứ 3 của doanh thu.

Giải

Xác định phân vị thứ 1: Vì $\frac{n+1}{4} = \frac{80}{4} = 20.$ Ta có $S_2 = 20 \geq 20$. Vậy tổ chứa phân vị thứ 1 là tổ 2. Áp dụng công thức (3.17), ta có

$$Q_1 = 400 + 100\frac{79:4-8}{12} = 497,92.$$

Xác định phân vị thứ 3: Vì $\frac{n+1}{4}3 = \frac{80}{4}3 = 60.$ Ta có $S_4 = 70 \geq 60$. Vậy tổ chứa phân vị thứ 3 là tổ 4. Áp dụng công thức (3.18), ta có

$$Q_3 = 600 + 200\frac{(3:4)79-45}{25} = 714.$$

3.3. Ước lượng tham số

Giả sử biến ngẫu nhiên gốc X có quy luật phân phối xác suất đã biết nhưng chưa biết tham số θ nào đó của nó. Phải ước lượng (xác định 1 cách gần đúng) giá trị của θ.

Có hai phương pháp là phương pháp ước lượng điểm và phương pháp ước lượng bằng khoảng tin cậy.

3.3.1. Phương pháp ước lượng điểm

Phương pháp ước lượng điểm dùng một thống kê $\hat{\theta}$ nào đó của mẫu ngẫu nhiên để thay thế cho tham số θ chưa biết của tổng thể. Có nhiều cách chọn thống kê $\hat{\theta}$ khác nhau tạo nên những phương pháp ước lượng điểm khác nhau.

3.3.1.1. Phương pháp hàm ước lượng (phương pháp mômen)

a. Khái niệm: Giả sử cần ước lượng tham số θ của biến ngẫu nhiên gốc X. Từ tổng thể lập mẫu ngẫu nhiên kích thước $n: W = (X_1, X_2, ... X_n)$. Chọn lập thống kê $\hat{\theta} = (X_1, X_2, ... X_n)$ mà thực chất là một thống kê đặc trưng mẫu tương ứng với tham số θ cần ước lượng. Chẳng hạn, để ước lượng kỳ vọng toán μ của biến ngẫu nhiên gốc thì chọn thống kê trung bình mẫu $\overline{X}$, để ước lượng phương sai σ^2 của biến ngẫu nhiên gốc thì chọn thống kê S_X^2,…Nếu lập một mẫu cụ thể và tính được giá trị $\hat{\theta} = f(X_1, X_2, ..., X_n)$ của thống kê $\hat{\theta}$ trên mẫu cụ thể đó thì nó là ước lượng của θ.

Thống kê $\hat{\theta}$ được gọi là *hàm ước lượng* của θ.

b. Các tiêu chuẩn lựa chọn hàm ước lượng

Ước lượng không chệch: Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng không chệch của tham số θ của biến ngẫu nhiên gốc X nếu $E(\hat{\theta}) = \theta$. Ngược lại nếu $E(\hat{\theta}) \neq \theta$. thì $\hat{\theta}$ được gọi là ước lượng chệch của θ.

Nhận xét:

- Trung bình mẫu $\overline{X}$ là ước lượng không chệch của trung bình tổng thể μ của biến ngẫu nhiên gốc, nghĩa là $E(\overline{X}) = \mu$.
- Tỷ lệ mẫu f là ước lượng không chệch của tỷ lệ tổng thể p của biến ngẫu nhiên gốc, nghĩa là $E(f) = p$.
- Phương sai mẫu S_X^2 là ước lượng không chệch của phương sai tổng thể σ^2.

Ước lượng hiệu quả: Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng hiệu quả nhất của tham số θ của biến ngẫu nhiên gốc X nếu nó là ước lượng không chệch và có phương sai nhỏ nhất so với mọi ước lượng không chệch khác được xây dựng trên cùng mẫu đó.

Khi hai ước lượng $\hat{\theta}_1$ và $\hat{\theta}_2$ nào đó đều là các ước lượng không chệch của θ song không phải là ước lượng hiệu quả nhất thì có thể so sánh phương sai của hai ước lượng đó để tìm ra ước lượng hiệu quả hơn. Giả sử $\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$, lúc đó độ hiệu quả của $\hat{\theta}_1$ so với $\hat{\theta}_2$ được xác định bằng biểu thức:

$$EF = \frac{\text{var}(\hat{\theta}_2)}{\text{var}(\hat{\theta}_1)} \tag{3.19}$$

Ước lượng vững: Thống kê $\hat{\theta}$ của mẫu được gọi là ước lượng vững của tham số θ của biến ngẫu nhiên gốc X nếu $\hat{\theta}$ tiến về θ khi mẫu lớn.

Ước lượng đủ: Một ước lượng $\hat{\theta}$ được gọi là ước lượng đủ nếu nó chứa đựng toàn bộ các thông tin trong mẫu về tham số θ của ước lượng.

c. Một vài kết luận của phương pháp hàm ước lượng

Dùng những tiêu chuẩn trên để đánh giá các thống kê đặc trưng mẫu khác nhau cho phép ta lựa chọn được những thống kê tốt nhất, tức là ước lượng một cách chính xác nhất các tham số đặc trưng của tổng thể. Có thể đưa ra một số kết luận chung sau đây:

- Vì trung bình mẫu $\overline{X}$ là ước lượng không chệch, hiệu quả nhất và vững của trung bình tổng thể μ và đồng thời là ước lượng tuyến tính không chệch tốt nhất, do đó nếu chưa biết μ có thể dùng $\overline{X}$ để ước lượng nó.

- Vì tần suất mẫu f là ước lượng không chệch, hiệu quả nhất và vững của tần suất tổng thể p và đồng thời là ước lượng tuyến tính không chệch tốt nhất, do đó nếu chưa biết p có thể dùng f để ước lượng nó.

- Vì phương sai S_X^2 là các ước lượng không chệch của phương sai tổng thể σ^2, do đó nếu chưa biết σ^2 có thể dùng S_X^2.

3.3.1.2. Phương pháp ước lượng hợp lý tối đa

Giả sử đã biết quy luật phân phối xác suất của biến ngẫu nhiên gốc X dưới dạng hàm mật độ $f(x,\theta)$ hoặc biểu thức xác suất nếu X là biến ngẫu nhiên rời rạc. Cần phải ước lượng tham số θ nào đó của X.

Lập mẫu ngẫu nhiên kích thước n: $W = (X_1, X_2, ... X_n)$ và xây dựng hàm hợp lý tại một giá trị cụ thể của mẫu: $L(X_1, X_2, ..., X_n, \theta) = f(X_1, \theta) \cdot f(X_2, \theta) \cdots f(X_n, \theta)$. Giá trị của thống kê θ tại điểm đó: $\hat{\theta} = f(X_1, X_2, ..., X_n)$ được gọi là ước lượng hợp lý tối đa của θ nếu ứng với giá trị này hàm hợp lý đạt cực đại.

Cách tìm giá trị của θ *để hàm hợp lý đạt cực đại:*

- Tìm đạo hàm bậc nhất của $\ln[L(X, \theta)]$ theo θ.

- Giải phương trình: $\dfrac{d \ln[L(X, \theta)]}{d\theta} = 0$, giả sử có nghiệm $\theta = \hat{\theta} = f(X_1, X_2, ..., X_n)$.

- Tìm đạo hàm bậc hai $\dfrac{d^2 \ln[L(X, \theta)]}{d\theta^2}$, nếu tại điểm $\theta = \hat{\theta}$ đạo hàm bậc hai âm thì tại điểm này hàm lnL đạt cực đại, do đó $\hat{\theta} = f(X_1, X_2, ..., X_n)$ là ước lượng điểm hợp lý tối đa cần tìm của θ.

3.3.2. Phương pháp ước lượng bằng khoảng tin cậy

Trong phần này, ta chỉ đề cập đến ước lượng trung bình và phương sai trong phân phối chuẩn $N(\mu, \sigma^2)$ và ước lượng tỷ lệ trong phân phối Bernoulli $B(1; p)$.

Cụ thể, để ước lượng tham số θ của một tổng thể có phân phối chuẩn hay có phân phối Bernoulli, ta lập mẫu ngẫu nhiên $X_1, X_2, ..., X_n$ và xác định một thống kê $G = f(X_1, X_2, ..., X_n, \theta)$ sao cho mặc dù θ chưa được xác định nhưng quy luật phân phối của G vẫn được xác định. Từ phân phối của G và với mức xác suất α cho trước, ta tìm được khoảng tin cậy $[a, b]$ sao cho

$$P(a \le G \le b) = 1 - \alpha \tag{3.20}$$

và từ các bất đẳng thức $a \le f(X_1, X_2, ..., X_n, \theta) \le b$, ta suy ra ước lượng khoảng cho θ.

3.3.3. Bài toán ước lượng khoảng tin cậy cho giá trị trung bình

3.3.3.1. Ước lượng μ trong phân phối $N(\mu, \sigma_0^2)$, (σ_0^2 biết)

Để ước lượng trung bình μ trong phân phối chuẩn có phương sai đã biết σ_0^2, ta dùng $X_1, X_2, ..., X_n$ là mẫu độc lập, $X_i \sim N(\mu, \sigma_0^2)$ để ước lượng μ. Ta có, do định lý Lindeberg-Lévy,

$$\overline{X} \sim N\left(\mu, \frac{\sigma_0^2}{n}\right)$$

nên

$$Z = \frac{\left(\overline{X} - \mu\right)\sqrt{n}}{\sigma_0} \sim N\left(0,1\right).$$

Với độ tin cậy bằng $\left(1-\alpha\right)$ cho trước tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$, từ đó tìm được hai giá trị tới hạn tương ứng của phân phối chuẩn hóa là $a = z_{1-\alpha_1}$ và $b = z_{\alpha_2}$ thoả mãn: $P\left(Z < a\right) = 1 - \alpha_1$ và $P\left(Z > b\right) = \alpha_2$.

$$\Leftrightarrow P\left(a \le Z \le b\right) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha.$$

$$\Leftrightarrow P\left(-a \le Z \le b\right) = 1 - \alpha.$$

Thay biểu thức của Z vào và biến đổi ta được:

$$P\left(\overline{X} - \frac{b\sigma_0}{\sqrt{n}} \le \mu \le \overline{X} + \frac{a\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

Với độ tin cậy bằng $\left(1-\alpha\right)$ tham số μ của biến ngẫu nhiên gốc X sẽ nằm trong khoảng

$$\left[\overline{X} - \frac{b\sigma_0}{\sqrt{n}} \le \mu \le \overline{X} + \frac{a\sigma_0}{\sqrt{n}}\right]. \tag{3.21}$$

Biểu thức (3.21) mới chỉ cho ta khoảng tin cậy tổng quát. Với độ tin cậy $\left(1-\alpha\right)$ cho trước, có thể tìm được vô số cặp giá trị α_1 và α_2 thỏa mãn điều kiện: $\alpha_1 + \alpha_2 = \alpha$, từ đó sẽ có vô số khoảng tin cậy tương ứng. Trong thực tế người ta thường dùng các trường hợp đặc biệt sau:

Khoảng tin cậy đối xứng: Chọn $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, ta có giá trị tới hạn $C = z_{\alpha/2}$, khoảng tin cậy của μ là:

$$\left[\overline{X} - \frac{C\sigma_0}{\sqrt{n}}; \overline{X} + \frac{C\sigma_0}{\sqrt{n}}\right] \tag{3.22}$$

Với sai số ước lượng : $\varepsilon = \frac{C\sigma_0}{\sqrt{n}}.$

ε được gọi là *sai số* hay *độ chính xác* của ước lượng. Nó phản ánh mức độ sai lệch của trung bình mẫu so với trung bình tổng thể với xác suất $(1-\alpha)$ cho trước.

Khoảng tin cậy phía phải (ước lượng tối thiểu): Chọn $\alpha_1 = 0$ và $\alpha_2 = \alpha$ thì $z_{\alpha_1} = z_0 = +\infty$, ta có giá trị tới hạn $C = z_\alpha$, khoảng tin cậy của μ là:

$$\left[\overline{X} - \frac{C\sigma_0}{\sqrt{n}}; +\infty\right) \tag{3.23}$$

Khoảng tin cậy phía trái (ước lượng tối đa): Chọn $\alpha_1 = \alpha$ và $\alpha_2 = 0$ thì $z_{\alpha_2} = z_0 = +\infty$, ta có giá trị tới hạn $C = z_\alpha$, khoảng tin cậy của μ là:

$$\left(-\infty; \overline{X} + \frac{C\sigma_0}{\sqrt{n}}\right] \tag{3.24}$$

Với cùng độ tin cậy $(1-\alpha)$, khoảng tin cậy nào ngắn hơn sẽ tốt hơn. Trong trường hợp này độ dài khoảng tin cậy I sẽ là ngắn nhất khi khoảng tin cậy là đối xứng.

Khi đó:

$$I = 2\varepsilon = 2\frac{C\sigma_0}{\sqrt{n}} \tag{3.25}$$

Ví dụ 3.22. Trọng lượng một loại sản phẩm là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn là 1 gam. Cân thử 25 sản phẩm loại này ta thu được kết quả sau:

Trọng lượng (gam)	18	19	20	21
Số sản phẩm	3	5	15	2

Với độ tin cậy 95%, hãy tìm khoảng tin cậy đối xứng của trọng lượng trung bình của loại sản phẩm nói trên.

Giải

Gọi X là trọng lượng sản phẩm, $X \sim N(\mu, 1)$. Đây là bài toán ước lượng tham số μ của biến ngẫu nhiên phân phối chuẩn bằng khoảng tin cậy đối xứng khi đã biết σ_0^2

Ta có: $n = 25$, $\overline{X} = 19,64$; $\sigma_0 = 1$

Với độ tin cậy: $\gamma = 0,95$ suy ra mức ý nghĩa $\alpha = 1 - \gamma = 1 - 0,95 = 0,05$

Ta có $\phi_0(C) = \frac{1-\alpha}{2} = \frac{1-0,05}{2} = 0,475 \to C = 1,96$

Vậy khoảng tin cậy đối xứng của trọng lượng trung bình là:

$$\mu \in \left[\overline{X} - \frac{C\sigma_0}{\sqrt{n}}; \overline{X} + \frac{C\sigma_0}{\sqrt{n}}\right] = [19,248; 20,032].$$

Kết luận: Với độ tin cậy 95%, từ mẫu cụ thể đã cho, trọng lượng trung bình của sản phẩm nói trên nằm trong khoảng [19,248; 20,032].

3.3.3.2. Ước lượng μ trong phân phối $N(\mu,\sigma^2)$, (σ^2 chưa biết)

Để ước lượng trung bình μ trong phân phối chuẩn có phương sai σ^2 chưa biết, ta cũng dùng $X_1, X_2, ..., X_n$ là mẫu độc lập, $X_i \sim N(\mu, \sigma^2)$ để ước lượng μ. Ta có, do định lý Lindeberg - Lévy,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

nên

$$Z = \frac{(\overline{X}-\mu)\sqrt{n}}{\sigma} \sim N(0,1).$$

Ta không thể dùng biến số này để ước lượng μ vì trong biểu thức còn chứa tham số σ chưa biết. Do đó, ta dùng phương sai mẫu S_X^2 làm ước lượng điểm thay thế σ^2, với lưu ý rằng, cũng do định lý Lindeberg-Lévy,

$$Y = \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1).$$

Do Y, Z độc lập nên bằng cách đặt

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{(\overline{X}-\mu)\sqrt{n}}{S_X}$$

thì T có phân phối $St(n-1)$.

Với độ tin cậy bằng $(1-\alpha)$ cho trước tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1+\alpha_2=\alpha$, từ đó tìm được hai giá trị tới hạn Student tương ứng là $a = t_{1-\alpha_1}(n-1)$ và $b = t_{\alpha_2}(n-1)$ thoả mãn: $P(T > a) = 1-\alpha_1$ và $P(T > b) = \alpha_2$.

$$\Leftrightarrow P(a \le T \le b) = 1-(\alpha_1+\alpha_2) = 1-\alpha.$$

$$\Leftrightarrow P(-a \le T \le b) = 1-\alpha.$$

Thay biểu thức của T vào và biến đổi ta được:

$$P\left(\overline{X} - \frac{bS_X}{\sqrt{n}} \le T \le \overline{X} + \frac{aS_X}{\sqrt{n}}\right) = 1-\alpha. \quad (3.26)$$

Với độ tin cậy bằng $(1-\alpha)$ tham số μ của biến ngẫu nhiên gốc X sẽ nằm trong khoảng

$$\left[\overline{X}-\frac{bS_X}{\sqrt{n}};\overline{X}+\frac{aS_X}{\sqrt{n}}\right] \tag{3.27}$$

Biểu thức (3.27) mới chỉ cho ta khoảng tin cậy tổng quát. Với độ tin cậy $(1-\alpha)$ cho trước, có thể tìm được vô số cặp giá trị α_1 và α_2 thỏa mãn điều kiện: $\alpha_1+\alpha_2=\alpha$, từ đó sẽ có vô số khoảng tin cậy tương ứng. Trong thực tế người ta thường dùng các trường hợp đặc biệt sau:

Khoảng tin cậy đối xứng: Chọn $\alpha_1=\alpha_2=\dfrac{\alpha}{2}$, ta có giá trị tới hạn $C=t_{\alpha/2}(n-1)$, khoảng tin cậy của μ là:

$$\left[\overline{X}-\frac{CS_X}{\sqrt{n}};\overline{X}+\frac{CS_X}{\sqrt{n}}\right]. \tag{3.28}$$

Với sai số ước lượng: $\varepsilon=\dfrac{CS_X}{\sqrt{n}}$.

ε được gọi là *sai số* hay *độ chính xác* của ước lượng. Nó phản ánh mức độ sai lệch của trung bình mẫu so với trung bình tổng thể với xác suất $(1-\alpha)$ cho trước.

Khoảng tin cậy phía phải (ước lượng tối thiểu): Chọn $\alpha_1=0$ và $\alpha_2=\alpha$, ta có giá trị tới hạn $C=t_\alpha(n-1)$, khoảng tin cậy của μ là:

$$\left[\overline{X}-\frac{CS_X}{\sqrt{n}};+\infty\right) \tag{3.29}$$

Khoảng tin cậy phía trái (ước lượng tối đa): Chọn $\alpha_1=\alpha$ và $\alpha_2=0$, ta có giá trị tới hạn $C=t_\alpha(n-1)$, khoảng tin cậy của μ là:

$$\left(-\infty;\overline{X}+\frac{CS_X}{\sqrt{n}}\right] \tag{3.30}$$

Trong trường hợp này độ dài khoảng tin cậy I cũng là ngắn nhất khi khoảng tin cậy là đối xứng, khi đó:

$$I=2\varepsilon=2\frac{CS_X}{\sqrt{n}} \tag{3.31}$$

Ví dụ 3.23. Năng suất một loại cây trồng là biến ngẫu nhiên có phân phối chuẩn. Thu hoạch tại một số điểm được kết quả sau:

Năng suất (tấn/ha)	30	31	32	33	34	35
Số điểm	4	5	6	3	4	3

Với độ tin cậy 95%, hãy ước lượng năng suất trung bình bằng khoảng tin cậy đối xứng.

Giải

Gọi X là năng suất loại cây trồng

Từ mẫu đã cho, ta tính được: $n = 25$; $\overline{X} = 32,28$; $S_X = 1,646$

Với độ tin cậy: $\gamma = 0,95$ suy ra mức ý nghĩa $\alpha = 1 - \gamma = 1 - 0,95 = 0,05$. Suy ra giá trị tới hạn $C = t_{0,025}(24) = 2,064$

Vậy khoảng tin cậy đối xứng của năng suất trung bình có dạng:

$$\mu \in \left[\overline{X} - \frac{CS_X}{\sqrt{n}}; \overline{X} + \frac{CS_X}{\sqrt{n}}\right] = [31,6; 32,96].$$

Kết luận: Với độ tin cậy 95%, từ mẫu cụ thể đã cho, năng suất trung bình của loại cây trồng nói trên nằm trong khoảng [31,6; 32,96].

Ví dụ 3.24. Để xác định kích thước trung bình của chi tiết do một máy sản xuất người ta lấy ngẫu nhiên 200 chi tiết để đo kích thước và thu được bảng số liệu sau:

Kích thước chi tiết (cm)	Số chi tiết tương ứng
54,795 – 54,805	6
54,805 – 54,815	14
54,815 – 54,825	33
54,825 – 54,835	47
54,835 – 54,845	45
54,845 – 54,855	33
54,855 – 54,865	15
54,865 – 54,875	7

Giả thiết kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn.

a) Với độ tin cậy 95% hãy ước lượng bằng khoảng tin cậy đối xứng kích thước trung bình của chi tiết do máy đó sản xuất.

b) Với độ tin cậy 95% hãy ước lượng bằng khoảng tin cậy tối thiểu và tối đa kích thước trung bình của chi tiết do máy đó sản xuất.

Giải

Gọi X là kích thước chi tiết do máy đó sản xuất.

Ta có $n = 200;\ \overline{X} = 54,8353;\ S_X = 0,0164$

a) Đây là bài toán ước lượng tham số μ của biến ngẫu nhiên phân phối chuẩn bằng khoảng tin cậy đối xứng trong trường hợp chưa biết σ^2.

Với độ tin cậy: $\gamma = 0,95$ suy ra mức ý nghĩa $\alpha = 1 - \gamma = 1 - 0,95 = 0,05$. Suy ra giá trị tới hạn $C = t_{0,025}(199) = 1,96$

Khoảng tin cậy đối xứng kích thước giá trị trung bình là:

$$\mu \in \left[\overline{X} - \frac{CS_X}{\sqrt{n}};\ \overline{X} + \frac{CS_X}{\sqrt{n}}\right] = [54,83294; 54,83752].$$

b) Khoảng ước lượng bằng khoảng tin cậy tối thiểu và tối đa kích thước trung bình của chi tiết do máy đó sản xuất

Với độ tin cậy: $\gamma = 0,95$ suy ra mức ý nghĩa $\alpha = 1 - \gamma = 1 - 0,95 = 0,05$. Suy ra giá trị tới hạn $C = t_{0,05}(199) = 1,64$

Khoảng ước lượng bằng khoảng tin cậy tối thiểu kích thước giá trị trung bình là:

$$\mu \in \left[\overline{X} - \frac{CS_X}{\sqrt{n}};\ +\infty\right) = [54,8334;\ +\infty).$$

Khoảng ước lượng bằng khoảng tin cậy tối đa kích thước giá trị trung bình là:

$$\mu \in \left(-\infty;\ \overline{X} + \frac{CS_X}{\sqrt{n}}\right] = (-\infty;\ 54,8372].$$

3.3.4. Bài toán ước lượng khoảng tin cậy cho phương sai

3.3.4.1. Ước lượng σ^2 trong phân phối $N(\mu_0, \sigma^2)$, (μ_0 biết).

Để ước lượng phương sai σ^2 trong phân phối chuẩn có trung bình μ_0 biết, ta cũng dùng mẫu độc lập $X_1, X_2, \ldots X_n \sim N(\mu_0, \sigma^2)$ để ước lượng σ^2. Trước hết, lưu ý rằng

$$X_i \sim N(\mu_0, \sigma^2) \Rightarrow \frac{X_i - \mu_0}{\sigma} \sim N(0,1)$$

và

$$\left(\frac{X_i - \mu_0}{\sigma}\right)^2 \sim \chi^2(1).$$

Chọn thống kê:

$$Y = \sum_{i=1}^{n} \frac{(X_i - \mu_0)^2}{\sigma^2} \sim \chi^2(n)$$

Với độ tin cậy $1-\alpha$ cho trước, ta tìm được α_1 và α_2 sao cho $\alpha_1+\alpha_2=\alpha$. Từ đó tìm được hai giá trị tới hạn chi bình phương $a=\chi^2_{1-\alpha_1}(n)$ và $b=\chi^2_{\alpha_2}(n)$ sao cho:

$$P(a \le Y \le b) = 1-\alpha$$

$$\Leftrightarrow P\left[a \le \sum_{i=1}^{n}\frac{(X_i-\mu_0)^2}{\sigma^2} \le b\right] = 1-\alpha$$

$$\Leftrightarrow P\left[\frac{1}{b}\sum_{i=1}^{n}(X_i-\mu_0)^2 \le \sigma^2 \le \frac{1}{a}\sum_{i=1}^{n}(X_i-\mu_0)^2\right] = 1-\alpha.$$

Như vậy, với độ tin cậy bằng $(1-\alpha)$ tham số σ^2 của biến ngẫu nhiên gốc X sẽ nằm trong khoảng

$$\left[\frac{1}{b}\sum_{i=1}^{n}(X_i-\mu_0)^2; \frac{1}{a}\sum_{i=1}^{n}(X_i-\mu_0)^2\right] \tag{3.32}$$

Biểu thức (3.32) mới chỉ cho ta khoảng tin cậy tổng quát. Với độ tin cậy $(1-\alpha)$ cho trước, có thể tìm được vô số cặp giá trị α_1 và α_2 thỏa mãn điều kiện: $\alpha_1+\alpha_2=\alpha$, từ đó sẽ có vô số khoảng tin cậy tương ứng. Trong thực tế người ta thường dùng các trường hợp đặc biệt sau:

Khoảng tin cậy hai phía: Chọn $\alpha_1=\alpha_2=\dfrac{\alpha}{2}$ suy ra giá trị tới hạn $a=\chi^2_{1-\frac{\alpha}{2}}(n)$ và $b=\chi^2_{\frac{\alpha}{2}}(n)$, khoảng tin cậy của σ^2 là:

$$\left[\frac{1}{b}\sum_{i=1}^{n}(X_i-\mu_0)^2; \frac{1}{a}\sum_{i=1}^{n}(X_i-\mu_0)^2\right] \tag{3.33}$$

Khoảng tin cậy phía phải (ước lượng tối thiểu): Chọn $\alpha_1=0$ và $\alpha_2=\alpha$, khoảng tin cậy của σ^2 là:

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{\alpha}(n)}; +\infty\right) \tag{3.34}$$

Khoảng tin cậy phía trái (ước lượng tối đa): Chọn $\alpha_1=\alpha$ và $\alpha_2=0$, khoảng tin cậy của σ^2 là:

$$\left(-\infty\ ;\ \frac{\sum_{i=1}^{n}\left(X_i-\mu_0\right)^2}{\chi^2_{1-\alpha}(n)}\right). \tag{3.35}$$

3.3.4.2. Ước lượng σ^2 trong phân phối $N\left(\mu,\sigma^2\right)$, ($\mu$ chưa biết).

Để ước lượng phương sai σ^2 trong phân phối chuẩn có trung bình μ chưa biết, ta cũng dùng mẫu độc lập $X_1, X_2, ... X_n \sim N\left(\mu,\sigma^2\right)$ để ước lượng σ^2.

Ta có, do định lý Lindeberg-Lévy

$$Y=\frac{(n-1)S_X^2}{\sigma^2}\sim\chi^2(n-1).$$

Với độ tin cậy $1-\alpha$ cho trước, ta tìm được α_1 và α_2 sao cho $\alpha_1+\alpha_2=\alpha$. Từ đó tìm được hai giá trị tới hạn chi bình phương $a=\chi^2_{1-\alpha_1}(n-1)$ và $b=\chi^2_{\alpha_2}(n-1)$ sao cho:

$$P(a\ \le Y\le\ b)=1-\alpha$$

$$\Leftrightarrow P\left[a\le\frac{(n-1)S_X^2}{\sigma^2}\le b\right]=1-\alpha$$

$$\Leftrightarrow P\left[\frac{(n-1)S_X^2}{b}\le\sigma^2\le\frac{(n-1)S_X^2}{a}\right]=1-\alpha.$$

Như vậy, với độ tin cậy bằng $(1-\alpha)$ tham số σ^2 của biến ngẫu nhiên gốc X sẽ nằm trong khoảng

$$\left[\frac{(n-1)S_X^2}{b};\frac{(n-1)S_X^2}{a}\right] \tag{3.36}$$

Biểu thức (3.36) mới chỉ cho ta khoảng tin cậy tổng quát. Với độ tin cậy $(1-\alpha)$ cho trước, có thể tìm được vô số cặp giá trị α_1 và α_2 thỏa mãn điều kiện: $\alpha_1+\alpha_2=\alpha$, từ đó sẽ có vô số khoảng tin cậy tương ứng. Trong thực tế người ta thường dùng các trường hợp đặc biệt sau:

Khoảng tin cậy hai phía: Chọn $\alpha_1=\alpha_2=\frac{\alpha}{2}$, suy ra hai giá trị tới hạn $a=\chi^2_{1-\frac{\alpha}{2}}(n-1)$ và $b=\chi^2_{\frac{\alpha}{2}}(n-1)$, khoảng tin cậy của σ^2 là:

$$\left[\frac{(n-1)S_X^2}{b};\frac{(n-1)S_X^2}{a}\right] \tag{3.37}$$

Khoảng tin cậy phía phải (ước lượng tối thiểu): Chọn $\alpha_1 = 0$ và $\alpha_2 = \alpha,$ khoảng tin cậy của σ^2 là:

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{\alpha}(n-1)};+\infty\right) \tag{3.38}$$

Khoảng tin cậy phía trái (ước lượng tối đa): Chọn $\alpha_1 = \alpha$ và $\alpha_2 = 0,$ khoảng tin cậy của σ^2 là:

$$\left(-\infty;\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{1-\alpha}(n-1)}\right) \tag{3.39}$$

Ví dụ 3.25. Khảo sát chỉ tiêu X của một loại sản phẩm, người ta quan sát một mẫu và có kết quả 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

Giả sử X có phân phối chuẩn. Hãy ước lượng phương sai của X với độ tin cậy 95% trong các trương hợp sau:

a) Biết trung bình tổng thể của X là 25 cm.

b) Chưa biết giá trị trung bình của X.

Giải

a) Để ước lượng phương sai tổng thể khi biết trung bình tổng thể $\mu_0 = 25,$ ta dùng thống kê

$$Y = \sum_{i=1}^{n}\frac{(X_i-\mu_0)^2}{\sigma^2} \sim \chi^2(n).$$

Với độ tin cậy $\gamma = 95\%,$ ta chọn khoảng tin cậy cho Y : $[a,b]$

Với $a = \chi^2_{0,975}(100) = 77,929;\ a = \chi^2_{0,025}(100) = 129,561$

Ta lập bảng

$X_i-\mu_0$	-12	-8	-4	0	4	8	12

n	8	9	20	16	16	13	18

Từ đó ta tìm được cỡ mẫu $n = 100; \sum_{i=1}^{k} n_i (X_i - \mu_0)^2 = 5728$

Khoảng ước lượng của phương sai tổng thể σ^2

$$\sigma^2 \in \left[\frac{1}{b} \sum_{i=1}^{n} (X_i - \mu_0)^2 ; \frac{1}{a} \sum_{i=1}^{n} (X_i - \mu_0)^2 \right] = [44,211; 73,503].$$

b) Để ước lượng phương sai tổng thể khi chưa biết trung bình của tổng thể, ta dùng thống kê

$$Y = \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1),$$

Với độ tin cậy $\gamma = 95\%,$ ta chọn khoảng tin cậy cho $Y : [a,b]$

Với $a = \chi^2_{0,975}(99) = 77,929; a = \chi^2_{0,025}(99) = 129,561$

Khoảng ước lượng của phương sai tổng thể σ^2

$$\sigma^2 \in \left[\frac{(n-1)S_X^2}{b} ; \frac{(n-1)S_X^2}{a} \right] = [42,784; 71,130].$$

3.3.5. Bài toán ước lượng khoảng tin cậy cho tỷ lệ

Để ước lượng tỷ lệ p trong phân phối Bernoulli, ta cũng dùng mẫu độc lập $X_1, X_2, ..., X_n \sim B(1;p)$ để ước lượng p. Đặt

$$f = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

thì

$$Z = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \sim N(0,1)$$

Với độ tin cậy $\gamma = 1 - \alpha$ cho trước, từ bảng phân phối Gauss, ta tìm được C sao cho

$$P(-C \le Z \le C) = 1 - \alpha$$

$$\Leftrightarrow P\left(-C \le \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}} \le C \right) = 1 - \alpha$$

$$\Leftrightarrow P\left(f - C\sqrt{\frac{p(1-p)}{n}} \le p \le f + C\sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha$$

Biểu thức này chưa cho ta khoảng ước lượng của p vì các đầu mút còn lệ thuộc vào p chưa biết. Do đó, ta thay p bằng ước lượng điểm, $\overline{p} = f$, của nó và ta có

Khoảng ước lượng đối xứng : Với độ tin cậy $\gamma = 1 - \alpha$ cho trước, ta tìm được giá trị tới hạn C thỏa mãn $\phi_0(C) = \frac{1-\alpha}{2}$, khoảng tin cậy đối xứng của p là:

$$p \in \left[f - C\sqrt{\frac{f(1-f)}{n}} ; f + C\sqrt{\frac{f(1-f)}{n}} \right]. \tag{3.40}$$

Khoảng tin cậy phía phải (ước lượng tối thiểu): Với độ tin cậy $\gamma = 1 - \alpha$ cho trước, ta tìm được giá trị tới hạn C thỏa mãn $\phi_0(C) = \frac{1-2\alpha}{2}$, khoảng tin cậy tối tiểu của p là:

$$p \in \left[f - C\sqrt{\frac{f(1-f)}{n}} \; ; 1 \right] \tag{3.41}$$

Khoảng tin cậy phía trái (ước lượng tối đa): Với độ tin cậy $\gamma = 1 - \alpha$ cho trước, ta tìm được giá trị tới hạn C thỏa mãn $\phi_0(C) = \frac{1-2\alpha}{2}$, khoảng tin cậy tối đa của p là:

$$p \in \left[0 \; ; \; f + C\sqrt{\frac{f(1-f)}{n}} \right] \tag{3.42}$$

3.4. Bài toán xác định cỡ mẫu

Ta cũng có thể giải quyết bài toán tìm cỡ mẫu n khi muốn ước lượng tỷ lệ p với độ tin cậy γ và sai số ε_0 cho trước như sau

$$C\sqrt{\frac{f(1-f)}{n}} \le \varepsilon_0 \Leftrightarrow n \ge \frac{C^2 f(1-f)}{\varepsilon_0^2} \tag{3.43}$$

Xét rằng hàm số $y = x(1-x)$, hàm số này đạt giá trị lớn nhất là $\frac{1}{4}$ tại điểm $x = \frac{1}{2}$, cho nên từ sai số ước lượng ta có thể viết

$$C\sqrt{\frac{f(1-f)}{n}} \le C\sqrt{\frac{1}{4n}} \le \varepsilon_0 \Leftrightarrow n \ge \frac{C^2}{4\varepsilon_0^2}$$

Do đó, ta có thể chọn cỡ mẫu

$$n \ge \frac{C^2}{4\varepsilon_0^2}. \tag{3.44}$$

Vậy

- Nếu biết tỷ lệ mẫu f, ta tìm n thỏa mãn $n \ge \dfrac{C^2 f(1-f)}{\varepsilon_0^2}$.

- Nếu chưa biết tỷ lệ mẫu f, ta tìm n thỏa mãn $n \ge \dfrac{C^2}{4\varepsilon_0^2}$.

Ví dụ 3.26. Để ước lượng tỷ lệ bệnh sốt rét ở Đồng Bằng Sông Cửu Long.

1) Nếu muốn sai số ước lượng không quá 2% ở độ tin cậy 95% thì cần quan sát ít nhất bao nhiêu người ?

2) Ta quan sát ngẫu nhiên 200 người, thấy có 24 người mắc bệnh sốt rét.

a) Tìm khoảng ước lượng đối xứng của p với độ tin cậy 95%,

b) Tìm khoảng ước lượng tối đa, tối thiểu của p với độ tin cậy 95%.

c) Nếu muốn sai số ước lượng không quá 2% ở độ tin cậy 95% thì cần quan sát ít nhất mấy người ?

Giải

1) Với độ tin cậy $\gamma = 1-\alpha = 0,95 \to \alpha = 0,05$.

Ta có $\phi_0(C) = \dfrac{1-\alpha}{2} = 0,475 \to C = 1,96$ và với $\varepsilon_0 = 0,02$ nên áp dụng công thức (3.44), ta có

$$n \ge \frac{C^2}{4\varepsilon_0^2} = \frac{(1,96)^2}{4(0,02)^2} = 2401.$$

Vậy cần quan sát ít nhất 2041 người.

2) Ta có $n = 200$, $k = 24$, $f = \dfrac{24}{200} = 0,12$,

a) Khoảng ước lượng đối xứng của p.

Với độ tin cậy $\gamma = 95\% = 0,95$ suy ra mức ý nghĩa $\alpha = 5\% = 0,05$.

Ta có $\phi_0(C) = \dfrac{1-\alpha}{2} = 0,475 \to C = 1,96$

Ta có ước lượng của tỷ lệ p là

$$p \in \left[f - C\sqrt{\frac{f(1-f)}{n}}; f + C\sqrt{\frac{f(1-f)}{n}} \right] = [0,075; 0,165].$$

b) Khoảng ước lượng tối đa, tối thiểu của p

Với độ tin cậy $\gamma = 95\% = 0,95$ suy ra mức ý nghĩa $\alpha = 5\% = 0,05$.

Ta có $\phi_0(C) = \dfrac{1-2\alpha}{2} = 0,45 \to C = 1,64$

Khoảng ước lượng tối đa của p

$$p \in \left[0\ ;\ f + C\sqrt{\frac{f(1-f)}{n}}\right] = [0;\ 0,1577]$$

Khoảng ước lượng tối thiểu của p

$$p \in \left[f - C\sqrt{\frac{f(1-f)}{n}};\ 1\right] = [0,0823;\ 1].$$

c) Ta có $\phi_0(C) = \dfrac{1-\alpha}{2} = 0,475 \to C = 1,96$ và với $\varepsilon_0 = 0,02$ nên áp dụng công thức (3.43), ta có

$$n \geq \frac{C^2 f(1-f)}{\varepsilon_0^2} = 1014,18.$$

Vậy cần quan sát ít nhất 1015 người.

3.5. Tóm tắt chương 3

1. Các đặc trưng đo lường khuynh hướng tập trung

i) Số trung bình mẫu: $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{k} f_i x_i,\ n = f_1 + f_2 + \cdots + f_k.$

ii) Mốt (Mode): Mốt là lượng biến có tần số lớn nhất.

+) Đối với đại lượng biến có khoảng cách tổ

$$M_0 = x_{M_{0(\min)}} + d_{M_0}\frac{F_{M_0} - F_{M_{0-1}}}{\left(F_{M_0} - F_{M_{0-1}}\right) + \left(F_{M_0} - F_{M_{0+1}}\right)}$$

ii) Số trung vị:

+) Trường hợp số đơn vị tổng thể lẻ ($n = 2m+1$). Số trung vị sẽ là đại lượng ở vị trí thứ ($m+1$): $Me = X_{m+1}$.

+) Trường hợp số đơn vị tổng thể chẵn ($n = 2m$). Số trung vị sẽ là đại lượng giữa vị trí thứ (m) và ($m+1$): $Me = \dfrac{X_m + X_{m+1}}{2}$.

Đối với dãy số có lượng biến là khoảng cách tổ

$$M_e = X_{M_{e(\min)}} + d_{M_e}\frac{\frac{1}{2}\sum f_i - S_{M_{e-1}}}{f_{M_e}}$$

2. Các đặc trưng đo lường khuynh hướng phân tán

i) Phương sai hiệu chỉnh: $S_X^2 = \dfrac{1}{n-1}\sum_{i=1}^{k} f_i\left(x_i - \overline{x}\right)^2.$

ii) Độ lệch chuẩn có hiệu chỉnh: $S_X = \sqrt{S_X^2}.$

ii) Tứ phân vị (Quartiles): là chia dãy số thành 4 phần, mỗi phần có số đơn vị bằng nhau.

Cách xác định tứ phân vị cho tài liệu phân tổ có khoảng cách tổ

Tổ chứa phân vị thứ i có tần số tích lũy $\geq \dfrac{n+1}{4}i.$

Tứ phân vị thứ 1: $Q_1 = X_{Q_{1(min)}} + d_{Q_1}\dfrac{\frac{1}{4}\sum f_i - S_{Q_{1-1}}}{f_{Q_1}}.$

Tứ phân vị thứ 3 : $Q_3 = X_{Q_{3(min)}} + d_{Q_3}\dfrac{\frac{3}{4}\sum f_i - S_{Q_{3-1}}}{f_{Q_3}}.$

3. Bài toán ước lượng khoảng tin cậy cho giá trị trung bình

i) Ước lượng μ trong phân phối $N\left(\mu, \sigma_0^2\right)$, ($\sigma_0^2$ biết)

+) Với mức ý nghĩa α. Ta có : $\phi_0(C) = \dfrac{1-\alpha}{2}$

Khoảng tin cậy đối xứng: $\left[\overline{X} - \dfrac{C\sigma_0}{\sqrt{n}}; \overline{X} + \dfrac{C\sigma_0}{\sqrt{n}}\right].$

+) Với mức ý nghĩa α. Ta có : $\phi_0(C) = \dfrac{1-2\alpha}{2}$

Khoảng tin cậy phía phải: $\left[\overline{X} - \dfrac{C\sigma_0}{\sqrt{n}}; +\infty\right).$

Khoảng tin cậy phía trái: $\left(-\infty; \overline{X} + \dfrac{C\sigma_0}{\sqrt{n}}\right].$

ii) Ước lượng μ trong phân phối $N\left(\mu, \sigma^2\right)$, ($\sigma^2$ chưa biết)

+) Với mức ý nghĩa α. Ta có: $C = t_{\alpha/2}(n-1)$

Khoảng tin cậy đối xứng: $\left[\overline{X} - \dfrac{CS_X}{\sqrt{n}}; \overline{X} + \dfrac{CS_X}{\sqrt{n}}\right].$

+) Với mức ý nghĩa α. Ta có: $C = t_{\alpha}(n-1)$

Khoảng tin cậy phía phải: $\left[\overline{X}-\frac{CS_X}{\sqrt{n}};+\infty\right)$.

Khoảng tin cậy phía trái: $\left(-\infty;\overline{X}+\frac{CS_X}{\sqrt{n}}\right]$.

4. Bài toán ước lượng khoảng tin cậy cho phương sai

i) Ước lượng σ^2 trong phân phối $N\left(\mu_0,\sigma^2\right)$, ($\mu_0$ biết).

+) Với mức ý nghĩa α, ta có $a=\chi^2_{1-\frac{\alpha}{2}}\left(n\right)$ và $b=\chi^2_{\frac{\alpha}{2}}\left(n\right)$

Khoảng tin cậy hai phía: $\left[\frac{1}{b}\sum_{i=1}^{n}\left(X_i-\mu_0\right)^2;\frac{1}{a}\sum_{i=1}^{n}\left(X_i-\mu_0\right)^2\right]$.

+) Với mức ý nghĩa α, ta có $\chi^2_\alpha(n)$

Khoảng tin cậy phía phải: $\left[\frac{\sum_{i=1}^{n}\left(X_i-\mu_0\right)^2}{\chi^2_\alpha(n)};+\infty\right)$.

Khoảng tin cậy phía trái: $\left(-\infty;\frac{\sum_{i=1}^{n}\left(X_i-\mu_0\right)^2}{\chi^2_{1-\alpha}(n)}\right)$.

ii) Ước lượng σ^2 trong phân phối $N\left(\mu,\sigma^2\right)$, ($\mu$ chưa biết).

+) Với mức ý nghĩa α, ta có $a=\chi^2_{1-\frac{\alpha}{2}}\left(n-1\right)$ và $b=\chi^2_{\frac{\alpha}{2}}\left(n-1\right)$

Khoảng tin cậy hai phía: $\left[\frac{(n-1)S_X^2}{b};\frac{(n-1)S_X^2}{a}\right]$.

+) Với mức ý nghĩa α, ta có $\chi^2_\alpha(n-1)$

Khoảng tin cậy phía phải: $\left(\frac{\sum_{i=1}^{n}\left(X_i-\mu_0\right)^2}{\chi^2_\alpha(n-1)};+\infty\right)$.

Khoảng tin cậy phía : $\left(-\infty\ ;\ \dfrac{\sum_{i=1}^{n}\left(X_i-\mu_0\right)^2}{\chi^2_{1-\alpha}(n-1)}\right)$.

5. Bài toán ước lượng khoảng tin cậy cho tỷ lệ

i) Với mức ý nghĩa α. Ta có : $\phi_0\left(C\right)=\dfrac{1-\alpha}{2}$

Khoảng ước lượng đối xứng : $p\in\left[f-C\sqrt{\dfrac{f\left(1-f\right)}{n}};f+C\sqrt{\dfrac{f\left(1-f\right)}{n}}\right]$.

ii) Với mức ý nghĩa α. Ta có : $\phi_0\left(C\right)=\dfrac{1-2\alpha}{2}$

Khoảng tin cậy phía phải: $p\in\left[f-C\sqrt{\dfrac{f\left(1-f\right)}{n}}\ ;\ 1\right]$.

Khoảng tin cậy phía trái: $p\in\left[0\ ;\ f+C\sqrt{\dfrac{f\left(1-f\right)}{n}}\right]$.

3.6. Bài tập

Bài số 1. Quan sát thời gian cần thiết để sản xuất một chi tiết máy, ta thu được số liệu cho bảng sau

Khoảng thời gian (phút)	Số lần quan sát
20 – 25	2
25 – 30	14
30 – 35	26
35 – 40	32
40 – 45	14
45 – 50	8
50 – 55	4

Tính trung bình mẫu $\overline{X}$, phương sai mẫu có hiệu chỉnh S^2_X.

Đáp số: $\overline{X}=36,6;\ S^2_X=45,1414$.

Bài số 2. Có tài liệu phân tổ về năng suất lao động của công nhân một doanh nghiệp trong kỳ nghiên cứu như sau:

Năng suất lao động (Sp/ca)	Số công nhân

20 – 22	10
22 – 24	40
24 – 26	80
26 – 28	50
28 – 30	20

Hãy tính trung bình, phương sai có hiệu chỉnh, mốt, tứ phân vị của năng suất lao động một công nhân.

Đáp số: 35,3; 4,131; 25,143; 24; 25,25; 26,8.

Bài số 3. Có tài liệu dưới đây của một doanh nghiệp

Năng suất lao động (kg)	Số công nhân
110 – 120	10
120 – 130	30
130 – 140	50
140 – 150	60
150 – 160	145
160 – 170	110
170 – 180	80
180 – 190	15

Hãy tính trung bình, phương sai có hiệu chỉnh, mốt, tứ phân vị của năng suất lao động một công nhân.

Đáp số: 155,5; 251,253; 157,083; 145,83; 156,9; 167,3.

Bài số 4. Chiều dài của một loại sản phẩm được xuất khẩu hàng loạt là biến ngẫu nhiên phân phối chuẩn với $\mu = 100\text{mm}; \sigma^2 = 4^2\text{mm}^2$. Kiểm tra ngẫu nhiên 25 sản phẩm. Khả năng chiều dài trung bình của số sản phẩm kiểm tra nằm trong khoảng từ 98mm đến 101mm là bao nhiêu.

Đáp số: 0,8828.

Bài số 5. Đo độ dài của một loại trục xe, ta có kết quả

Nhóm	18,4-18,6	18,6-18,8	18,8 -19	19 -19,2	19,2-19,4	19,4-19,6	19,6-19,8
n_i	1	4	20	41	19	8	4

Hãy ước lượng điểm độ dài trung bình và phương sai của trục xe.

Đáp số: $\overline{X} = 19{,}133; S_X^2 = 0{,}539$.

Bài số 5. Đo sức bền chịu lực của 1 loại ống thí nghiệm, người ta thu được bộ số liệu sau

4500	6500	5200	4800	4900	5125	6200	5375

Từ kinh nghiệm nghề nghiệp, người ta cũng biết rằng sức bền đó có phân phối chuẩn với độ lệch chuẩn $\sigma = 300$. Hãy ước lượng sức bền trung bình của loại ống trên, với độ tin cậy 90%.

Đáp số: $\mu \in [5151,0517;5498,9483]$.

Bài số 7. Trên tập mẫu gồm 100 số liệu, người ta tính được $\overline{X} = 0,1; S_X = 0,014$. Hãy ước lượng giá trị trung bình tổng thể, với độ tin cậy 95%.

Đáp số: $\mu \in [0,0973;0,1027]$.

Bài số 8. Chọn ngẫu nhiên 36 công nhân của xí nghiệp thì thấy lương trung bình là 380 ngàn đ/tháng. Giả sử lương công nhân tuân theo luật chuẩn với $\sigma = 14$ ngàn đồng. Với độ tin cậy 95%, hãy ước lượng mức lương trung bình của công nhân trong toàn xí nghiệp.

Đáp số: $\mu \in [375,4267;384,5733]$.

Bài số 9. Điểm trung bình môn toán của 100 thí sinh dự thi vào ĐHKT là 5 với độ lệch chuẩn mẫu đã điều chỉnh $S_X = 2,5$.

a) Ước lượng điểm trung bình môn toán của toàn thể thí sinh với độ tin cậy là 95%.

b) Với sai số 0,25 điểm. Hãy xác định độ tin cậy.

Đáp số: a) $\mu \in [4,51;5,49]$*; b) 68,26%.*

Bài số 10. Tuổi thọ của một loại bóng đèn được biết theo quy luật chuẩn với độ lệch chuẩn 100 giờ.

a) Chọn ngẫu nhiên 100 bóng đèn để thử nghiệm, thấy mỗi bóng tuổi thọ trung bình là 1000 giờ. Hãy ước lượng tuổi thọ trung bình của bóng đèn xí nghiệp A sản xuất với độ tin cậy là 95%.

b) Với độ chính xác là 15 giờ. Hãy xác định độ tin cậy.

c) Với độ chính xác là 25 giờ và độ tin cậy là 95% thì cần thử nghiệm bao nghiêu bóng.

Đáp số: a) $\mu \in [980,4;1019,6]$*; b) 86,64%; c) 62.*

Bài số 11. Khối lượng các bao bột mì tại một cửa hàng lương thực theo quy luật chuẩn. Kiểm tra 20 bao, thấy khối lượng trung bình của mỗi bao bột mì là 48kg, và phương sai mẫu có điều chỉnh là $S_X^2 = (0,5\text{kg})^2$.

a) Với độ tin cậy 95%, hãy ước lượng khối lượng trung bình của một bao bột mì thuộc cửa hàng.

b) Với độ chính xác là 0,26kg. Hãy xác định độ tin cậy.

c) Với độ chính xác là 160g và độ tin cậy là 95%, tính cỡ mẫu.

Đáp số: a) $\mu \in [47,766; 48,234]$*; b) 97%; c) 43.*

Bài số 12. Để ước lượng tỷ lệ sản phẩm xấu của một kho đồ hộp, người ta kiểm tra ngẫu nhiên 100 hộp thấy có 11 hộp xấu.

a) Ước lượng tỷ lệ sản phẩm xấu của kho đồ hộp với độ tin cậy 94%.

b) Với sai số cho phép $\varepsilon = 3\%$, hãy xác định độ tin cậy.

Đáp số: a) $p \in [0,051; 0,169]$*; b) 66,3%.*

Bài số 13. a) Muốn ước lượng tỷ lệ bệnh sốt xuất huyết ở Tp. Hồ Chí Minh với sai số không quá 3% ở độ tin cậy 95% thì phải quan sát ít nhất bao nhiêu người ?

b) Giả sử quan sát 100 người thấy có 20 người bị bệnh sốt xuất huyết. Hãy ước lượng tỷ lệ bệnh sốt xuất huyết ở Tp. Hồ Chí Minh ở độ tin cậy 97%. Nếu muốn sai số ước lượng không quá 3% ở độ tin cậy 95% thì phải quan sát ít nhất bao nhiêu người ?

Đáp số: a) 1068 người; b) $p \in [0,1132; 0,2868]$*; 683 người.*

Bài số 14. Để ước lượng xác suất mắc bệnh gan với độ tin cậy 90% và sai số không vượt quá 2% thì cần phải khám ít nhất bao nhiêu người, biết rằng tỷ lệ mắc bệnh gan thực nghiệm đã cho bằng 0,9.

Đáp số: 606 người.

Bài số 15. Trước bầu cử, người ta phỏng vấn ngẫu nhiên 2000 cử tri thì thấy có 1380 người ủng hộ một ứng cử viên K. Với độ tin cậy 95%, hỏi ứng cử viên đó thu được tối thiểu bao nhiêu phần trăm phiếu bầu ?

Đáp số: 67,3%.

Bài số 16. Gieo thử 400 hạt giống thì thấy có 20 hạt không nảy mầm. Tỷ lệ hạt giống không nảy mầm tối đa là bao nhiêu, với độ tin cậy 95%.

Đáp số: 6,78715%.

Bài số 17. Muốn biết trong hồ có bao nhiêu cá, người ta bắt lên 2000 con, đánh dấu xong lại thả xuống hồ. Sau một thời gian, người ta bắt lên 500 con và thấy có 20 con cá có đánh dấu của lần bắt trước. Dựa vào kết quả đó, hãy ước lượng số cá có trong hồ với độ tin cậy 95%.

Đáp số: $[34978; 87642]$.

Bài số 18. Để có thể dự đoán được số lượng chim thường nghỉ tại vườn nhà mình, người chủ bắt 89 con, đem đeo khoen cho chúng rồi thả đi. Sau một thời gian, ông bắt ngẫu nhiên được 120 con và thấy có 7 con có đeo khoen. Hãy dự đoán số chim giúp ông chủ vườn ở độ tin cậy 99%.

Đáp số: $[785;28525]$.

Bài số 19. Sản lượng mỗi ngày của một phân xưởng là biến ngẫu nhiên tuân theo luật chuẩn. Kết quả thống kê của 9 ngày cho ta :

27	26	21	28	25	30	26	23	26

Hãy ước lượng sản lượng trung bình và phương sai mỗi ngày, với độ tin cậy 95%.

Đáp số: $\mu \in [23,7522;27,8034]$; $\sigma^2 \in [3,1682;25,4840]$.

Bài số 20. Cân thử 100 quả cam, ta có bộ số liệu sau :

Khối lượng (g)	32	33	34	35	36	37	38	39	40
Số quả	2	3	15	26	28	6	8	8	4

a) Hãy ước lượng khối lượng trung bình các quả cam ở độ tin cậy 95%.

b) Cam có khối lượng dưới 34g được coi là cam loại 2. Tìm ước lượng tỷ lệ cam loại 2 với độ tin cậy 90% .

Đáp số: a) $\mu \in [35,539;36,241]$*; b)* $p \in [0,0143;0,0857]$.

Bài số 21. Lô trái cây của một chủ cửa hàng được đóng thành sọt mỗi sọt 100 trái. Kiểm tra 50 sọt thấy có 450 trái không đạt tiêu chuẩn.

a) Ước lượng tỷ lệ trái cây không đạt tiêu chuẩn của lô hàng với độ tin cậy 95%.

b) Muốn ước lượng tỷ lệ trái cây không đạt tiêu chuẩn với độ chính xác 0,5%, độ tin cậy đạt được là bao nhiêu.

c) Muốn ước lượng tỷ lệ trái cây không đạt tiêu chuẩn với độ tin cậy 99% và độ chính xác 1% thì cần kiểm tra bao nhiêu sọt.

d) Muốn ước lượng tỷ lệ trái cây không đạt tiêu chuẩn với độ tin cậy 99% thì độ chính xác đạt được là bao nhiêu?

Đáp số: a) $p \in [0,082;0,098]$*; b) 78,5%; c) 0,012; d) 55 sọt.*

Bài số 22. Điều tra năng suất lúa trên diện tích 100 hec ta trồng lúa của một vùng, ta thu được bảng số liệu sau :

Năng suất (tạ/ha)	41	44	45	46	48	52	54
Diện tích (ha)	10	20	30	15	10	10	5

a) Hãy ước lượng năng suất lúa trung bình của vùng đó với độ tin cậy 95%?

b) Những thửa ruộng có năng suất từ 48tạ/ha trở lên là những thửa có năng suất cao. Hãy ước lượng tỉ lệ diện tích có năng suất cao trong vùng với độ tin cậy 97%.

Đáp số: a) $\mu \in [45,353; 46,647]$; *b)* $p \in [0,156; 0,344]$.

Bài số 23. Đo đường kính của 100 chi tiết do một máy sản xuất kết quả cho ở bảng sau :

Đường kính (mm)	Số chi tiết
19,80 - 19,85	3
19,85 - 19,90	5
19,90 - 19,95	16
19,95 - 20,00	28
20,00 - 20,05	23
20,05 - 20,10	14
20,10 - 20,15	7
20,15 - 20,20	4

Quy định những chi tiết có đường kính 19,9mm đến 20,1mm là những chi tiết đạt tiêu chuẩn.

a) Ước lượng đường kính trung bình của những chi tiết đạt tiêu chuẩn với độ tin cậy 95%.

b) Ước lượng tỷ lệ chi tiết đạt tiêu chuẩn với độ tin cậy 95%.

c) Muốn ước lượng đường kính trung bình của chi tiết đạt tiêu chuẩn muốn độ chính xác đạt 0,02mm và khi ước lượng tỷ lệ chi tiết đạt tiêu chuẩn muốn độ chính xác là 5%, với cùng độ tin cậy là 99% thì cần đo thêm bao nhiêu chi tiết nữa.

Đáp số: a) $\mu \in [19,986; 20,008]$; *b)* $p \in [0,733; 0,887]$; *c) 310.*

Bài số 24. Kích thước của một chi tiết máy là một đại lượng ngẫu nhiên có phân phối chuẩn. Trong một mẫu gồm 30 chi tiết máy được kiểm tra, ta tính được $\overline{X} = 0,47\text{cm}$ và $S_X = 0,032\text{ cm}$. Tìm khoảng tin cậy cho phương sai và trung bình chuẩn của kích thước của toàn bộ các chi tiết máy với độ tin cậy 95%.

Đáp số: $\mu \in [0,482; 0,458]$; $\sigma^2 \in [0,00065; 0,00185]$.

Bài số 25. Lấy 28 mẫu xi măng của một nhà máy sản xuất xi măng để kiểm tra. Kết quả kiểm tra về sức chịu lực R (kg/cm^2) như sau:

10,0	13,0	13,7	11,5	11,0	13,5	12,2

13,0	10,0	11,0	13,5	11,5	13,0	12,2
13,5	10,0	10,0	11,5	13,0	13,7	14,0
13,0	13,7	13,0	11,5	10,0	11,0	13,0

Với độ tin cậy 95% hãy ước lượng:

a) Sức chịu lực trung bình của xi măng do nhà máy sản xuất.

b) Phương sai của sức chịu lực.

Đáp số: a) $\mu \in [11,64;12,64]$; $\sigma^2 \in [1,156;3,427]$.

Bài số 26. Để định mức thời gian gia công một chi tiết máy người ta theo dõi ngẫu nhiên quá trình gia công 25 chi tiết và thu được bảng số liệu sau:

Thời gian (phút)	Số chi tiết
15-17	1
17-19	3
19-21	4
21-23	12
23-25	3
25-27	2

Với độ tin cậy 95%, hãy ước lượng thời gian gia công trung bình một chi tiết máy. Giả thiết thời gian gia công chi tiết là biến ngẫu nhiên phân phối chuẩn.

Đáp số: $\mu \in [20,5293;21,5107]$.

Bài số 27. Kiểm tra 300 gói hàng do máy tự động đóng gói thì thấy trọng lượng trung bình là 1404g, với độ lệch tiêu chuẩn là 83,4g.

a) Hãy tính độ tin cậy với sai số cho phép là 5g.

b) Với độ tin cậy là 99 % hãy ước lượng trọng lượng trung bình của toàn bộ gói hàng do máy đó đóng gói.

Đáp số: a) 70,2%; b) $\mu \in [1391,577;1416,423]$.

Bài số 28. Với độ tin cậy 95%, hãy ước lượng lượng xăng hao phí trung bình cho một ô tô chạy từ A đến B nếu chạy thử 30 lần trên đoạn đường này người ta ghi nhận được lượng xăng hao phí như sau:

Lượng xăng hao phí (lít)	Số lần tương ứng
9,6-9,8	3
9,8-10,0	5

10,0-10,2	10
10,2-10,4	8
10,4-10,6	4

Biết rằng lượng xăng hao phí là biến ngẫu nhiên phân phối chuẩn.

Đáp số: $\mu \in [10,0454;10,2212]$.

Bài số 29. Điều tra doanh số hàng tháng của 100 hộ kinh doanh một loại hàng, ta có bảng số liệu sau:

Doanh số (triệu đồng)	Số hộ tương ứng
11,5	10
11,6	15
11,7	20
11,8	30
11,9	15
12,0	10

Hãy ước lượng doanh số trung bình hàng tháng của các hộ kinh doanh mặt hàng này bằng khoảng tin cậy đối xứng với hệ số tin cậy 95%, với giả thiết doanh số là biến ngẫu nhiên phân phối chuẩn.

Đáp số: $\mu \in [11,7268;11,7832]$.

Bài số 30. Để xác định giá trung bình đối với một loại hàng hoá trên thị trường, người ta điều tra ngẫu nhiên tại 100 cửa hàng thu được bảng số liệu sau:

Giá (ngàn đồng)	83	85	87	89	91	93	95	97	99	101
Số cửa hàng	6	7	12	15	30	10	8	6	4	2

Với độ tin cậy 95%, hãy ước lượng giá trung bình của loại hàng đó tại thời điểm đang xét.

Đáp số: $\mu \in [89,903;91,537]$.

Bài số 31. Để nghiên cứu độ ổn định của một máy gia công, người ta lấy ngẫu nhiên 25 chi tiết do máy đó gia công, đem đo và thu được các kích thước sau:

24,1	27,2	26,7	23,6	26,4
25,8	27,3	23,2	26,9	27,1
22,7	26,9	24,8	24,0	23,4
24,5	26,1	25,9	25,4	22,9

26,4	25,4	23,3	23,0	24,3

Với độ tin cậy 95% hãy ước lượng độ phân tán của kích thước các chi tiết do máy đó gia công. Biết kích thước các chi tiết là biến ngẫu nhiên phân phối chuẩn.

Đáp số: $\sigma^2 \in [1,46683;\ 4,7781]$.

Bài số 32. Lãi suất cổ phiếu của một công ty trong vòng 5 năm qua là 15%, 10%, 20%, 7%, 14%. Với độ tin cậy 90% hãy ước lượng độ phân tán của lãi suất cổ phiếu của công ty đó. Biết lãi suất cổ phiếu là biến ngẫu nhiên phân phối chuẩn.

Đáp số: $\sigma^2 \in [10,4132;\ 138,959]$.

Bài số 33. Để nghiên cứu một loại bệnh ở một địa phương, người ta đã khám 2500 người và thấy 500 người mắc bệnh.

a) Hãy ước lượng tỷ lệ mắc bệnh này ở toàn bộ địa phương đó với độ tin cậy 95%.
b) Cần khám cho bao nhiêu người để có thể nhận định về tỷ lệ mắc bệnh của toàn bộ địa phương với sai số không quá 0,01 và với độ tin cậy 99 %.

Đáp số: a) $p \in [0,18432; 0,21568]$*; b) 10651.*

Bài số 34. Một doanh nghiệp có dự định đưa một sản phẩm mới vào một thị trường có 1500000 người tiêu dùng. Nghiên cứu thị trường đối với 2500 khách hàng thấy 800 người sẵn sàng mua sản phẩm đó.

a) Với độ tin cậy 95% hãy ước lượng thị phần tiềm năng của doanh nghiệp.
b) Số lượng khách hàng tiềm năng mà doanh nghiệp hy vọng sẽ có được ở thị trường mới là bao nhiêu?

Đáp số: a) $p \in [0,301714; 0,338286]$*; b)* $[452572; 507429]$.

Bài số 35. Khảo sát ngẫu nhiên 400 sản phẩm của một nhà máy sản xuất, thấy có 92 sản phẩm đạt chất lượng loại A.

a) Với độ tin cậy 95% hãy ước lượng tỷ lệ sản phẩm loại A của nhà máy.
b) Nếu nhà máy sản xuất tổng cộng 100000 sản phẩm thì có tối đa bao nhiêu sản phẩm loại A?

Đáp số: a) $p \in [0,1888; 0,2712]$*; b)* 26460.

3.7. Tài liệu tham khảo

[1] Nguyễn Cao Văn, Trần Thái Ninh, Ngô Văn Thứ, Lý thuyết xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[2] Nguyễn Cao Văn, Trần Thái Ninh, Nguyễn Thế Hệ, Bài tập xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[3] Phạm Văn Chững, Lê Thanh Hoa, Nguyễn Đình Uông, Thống kê ứng dụng, NXB Đại học Quốc gia Thành Phố Hồ Chí Minh, 2016.

[4] Hà văn Sơn, Giáo trình Lý thuyết Thống kê, ứng dụng trong Quản trị và kinh tế.

[5] Anderson, Sweeney, and William [2010], Statistics for Business and Economics, South-Western Cengage Learning (11^{th} Edition).

[6] Michael Barrow, Statistics for Economics, Accounting and Business Studies-Prentice Hall, 2006.

[7] Newbold Paul - Statistics for Bussiness and Economics, 5th edition - Prentice Hall, 2005.

Thuật ngữ chính chương 3	
Tiếng Anh	**Tiếng Việt**
Approximately	Xấp xỉ
Biased Estimator	Ước lượng chệch
Confidence interval	Khoảng tin cậy
Central limit theorem	Định lý giới hạn trung tâm
Convenience Sampling	Lấy mẫu thuận tiện
Infinite population	Tổng thể vô hạn
Expected value	Giá trị kỳ vọng
Finite population	Tổng thể hữu hạn
Random sample	Mẫu ngẫu nhiên
Relative efficiency	Hiệu quả tương đối
Judgment sampling	Lấy mẫu phán đoán
Point Estimation	Ước lượng điểm
Population Parameter	Tham số tổng thể
Sample mean	Trung bình mẫu
Population proportion	Tỷ lệ tổng thể
Properties of point estimator	Tính chất của ước lượng điểm
Target population	Tổng thể mục tiêu
Simple random sample	Mẫu ngẫu nhiên đơn giản
Sample proportion	Tỷ lệ mẫu
Sampling distribution	Phân phối mẫu
Sample size	Cỡ mẫu
Sampling method	Phương pháp lấy mẫu
Standard deviation of the population	Độ lệch chuẩn tổng thể
Stratified random sampling	Lấy mẫu ngẫu nhiên phân tầng
Standard error	Độ lệch chuẩn
Statistic	Thống kê
Sample statistic	Thống kê mẫu
Sample variance	Phương sai mẫu
Unbiased estimator	Ước lượng không chệch

Chương 4

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

Mục tiêu chương 4

Chương này giúp sinh viên:

- Hiểu được thế nào là giả thuyết, đối thuyết và kiểm định giả thuyết thống kê.
- Các loại sai lầm thường gặp trong kiểm định giả thuyết thống kê...
- Nắm và áp dụng được một số bài toán kiểm định tham số như kiểm định trung bình, kiểm định phương sai và kiểm định tỷ lệ.
- Nắm và áp dụng được một số bài toán kiểm định phi tham số như kiểm định luật phân phối, kiểm định tính độc lập, kiểm định dấu – tổng hạng Wilconxon và kiểm định Kruskal – Wallis.

4.1. Bài toán kiểm định giả thuyết thống kê

4.1.1. Đặt vấn đề, giả thuyết, đối thuyết, kiểm định giả thuyết thống kê

Giả sử ta đi tiếp nhận một lô hàng (rất nhiều) và ta chỉ bằng lòng nhận nếu tỷ lệ hỏng $p \le 0,05$ và từ chối nếu $p > 0,05$.

Vậy ta có bài toán kiểm định

$$\begin{cases} H_0 : p \le 0,05 \\ H_1 : p > 0,05 \end{cases}$$

Mô hình tổng quát của bài toán kiểm định là : ta nêu lên hai mệnh đề trái ngược nhau, một mệnh đề được gọi là *giả thuyết* H_0 và mệnh đề ngược lại được gọi là *nghịch thuyết (đối thuyết)* H_1. Giải quyết một bài toán kiểm định là nêu lên một quy tắc hành động (chấp nhận giả thuyết H_0 hoặc bác bỏ giả thuyết H_0) bằng cách dựa vào mẫu quan sát.

Ta nói rằng : *chấp nhận giả thuyết* H_0, có nghĩa là ta tin rằng H_0 đúng; *từ chối* H_0 có nghĩa là ta tin rằng H_0 sai. Ở đây, ta không thể khẳng định H_0 đúng hay sai, ta chỉ quan sát ngẫu nhiên một số trường hợp nên không thể khẳng định chắc chắn điều gì cho cả tổng thể.

Giả sử dấu hiệu nghiên cứu trong tổng thể có thể xem như biến ngẫu nhiên X. Nếu chưa biết dạng phân phối xác suất của nó, song có cơ sở để giả thiết rằng X phân phối theo một quy luật A nào đó, người ta đưa ra giả thuyết: Biến ngẫu nhiên X tuân theo quy luật A.

Cũng có trường hợp dạng phân phối xác suất của X đã biết song tham số đặc trưng của nó lại chưa biết, nếu có cơ sở để giả thiết rằng giá trị của tham số bằng θ_0, người ta đưa ra giả thuyết: $\theta = \theta_0$.

Khi nghiên cứu hai hay nhiều biến ngẫu nhiên thuộc các tổng thể khác nhau hay thuộc cùng một tổng thể thường phải xét xem chúng độc lập hay phụ thuộc nhau, các tham số đặc trưng của chúng có bằng nhau hay không. Nếu chưa biết một cách chắc chắn song có cơ sở để nhận định về các vấn đề đó cũng có thể đưa ra các giả thuyết tương ứng.

Định nghĩa: *Giả thuyết thống kê là giả thuyết về quy luật phân phối xác suất của biến ngẫu nhiên, về các tham số đặc trưng của biến ngẫu nhiên, hoặc về tính độc lập của các biến ngẫu nhiên.*

Ví dụ 4.1. Khi nghiên cứu nhu cầu thị trường X về một loại hàng hóa nào đó, ta có thể có các giả thuyết:

H_0: X phân phối chuẩn

H_0: Nhu cầu trung bình $\mu = 50$ tấn/tháng.

H_0: Nhu cầu X và giá Y là độc lập.

Giả thuyết thống kê có thể là đúng hoặc sai nên phải kiểm định gọi là phép kiểm định giả thuyết thống kê.

Giả thuyết thống kê đưa ra được gọi là giả thuyết gốc, ký hiệu là H_0. Để kiểm định giả thuyết H_0, người ta thành lập giả thuyết mâu thuẫn với nó gọi là giả thuyết đối hay nghịch thuyết, ký hiệu là H_1. Ta có H_0 và H_1 tạo nên cặp giả thuyết thống kê.

Ví dụ 4.2. Tiếp ví dụ 4.1 ta có đối thuyết đối của từng H_0 tương ứng:

H_1: X không phân phối chuẩn.

H_1: $\mu > 50$; H_1: $\mu < 50$; H_1: $\mu \neq 50$.

H_1: X và Y phụ thuộc.

Phương pháp chung để kiểm định giả thuyết thống kê như sau: Trước hết giả sử H_0 đúng và từ đó dựa vào thông tin của mẫu rút ra từ tổng thể có thể tìm được biến cố A nào đó, sao cho xác suất xảy ra biến cố A bằng α rất bé mà có thể coi A không xảy ra trong phép thử về biến cố này. Lúc đó trên một mẫu cụ thể thực hiện một phép thử đối với biến cố A, nếu A xảy ra thì chứng tỏ H_0 sai và ta bác bỏ nó, còn nếu A không xảy ra thì ta chưa có cơ sở để bác bỏ H_0.

4.1.2. Các loại sai lầm trong kiểm định giả thuyết thống kê

Khi kiểm định một giả thuyết thống kê, có thể mắc các sai lầm thuộc hai loại sau:

4.1.2.1. Sai lầm loại I: Bác bỏ giả thuyết H_0, trong khi H_0 đúng.

Mức ý nghĩa α chính là xác suất mắc sai lầm loại I.

$$P\left(G \in W_\alpha \middle| H_0\right) = \alpha \tag{4.1}$$

Thật vậy, mặc dù H_0 đúng nhưng xác suất để $(G \in W_\alpha)$ vẫn bằng α. Nhưng khi $G \in W_\alpha$, ta lại bác bỏ H_0. Do đó xác suất mắc sai lầm loại I đúng bằng α.

Sai lầm này có thể sinh ra do kích thước mẫu quá nhỏ, do phương pháp lấy mẫu,…

4.1.2.2. Sai lầm loại II: Thừa nhận giả thuyết H_0, trong khi H_0 sai, hay giá trị quan sát G_{qs} không thuộc miền bác bỏ W_α trong khi H_1 đúng.

Gọi xác suất mắc sai lầm loại II là β:

$$P\left(G \notin W_\alpha \middle| H_1\right) = 1 - \beta \tag{4.2}$$

Trên thực tế sai lầm loại I và loại II luôn mâu thuẫn nhau, tức là với một mẫu kích thước n xác định thì không thể cùng một lúc giảm xác suất mắc hai loại sai lầm nói trên được. Khi ta giảm α đi thì đồng thời sẽ làm tăng β và ngược lại.

Để dung hòa mâu thuẫn trên, người ta thường cho trước α, và trong số các miền W_α có thể lựa chọn miền nào có β nhỏ nhất, đó là miền bác bỏ tốt nhất.

Vậy miền bác bỏ tốt nhất W_α phải thỏa mãn:

$$\begin{cases} P\left(G \in W_\alpha \middle| H_0\right) = \alpha \\ P\left(G \in W_\alpha \middle| H_1\right) = 1 - \beta \to \max \end{cases} \tag{4.3}$$

Việc chọn α tùy thuộc vào hậu quả mà sai lầm loại I và loại II mang lại.

Ví dụ 4.3. Sau khi xây dựng xong một tòa nhà thì cơ quan chức năng phát hiện 20% số sắt đã bị "rút ruột".

Gọi H_0 : Chất lượng công trình đảm bảo

H_1 : Chất lượng công trình không đảm bảo.

Vậy sai lầm loại I hay loại II nghiêm trọng hơn.

Giải

Giả sử chất lượng công trình đảm bảo nhưng ta loại bỏ H_0 nhưng đập nhà đi do đó gây tốn kém tiền của.

Giả sử chất lượng công trình không đảm bảo nhưng ta vẫn thừa nhận H_0 loại bỏ H_1 nhưng vẫn đưa vào sử dụng dẫn tới nhà sập. Do đó vừa tốn kém tiền của vừa nguy hiểm đến tính mạng.

Vậy sai lầm loại II nghiêm trọng hơn suy ra chọn α lớn để β nhỏ.

4.1.3. Giải quyết vấn đề

Gần giống như trong lý thuyết về ước lượng khoảng, để giải quyết bài toán kiểm định, ta quan sát mẫu $X_1, X_2, ..., X_n$ và đưa ra giả thuyết H_0. Từ mẫu quan sát, ta chọn một thống kê $Q = f(X_1, X_2, ..., X_n, \theta)$ sao cho nếu H_0 đúng thì phân phối xác suất của Q hoàn toàn xác định. Ta còn nói thống kê Q là *tiêu chuẩn kiểm định giả thuyết* H_0.

Bấy giờ, với mức sai lầm α cho trước, ta tìm được khoảng tin cậy $[a, b]$ của Q với độ tin cậy $\gamma = 1 - \alpha$ và khi đó,

Nếu $Q \in [a, b]$: ta chấp nhận giả thuyết H_0, và

Nếu $Q \notin [a, b]$: ta bác bỏ giả thuyết H_0.

Trong ứng dụng, nếu hàm mật độ của Q có đồ thị đối xứng qua trục Oy, chẳng hạn như trong phân phối Gauss, $N(0,1)$, và phân phối Student, $St(n)$, thì ta chọn khoảng tin cậy đối xứng $[-C, C]$ với

$$P(Q \le -C) = P(Q \ge C) = \frac{\alpha}{2}$$

và do đó, ta có

Nếu $|Q| \le C$: Chấp nhận giả thuyết H_0, và

Nếu $|Q| > C$: Bác bỏ giả thuyết H_0.

Nếu hàm mật độ của Q không đối xứng, chẳng hạn như trong phân phối $\chi^2(n)$, và phân phối Fisher, $F(n, m)$, thì thay vì chọn khoảng tin cậy $[a, b]$ sao cho

$$P(a \le -C) = P(b \ge C) = \frac{\alpha}{2},$$

ta quy ước khoảng tin cậy trong phép kiểm định là $[0, C]$ với

$$P(Q \ge C) = \alpha.$$

Nếu $Q \le C$: Chấp nhận giả thuyết H_0, và

Nếu $Q > C$: Bác bỏ giả thuyết H_0.

Trong phần sau, dựa trên các bảng số liệu, ta lần lượt khảo sát các phép kiểm định :

- So sánh các bảng số liệu, mà người ta còn gọi là các phép kiểm định phi tham số.
- So sánh tham số đặc trưng của các bảng số liệu, mà người ta còn gọi là các phép kiểm định tham số.

4.2. Kiểm định giả thuyết về giá trị trung bình

Quan sát mẫu $X_1, X_2, ..., X_n$ độc lập và có cùng phân phối chuẩn $N(\mu; \sigma^2)$.

Trong đó tham số μ là chưa biết, song có cơ sở cho rằng giá trị của nó bằng μ_0, người ta đưa ra giả thuyết $H_0: \mu = \mu_0$. Để kiểm định giả thuyết trên từ tổng thể lập mẫu kích thước n: $(X_1, X_2, ..., X_n)$. (Nếu X không có phân phối chuẩn thì yêu cầu kích thước mẫu $n > 30$). Ta xét hai trường hợp sau:

4.2.1. Kiểm định giả thuyết về giá trị trung bình, nếu biết phương sai tổng thể σ_0^2

Chọn tiêu chuẩn kiểm định:

$$Z = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma_0} \tag{4.4}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma_0} \sim N(0;1)$$

Nếu cho trước mức ý nghĩa α thì tùy thuộc vào dạng của giả thuyết đối H_1, miền bác bỏ giả thuyết tốt nhất được xây dựng theo các trường hợp sau:

- Cặp giả thuyết thống kê : $\begin{cases} H_0: \ \mu = \mu_0 \\ H_1: \ \mu > \mu_o \end{cases}$

Với mức ý nghĩa α cho trước tìm được giá trị tới hạn chuẩn z_α sao cho

$$P(Z \in W_\alpha | H_0) = P(Z > z_\alpha) = \alpha$$

ta thu được miền bác bỏ bên phải:

$$W_\alpha = \left\{ Z = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma_0}; \ Z > z_\alpha \right\} \tag{4.5}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0: \ \mu = \mu_0 \\ H_1: \ \mu < \mu_o \end{cases}$

Với mức ý nghĩa α cho trước tìm được giá trị tới hạn chuẩn $z_{1-\alpha}$ sao cho

$$P\left(Z \in W_\alpha \middle| H_0\right) = P\left(Z < z_{1-\alpha}\right) = P\left(Z < -z_\alpha\right) = \alpha$$

ta thu được miền bác bỏ bên trái:

$$W_\alpha = \left\{ Z = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0}; \ Z < \ -z_\alpha \right\} \tag{4.6}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0: \ \mu = \mu_0 \\ H_1: \ \mu \neq \mu_o \end{cases}$

Với mức ý nghĩa α cho trước tìm được hai giá trị tới hạn chuẩn $z_{1-\alpha/2}$ và $z_{\alpha/2}$ sao cho

$$\begin{aligned} P\left(Z \in W_\alpha \middle| H_0\right) &= P\left(Z < z_{1-\alpha}\right) + P\left(Z > z_\alpha\right) \\ &= P\left(Z < -z_\alpha\right) + P\left(Z > z_\alpha\right) \\ &= P\left(\left|Z\right| > z_\alpha\right) = \alpha \end{aligned}$$

ta thu được miền bác bỏ hai phía:

$$W_\alpha = \left\{ Z = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0}; \ \left|Z\right| \ > \ z_{\alpha/2} \right\} \tag{4.7}$$

Từ một mẫu cụ thể $\left(X_1, X_2, ..., X_n\right)$ và tính giá trị quan sát của tiêu chuẩn kiểm định:

$Z_{qs} = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0}$ và so sánh với W_α để kết luận:

- Nếu $Z_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $Z_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Ví dụ 4.4. Trọng lượng mỗi gói sản phẩm do một nhà máy sản xuất là biến ngẫu nhiên phân phối chuẩn với độ lệch chuẩn 36g và trọng lượng trung bình 453g. Kiểm tra ngẫu nhiên 81 gói sản phẩm đó thấy trọng lượng trung bình là 448g. Với mức ý nghĩa 5% có thể kết luận các sản phẩm đóng gói có bị thiếu hay không.

Giải

Gọi X là trọng lượng gói sản phẩm $X \sim N\left(\mu, 36^2\right)$

Đây là bài toán kiểm định giả thuyết về tham số μ của biến ngẫu nhiên phân phối chuẩn khi đã biết phương sai tổng thể.

Ta có $n = 81, \overline{X} = 448, \mu_0 = 453, \sigma_0 = 36$

Cặp giả thuyết thống kê: $\begin{cases} H_0:\ \mu = 453 \\ H_1:\ \mu < 453 \end{cases}$

Nếu H_0 đúng, ta có thống kê

$$Z = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0} \sim N(0,1)$$

Thay $n = 81, \overline{X} = 448, \mu_0 = 453, \sigma_0 = 36$, ta có

$$Z_{qs} = \frac{(448 - 453)\sqrt{81}}{36} = -1,25$$

Với mức ý nghĩa $\alpha = 5\% = 0,05.$ Ta có $\phi_0(C) = \frac{1 - 2\alpha}{2} = 0,45 \rightarrow C = 1,64$

Miền bác bỏ: $W_\alpha = (-\infty; -1,64).$

Do $Z_{qs} \notin W_\alpha$ nên chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, các sản phẩm đóng gói bị thiếu.

4.2.2. Kiểm định giả thuyết về giá trị trung bình, nếu chưa biết phương sai tổng thể

Chọn tiêu chuẩn kiểm định:

$$T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X} \tag{4.8}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X} \sim St(n-1)$$

Nếu cho trước mức ý nghĩa α thì tùy thuộc vào dạng của giả thuyết đối H_1, miền bác bỏ giả thuyết tốt nhất được xây dựng theo các trường hợp sau:

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu > \mu_o \end{cases}$

Với mức ý nghĩa α cho trước tìm được giá trị tới hạn $t_\alpha(n-1)$ sao cho

$$P\left(T \in W_\alpha \middle| H_0\right) = P\left(T > t_\alpha(n-1)\right) = \alpha$$

ta thu được miền bác bỏ bên phải:

$$W_\alpha = \left\{ T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X};\ T > t_\alpha\left(n-1\right) \right\} \tag{4.9}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu < \mu_o \end{cases}$

Với mức ý nghĩa α cho trước tìm được giá trị tới hạn $t_{1-\alpha}\left(n-1\right)$ sao cho

$$P\left(T \in W_\alpha \middle| H_0\right) = P\left(T < t_{1-\alpha}\left(n-1\right)\right) = P\left(T < -t_\alpha\left(n-1\right)\right) = \alpha$$

ta thu được miền bác bỏ bên trái:

$$W_\alpha = \left\{ T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X};\ T < -t_\alpha\left(n-1\right) \right\} \tag{4.10}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu \neq \mu_o \end{cases}$

Với mức ý nghĩa α cho trước tìm được hai giá trị tới hạn $t_{1-\frac{\alpha}{2}}\left(n-1\right)$ và $t_{\frac{\alpha}{2}}\left(n-1\right)$ sao cho

$$\begin{aligned} P\left(T \in W_\alpha \middle| H_0\right) &= P\left(T < t_{1-\frac{\alpha}{2}}\left(n-1\right)\right) + P\left(T > t_{\frac{\alpha}{2}}\left(n-1\right)\right) \\ &= P\left(T < -t_{\frac{\alpha}{2}}\left(n-1\right)\right) + P\left(T > t_{\frac{\alpha}{2}}\left(n-1\right)\right) \\ &= P\left(\left|T\right| > t_{\frac{\alpha}{2}}\left(n-1\right)\right) = \alpha \end{aligned}$$

ta thu được miền bác bỏ hai phía:

$$W_\alpha = \left\{ T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X};\ \left|T\right| > t_{\frac{\alpha}{2}}(n-1) \right\} \tag{4.11}$$

Từ một mẫu cụ thể $\left(X_1, X_2, ..., X_n\right)$ và tính giá trị quan sát của tiêu chuẩn kiểm định:

$T_{qs} = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X}$ và so sánh với W_α để kết luận:

- Nếu $T_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $T_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Ví dụ 4.5. Thu hoạch thử 41 thửa ruộng trồng lúa, tính được năng suất trung bình 39,5 tạ/ha và độ lệch chuẩn mẫu hiệu chỉnh 1,2 tạ/ha. Trước đây, giống lúa này cho năng suất 39 tạ/ha. Với mức ý nghĩa 5%, có thể cho rằng năng suất lúa đã tăng lên hay không? Biết rằng năng suất lúa là biến ngẫu nhiên tuân theo quy luật chuẩn.

Giải

Gọi X là năng suất lúa, $X \sim N\left(\mu, \sigma^2\right)$.

Ta có : $n = 41; \overline{X} = 39,5; S_X = 1,2$

Đây là bài toán kiểm định giả thuyết về tham số μ của biến ngẫu nhiên phân phối chuẩn khi chưa biết phương sai tổng thể.

Xét giả thuyết H_0: "Năng suất lúa trung bình không tăng", nghĩa là ta có bài toán kiểm định

$$\begin{cases} H_0: \ \mu = 39 \\ H_1: \ \mu > 39 \end{cases}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X} \sim St\left(n-1\right)$$

Thay $n = 41; \overline{X} = 39,5; S_X = 1,2,$ ta có

$$T_{qs} = \frac{\left(39,5 - 39\right)\sqrt{41}}{1,2} = 2,667$$

Với mức ý nghĩa $\alpha = 5\% = 0,05.$ Ta có $t_\alpha\left(n-1\right) = t_{0,05}\left(40\right) = 1,645.$

Miền bác bỏ: $W_\alpha = \left(1,645; +\infty\right)$

Do $T_{qs} \in W_\alpha$ nên bác bỏ H_0. Vậy với mức ý nghĩa 5%, năng suất lúa trung bình đã tăng lên.

Ví dụ 4.6. Một máy đóng gói các sản phẩm có khối lượng 1kg. Nghi ngờ máy hoạt động không bình thường, người ta chọn ra một mẫu ngẫu nhiên gồm 100 sản phẩm thì thấy như sau :

Khối lượng	0,95	0,97	0,99	1,01	1,03	1,05
Số gói	9	31	40	15	3	2

Với mức ý nghĩa 5%, hãy kết luận về nghi ngờ trên.

Giải

Gọi X (kg) là khối lượng một gói sản phẩm

Từ số liệu của mẫu, ta có trung bình mẫu: $\overline{X} = 0,9856$, độ lệch chuẩn mẫu có hiệu chỉnh: $S_X = 0,021$, cỡ mẫu: $n = 100$.

Xét giả thuyết H_0: "máy hoạt động bình thường", nghĩa là ta có bài toán kiểm định

$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$T = \frac{(\overline{X} - \mu)\sqrt{n}}{S_X} \sim St(n-1)$$

Với mức ý nghĩa 5%, ta có $t_{\frac{\alpha}{2}}(n-1) = t_{0,025}(99) = 1,96$.

Miền bác bỏ : $W_\alpha = (-\infty; -1,96) \cup (1,96; +\infty)$.

Với số liệu trên, ta được

$$T_{qs} = \frac{(0,9856-1)\sqrt{100}}{0,021} = -6,86$$

Do $T_{qs} \in W_\alpha$ nên ta bác bỏ H_0. Vậy với mức ý nghĩa 5%, máy hoạt động không bình thường.

Ví dụ 4.7. Quan sát số hoa hồng bán ra trong một ngày của một cửa hàng bán hoa sau một thời gian, người ta ghi được số liệu sau :

Số hoa hồng (đoá)	12	13	15	16	17	18	19
Số ngày	3	2	7	7	3	2	1

Sau khi tính toán, ông chủ cửa hàng nói rằng nếu trung bình một ngày không bán được 15 đoá hoa thì chẳng thà đóng cửa còn hơn. Dựa vào số liệu trên, anh (chị) hãy kết luận giúp ông chủ cửa hàng xem có nên tiếp tục bán hay không ở mức ý nghĩa 5%.

Giải

Gọi X (đóa) là số hoa hồng bán ra trong một ngày

Ta có $\mu_0 = 15; \overline{X} = 15,4; S_X = 1,871; n = 25$.

Xét giả thiết H_0 : "nên bán tiếp", ta có bài toán kiểm định

$$\begin{cases} H_0 : \mu = \mu_0 = 15 \\ H_1 : \mu < \mu_0 = 15 \end{cases}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$T = \frac{(\overline{X} - \mu)\sqrt{n}}{S_X} \sim St(n-1).$$

Với mức ý nghĩa $\alpha = 5\% = 0,05.$ Ta có $t_\alpha(n-1) = t_{0,05}(24) = 1,711$

Miền bác bỏ : $W_\alpha = (-\infty; -1,711)$.

Thế có $\mu_0 = 15; \overline{X} = 15,4; S_X = 1,871; n = 25$ vào, ta có

$$T_{qs} = \frac{(15,4-15)\sqrt{5}}{1,871} = 1,07$$

Do $T_{qs} \notin W_\alpha$ nên chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, ông chủ nên tiếp tục bán.

4.3. Kiểm định giả thuyết về tỷ lệ

Giả sử trong tổng thể biến ngẫu nhiên $X \sim B(1,p)$, với tham số p. Nếu chưa biết p, song có thể cho rằng giá trị của nó bằng p_0 thì đưa ra giả thuyết thống kê.

$$H_0 : p = p_0$$

Để kiểm định giả thuyết trên, từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$(X_1, X_2, ..., X_n)$$

Chọn tiêu chuẩn kiểm định:

$$Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \tag{4.12}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê: $Z \sim N(0,1)$. Do đó với mức ý nghĩa α cho trước, các miền bác bỏ W_α được xác định như sau:

- Cặp giả thuyết thống kê : $\begin{cases} H_0: p = p_0 \\ H_1: p > p_0 \end{cases}$

Miền bác bỏ bên phải:

$$W_\alpha = \left\{ Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}};\ Z > z_\alpha \right\} \tag{4.13}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0: p = p_0 \\ H_1: p < p_0 \end{cases}$

Miền bác bỏ bên trái:

$$W_\alpha = \left\{ Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}};\ Z < -z_\alpha \right\} \tag{4.14}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ p = p_0 \\ H_1:\ p \neq p_0 \end{cases}$

Miền bác bỏ hai phía:

$$W_\alpha = \left\{ Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}};\ |Z| > z_{\alpha/2} \right\} \tag{4.15}$$

Từ một mẫu cụ thể và tính giá trị quan sát: $Z_{qs} = \dfrac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}$ và so sánh với W_α để kết luận.

- Nếu $Z_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $Z_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Ví dụ 4.8. Thống kê 10000 trẻ sơ sinh ở một địa phương, người ta thấy 5080 bé trai. Hỏi tỷ lệ sinh con trai có thực sự cao hơn tỷ lệ sinh con gái không? Cho kết luận với mức ý nghĩa 0,01.

Giải

Gọi X là số con trai, $X \sim B(1;p)$

Ta có: $n = 10000;\ k = 5080 \rightarrow f = \dfrac{5080}{10000} = 0,508$

Cặp giả thuyết thống kê: $\begin{cases} H_0: p = p_0 = 0,5 \\ H_1: p > p_0 = 0,5 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0,1)$$

Với mức ý nghĩa $\alpha = 1\% = 0,01$. Ta có $\phi_0(C) = \dfrac{1-2\alpha}{2} = 0,49 \rightarrow C = 2,33$

Miền bác bỏ: $W_\alpha = (2,33; +\infty)$.

Thay Ta có: $n = 10000;\ f = 0,508;\ p_0 = 0,5$. Ta có

$$Z_{qs} = \frac{(0,508 - 0,5)\sqrt{10000}}{\sqrt{0,5.0,5}} \approx 1,6$$

Do $Z_{qs} \notin W_\alpha$ nên chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 1%, tỷ lệ sinh con trai thực sự cao hơn tỷ lệ sinh con gái.

Ví dụ 4.9. Trong một vùng dân cư có 18 bé trai và 28 bé gái mắc bệnh B. Hỏi rằng tỷ lệ nhiễm bệnh của bé trai và bé gái có như nhau không ? (kết luận với ý nghĩa 5% và giả sử rằng số lượng bé trai và bé gái trong vùng tương đương nhau, và rất nhiều).

Giải

Ta có: $n = 46;\ k = 18 \to f = \dfrac{18}{46} = 0,391$

Xét giả thuyết H_0: "tỷ lệ mắc bệnh B của bé trai và bé gái là như nhau", nghĩa là ta có bài toán kiểm định

$$\begin{cases} H_0 : p = p_0 = 0,5 \\ H_1 : p \neq p_0 = 0,5 \end{cases}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0,1).$$

Thay $n = 46;\ f = 0,391;\ p_0 = 0,5$, ta có :

$$Z_{qs} = \frac{(0,391 - 0,5)\sqrt{46}}{\sqrt{0,5 \cdot 0,5}} = -1,48.$$

Với mức ý nghĩa $\alpha = 5\% = 0,05.$ Ta có $\phi_0(C) = \dfrac{1-\alpha}{2} = 0,475 \to C = 1,96$

Miền bác bỏ: $W_\alpha = (-\infty; -1,96) \cup (1,96; +\infty)$.

Do $Z_{qs} \notin W_\alpha$ nên chưa đủ cơ sở bác bỏ H_0, vậy với mức ý nghĩa 5%, tỷ lệ mắc bệnh B của bé trai và bé gái là như nhau.

4.4. Kiểm định giả thuyết về phương sai

Giả sử trong tổng thể, biến ngẫu nhiên gốc X phân phối $N(\mu, \sigma^2)$ với σ^2 chưa biết song có cơ sở để giả thiết rằng giá trị của nó bằng σ_0^2. Người ta đưa ra giả thuyết:

$$H_0: \sigma^2 = \sigma_0^2$$

Để kiểm định giả thuyết trên, từ tổng thể lập mẫu ngẫu nhiên kích thước n:

$$(X_1, X_2, ..., X_n)$$

và chọn tiêu chuẩn kiểm định.

$$Q = \frac{(n-1)S_X^2}{\sigma_0^2} \qquad (4.16)$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Q = \frac{(n-1)S_X^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Do đó với mức ý nghĩa α cho trước, tùy thuộc vào giả thuyết đối H_1, miền bác bỏ W_α được xây dựng như sau:

- Cặp giả thuyết thống kê : $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases}$

Miền bác bỏ bên phải:

$$W_\alpha = \left\{ Q = \frac{(n-1)S_X^2}{\sigma_0^2};\ \chi^2 > \chi_\alpha^2(n-1) \right\} \qquad (4.17)$$

- Cặp giả thuyết: $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$

Miền bác bỏ bên trái:

$$W_\alpha = \left\{ Q = \frac{(n-1)S_X^2}{\sigma_0^2};\ \chi^2 < \chi_{1-\alpha}^2(n-1) \right\} \qquad (4.18)$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$

Miền bác bỏ hai phía:

$$W_\alpha = \left\{ Q = \frac{(n-1)S_X^2}{\sigma_0^2};\ \chi^2 < \chi_{1-\alpha/2}^2(n-1) \text{ hoac } \chi^2 > \chi_{\alpha/2}^2(n-1) \right\} \qquad (4.19)$$

Từ một mẫu cụ thể $(X_1, X_2, ..., X_n)$ và tính giá trị quan sát của tiêu chuẩn kiểm định: $Q_{qs} = \frac{(n-1)S_X^2}{\sigma_0^2}$ và so sánh với W_α để kết luận:

- Nếu $Q_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $Q_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Ví dụ 4.10. Để kiểm tra độ chính xác của một máy người ta đo ngẫu nhiên kích thước của 15 chi tiết do máy đó sản xuất và tính được $S_X^2 = 14,6$. Với mức ý nghĩa 1% hãy kết luận máy đó có hoạt động bình thường không, biết rằng kích thước chi tiết là biến ngẫu nhiên phân phối chuẩn có phương sai theo thiết kế là $\sigma^2 = 12$.

Giải

Gọi X là kích thước chi tiết do máy đó sản xuất, $X \sim N\left(\mu, \sigma^2\right)$.

Ta có : $n = 15,\ S_X^2 = 14,6;\ \sigma_0^2 = 12$

Đây là bài toán kiểm định giả thuyết thống kê về tham số σ^2 của biến ngẫu nhiên phân phối chuẩn.

Cặp giả thuyết thống kê: $\begin{cases} H_0: \ \sigma^2 = 12 \\ H_1: \ \sigma^2 > 12 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Q = \frac{(n-1)S_X^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Thay $n = 15,\ S_X^2 = 14,6;\ \sigma_0^2 = 12.$ Ta có

$$Q_{qs} = \frac{14.14,6}{12} = 17,033$$

Với mức ý nghĩa 1%, ta có $\chi_\alpha^2(n-1) = \chi_{0,01}^2(14) = 29,141$

Miền bác bỏ: $W_\alpha = (29,141; +\infty)$.

Do $Q_{qs} \notin W_\alpha$ nên chưa có cơ sở để bác bỏ H_0, có thể nói máy móc vẫn làm việc bình thường.

4.5. Bài toán so sánh

4.5.1. So sánh hai trung bình μ_X và μ_Y của hai tổng thể

Để so sánh trung bình của hai tổng thể thỏa phân phối chuẩn và dựa vào hai mẫu quan sát độc lập lấy từ hai tổng thể này,

$$X_1, X_2, ..., X_n;\ X_i \sim N(\mu_X, \sigma_X^2);\ Y_1, Y_2, ..., Y_m;\ Y_j \sim N(\mu_Y, \sigma_Y^2).$$

Để so sánh hai trung bình μ_X và μ_Y

Trong đó μ_X và μ_Y chưa biết song có cơ sở để giả thiết rằng giá trị của chúng bằng nhau, người ta đưa ra giả thuyết $H_0 : \mu_X = \mu_Y$. Từ tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước n và m. (Nếu X và Y không có phân phối chuẩn thì yêu cầu hai kích thước mẫu n và m lớn hơn 30).

4.5.1.1. So sánh hai trung bình μ_X và μ_Y của hai tổng thể nếu đã biết $\sigma_X^2;\ \sigma_Y^2$

Chọn tiêu chuẩn kiểm định:

$$Z = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}} \tag{4.20}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}} \sim N(0;1)$$

Các miền bác bỏ mức α có dạng:

- Cặp giả thuyết: $\begin{cases} H_0:\ \mu_X = \mu_Y \\ H_1:\ \mu_X > \mu_Y \end{cases}$

Miền bác bỏ bên phải:

$$W_\alpha = \left\{ Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}};\ Z > z_\alpha \right\} \tag{4.21}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu_X = \mu_Y \\ H_1:\ \mu_X < \mu_Y \end{cases}$

Miền bác bỏ bên trái:

$$W_\alpha = \left\{ Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}};\ Z < -z_\alpha \right\} \tag{4.22}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu_X = \mu_Y \\ H_1:\ \mu_X \neq \mu_Y \end{cases}$

Miền bác bỏ hai phía:

$$W_\alpha = \left\{ Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}};\ |Z| > z_{\alpha/2} \right\} \tag{4.23}$$

Từ một mẫu cụ thể và tính giá trị quan sát: $Z_{qs} = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}}$ và so sánh với W_α để kết luận.

- Nếu $Z_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $Z_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

4.5.1.2. So sánh hai trung bình μ_X và μ_Y của hai tổng thể nếu chưa biết $\sigma_X^2;\ \sigma_Y^2$

Chọn tiêu chuẩn kiểm định:

$$Z = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}} \tag{4.24}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}} \sim N\left(0;1\right) \quad \text{với } n, m > 30$$

Các miền bác bỏ mức α có dạng:

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu_X = \mu_Y \\ H_1:\ \mu_X > \mu_Y \end{cases}$

Miền bác bỏ bên phải:

$$W_\alpha = \left\{ Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}};\ Z > z_\alpha \right\} \tag{4.25}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu_X = \mu_Y \\ H_1:\ \mu_X < \mu_Y \end{cases}$

Miền bác bỏ bên trái:

$$W_\alpha = \left\{ Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}};\ Z < -z_\alpha \right\} \tag{4.26}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu_X = \mu_Y \\ H_1:\ \mu_X \neq \mu_Y \end{cases}$

Miền bác bỏ hai phía:

$$W_\alpha = \left\{ Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}};\ |Z| > z_{\alpha/2} \right\} \tag{4.27}$$

Từ một mẫu cụ thể và tính giá trị quan sát: $Z_{qs} = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$ và so sánh với W_α để kết luận.

- Nếu $Z_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $Z_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Ví dụ 4.11. Kết quả điểm thi môn xác suất thống kê của hai lớp A và B như sau:

Lớp A: $n = 64;\ \overline{X} = 73,2;\ S_X^2 = 118,81$

Lớp B: $m = 68;\ \overline{Y} = 76,6;\ S_Y^2 = 125,44\, n_2$

Với mức ý nghĩa 5%, có thể cho rằng kết quả điểm thi trung bình của lớp B cao hơn lớp A được không, biết rằng kết quả điểm thi là biến ngẫu nhiên phân phối chuẩn.

Giải

Gọi X và Y lần lượt là kết quả thi của hai lớp A, B.

$$X \sim N\left(\mu_X, \sigma_X^2\right); Y \sim N\left(\mu_Y, \sigma_Y^2\right).$$

Cặp giả thuyết thống kê :

$$\begin{cases} H_0:\ \mu_X = \mu_Y \\ H_1:\ \mu_X < \mu_Y \end{cases}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim N(0;1)$$

Thay $n = 64;\ \overline{X} = 73,2;\ S_X^2 = 118,81$ và $m = 68;\ \overline{Y} = 76,6;\ S_Y^2 = 125,44$, ta có

$$Z_{qs} = \frac{73,2 - 76,6}{\sqrt{\frac{118,81}{64} + \frac{125,44}{68}}} \approx -1,7673$$

Với mức ý nghĩa $\alpha = 5\% = 0,05,$ ta có $\phi_0(C) = \dfrac{1-2\alpha}{2} = 0,45 \to C = 1,64$

Miền bác bỏ: $W_\alpha = (-\infty; -1,64).$

Do $Z_{qs} \in W_\alpha$ nên bác bỏ H_0, vậy với mức ý nghĩa 5%, kết quả điểm thi trung bình của lớp B cao hơn lớp A.

Ví dụ 4.12. Người ta cân trẻ sơ sinh ở hai khu vực thành thị và nông thôn, thu được kết quả như sau:

Khu vực	Số trẻ được cân	Trọng lượng trung bình	Phương sai
Nông thôn	2500	3,0	200
Thành thị	500	3,1	5

Với mức ý nghĩa 1%, có thể coi trọng lượng trung bình của trẻ sơ sinh ở hai khu vực bằng nhau được hay không?

Giải

Gọi X và Y lần lượt là trọng lượng của trẻ sơ sinh nông thôn và thành thị.

$$X \sim N(\mu_X, \sigma_X^2); Y \sim N(\mu_Y, \sigma_Y^2).$$

Ta có : $n = 2500; \bar{X} = 3; S_X^2 = 200$ và $m = 500; \bar{Y} = 3,1; S_Y^2 = 5$

Cặp giả thuyết thống kê : $\begin{cases} H_0: \mu_X = \mu_Y \\ H_1: \mu_X \neq \mu_Y \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}}} \sim N(0;1)$$

Thay $n = 2500; \bar{X} = 3; S_X^2 = 200$ và $m = 500; \bar{Y} = 3,1; S_Y^2 = 5$, ta có

$$Z_{qs} = \frac{3,0 - 3,1}{\sqrt{\dfrac{200}{2500} + \dfrac{5}{500}}} = -0,33$$

Với mức ý nghĩa $\alpha = 1\% = 0,01,$ ta có $\phi_0(C) = \dfrac{1-\alpha}{2} = 0,495 \to C = 2,58$

Miền bác bỏ: $W_\alpha = (-\infty; -2,58) \cup (2,58; +\infty).$

Do $Z_{qs} \notin W_\alpha$ nên chưa đủ cơ sở bác bỏ H_0, vậy với mức ý nghĩa 1%, trọng lượng của trẻ sơ sinh nông thôn và thành thị là như nhau.

4.5.2. So sánh hai tỷ lệ p_X và p_Y của hai tổng thể

Để so sánh tỷ lệ của hai tổng thể, ta cũng dựa vào các tỷ lệ lấy ra từ hai mẫu quan sát độc lập từ hai tổng thể này,

$$X_1, X_2, ..., X_n; \quad Y_1, Y_2, ..., Y_m$$

trong đó X_i, Y_j chỉ lấy các giá trị là 0 hay 1. Khi đó, $f_X = \frac{1}{n}\sum_{i=1}^{n} X_i$ là tỷ lệ (tần suất) của mẫu X và $f_Y = \frac{1}{n}\sum_{j=1}^{n} Y_j$ là tỷ lệ (tần suất) của mẫu Y.

Để so sánh hai tỷ lệ p_X và p_Y. Nếu tỷ lệ p_X và p_Y chưa biết song có cơ sở cho rằng giá trị của chúng bằng nhau, ta đưa ra giả thuyết $H_0 : p_X = p_Y$. Từ tổng thể rút ra hai mẫu kích thước n, m và chọn tiêu chuẩn kiểm định:

$$Z = \frac{(f_X - f_Y) - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_Y)}{n} + \frac{p_Y(1-p_Y)}{m}}} \tag{4.28}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$Z = \frac{f_X - f_Y}{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0;1) \quad \text{với } n, m > 30 \quad (p_X = p_Y = p)$$

p chưa biết nên được thay bằng ước lượng của nó là

$$f = \frac{nf_X + mf_Y}{n + m}$$

Các miền bác bỏ mức α được xác định như sau:

- Cặp giả thuyết thống kê : $\begin{cases} H_0: \ p_X = p_Y \\ H_1: \ p_X > p_Y \end{cases}$

$$\text{Miền bác bỏ: } W_\alpha = \left\{ Z = \frac{f_X - f_Y}{\sqrt{f(1-f)\left(\frac{1}{n} + \frac{1}{m}\right)}}; \quad Z > z_\alpha \right\} \tag{4.29}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0: \ p_X = p_Y \\ H_1: \ p_X < p_Y \end{cases}$

Miền bác bỏ: $W_\alpha = \left\{ Z = \dfrac{f_X - f_Y}{\sqrt{f(1-f)\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}};\ Z < -z_\alpha \right\}$ (4.30)

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ p_X = p_Y \\ H_1:\ p_X \neq p_Y \end{cases}$

Miền bác bỏ: $W_\alpha = \left\{ Z = \dfrac{f_X - f_Y}{\sqrt{f(1-f)\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}};\ |Z| > z_{\alpha/2} \right\}$ (4.31)

Từ một mẫu cụ thể và ta tính giá trị quan sát: $Z_{qs} = \dfrac{f_X - f_Y}{\sqrt{f(1-f)\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}}$ và so sánh với W_α để kết luận.

- Nếu $Z_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $Z_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Ví dụ 4.13. Bệnh A có thể được chữa khỏi bằng 2 loại thuốc H và K. Người ta dùng thử thuốc H cho 250 bệnh nhân thấy 210 người khỏi bệnh và dùng thử thuốc K cho 200 bệnh nhân thấy 170 người khỏi bệnh. Có thể cho rằng hiệu quả chữa bệnh của thuốc K là cao hơn thuốc H không với mức ý nghĩa 0,05.

Giải

Gọi p_X, p_Y lần lượt là tỷ lệ bệnh nhân khỏi bệnh khi dùng thuốc H và K

Ta có : $n = 250;\ f_1 = \dfrac{210}{250} = 0{,}84$; $m = 200;\ f_Y = \dfrac{170}{200} = 0{,}85$

Tỷ lệ chung: $f = \dfrac{nf_X + mf_Y}{n + m} = \dfrac{210+170}{250+200} = \dfrac{38}{45}$

Cặp giả thuyết thống kê: $\begin{cases} H_0:\ p_X = p_Y \\ H_1:\ p_X < p_Y \end{cases}$

Nếu giả thuyết H_0 đúng thì

$$Z = \frac{f_X - f_Y}{\sqrt{p(1-p)\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}} \sim N(0;1)$$

Thay $n = 250;\ f_1 = 0{,}84;\ m = 200;\ f_Y = 0{,}85;\ f = \frac{38}{45}$. Ta có

$$Z_{qs} = \frac{0{,}84 - 0{,}85}{\sqrt{\frac{38}{45}\left(1 - \frac{38}{45}\right)\left(\frac{1}{250} + \frac{1}{200}\right)}} = -\ 0{,}29$$

Với mức ý nghĩa $\alpha = 5\% = 0{,}05,$ ta có $\phi_0(C) = \frac{1-2\alpha}{2} = 0{,}45 \to C = 1{,}64$

Miền bác bỏ: $W_\alpha = (-\infty; -1{,}64)$.

Do $Z_{qs} \notin W_\alpha$ nên chưa có cơ sở để bác bỏ H_0, không thể cho rằng hiệu quả chữa khỏi bệnh của thuốc K là cao hơn thuốc H.

4.5.3. So sánh hai phương sai σ_X^2 và σ_Y^2 của hai tổng thể

Để so sánh phương sai của hai tổng thể thỏa phân phối chuẩn, ta dựa vào hai mẫu quan sát độc lập lấy từ hai tổng thể này,

$$X_1, X_2, ..., X_n;\ X_i \sim N\left(\mu_X, \sigma_X^2\right);\ Y_1, Y_2, ..., Y_m;\ Y_i \sim N\left(\mu_Y, \sigma_Y^2\right).$$

Nếu chưa biết σ_X^2 và σ_Y^2 nhưng có cơ sở để cho rằng giá trị của chúng bằng nhau thì ta đưa ra giả thuyết $H_0: \sigma_1^2 = \sigma_Y^2$.

Để kiểm định giả thuyết trên, từ hai tổng thể rút ra hai mẫu ngẫu nhiên độc lập kích thước n, m.

Chọn tiêu chuẩn kiểm định:

$$F = \frac{S_X^2 . \sigma_Y^2}{S_Y^2 . \sigma_X^2} \tag{4.32}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$F = \frac{S_X^2}{S_Y^2} \sim F(n-1; m-1)$$

Các miền bác bỏ mức α được xác định như sau:

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \sigma_X^2 = \sigma_Y^2 \\ H_1:\ \sigma_X^2 > \sigma_Y^2 \end{cases}$

Miền bác bỏ: $$W_\alpha = \left\{F = \frac{S_X^2}{S_Y^2};\ F > f_\alpha(n-1; m-1)\right\} \tag{4.33}$$

- Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \sigma_X^2 = \sigma_Y^2 \\ H_1:\ \sigma_X^2 \neq \sigma_Y^2 \end{cases}$

Miền bác bỏ:

$$W_\alpha = \left\{ F = \frac{S_X^2}{S_Y^2}; F < f_{1-\alpha/2}(n-1;m-)1;\ F > f_{\alpha/2}(n-1;m-1) \right\} \quad (4.34)$$

Từ một mẫu cụ thể và tính giá trị quan sát: $F_{qs} = \frac{S_X^2}{S_Y^2}$ và so sánh với W_α để kết luận.

- Nếu $F_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $F_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

Ví dụ 4.14. Một giống lúa được gieo cấy ở hai vùng A và B. Khi thu hoạch thử ở mỗi vùng 41 điểm, ta thu được kết quả sau :

Vùng A : năng suất trung bình 40 tạ/ha; độ lệch chuẩn 1,5 tạ/ha;

Vùng B : năng suất trung bình 39,7 tạ/ha; độ lệch chuẩn 1,2 tạ/ha.

Với mức ý nghĩa 5%, có thể cho rằng, độ ổn định về năng suất ở hai vùng là như nhau hay không? Giả sử năng suất lúa phân phối theo quy luật chuẩn.

Giải

Gọi X là năng suất lúa vùng A: $X \sim N(\mu_X, \sigma_Y^2)$

Y là năng suất lúa vùng B: $Y \sim N(\mu_Y, \sigma_Y^2)$

Ta có: $n = 41; S_X^2 = (1,5)^2 = 2,25$; $m = 41; S_Y^2 = (1,2)^2 = 1,44$

Cặp giả thuyết thống kê: $\begin{cases} H_0:\ \sigma_X^2 = \sigma_Y^2 \\ H_1:\ \sigma_X^2 \neq \sigma_Y^2 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$F = \frac{S_X^2}{S_Y^2} \sim F(n-1;m-1)$$

Thay $S_X^2 = 2,25$; $S_Y^2 = 1,44$, ta được $F_{qs} = \frac{2,25}{1,44} = 1,3625$

Với mức ý nghĩa $\alpha = 5\% = 0,05$ suy ra giá trị tới hạn

$$f_{0,025}(40;40) = 1,88\ ;\ f_{0,975}(40;40) = \frac{1}{f_{0,025}(40;40)} = \frac{1}{1,88} = 0,53$$

Miền bác bỏ: $W_\alpha = (-\infty; 0,53) \cup (1,88; +\infty)$

Do $F_{qs} \notin W_\alpha$ nên chưa có cơ sở để bác bỏ H_0, có thể cho rằng năng suất lúa ở hai vùng ổn định như nhau.

Ví dụ 4.15. Quan sát sức nặng của bé trai (X) và bé gái (Y) lúc sơ sinh (đơn vị gam), ta có kết quả

Khối lượng	3000-3200	3200-3400	3400-3600	3600-3800	3800-4000
Số bé trai	1	3	8	10	3
Số bé gái	2	10	10	5	1

So sánh các phương sai tổng thể. Kết luận với mức ý nghĩa 5%.

Giải

Từ số liệu của mẫu, ta có

$$n = 25,\ S_X^2 = 40266,67,\ \ m = 28,\ S_Y^2 = 37407,41 \qquad .$$

Ta xét bài toán kiểm định

$$\begin{cases} H_0: \sigma_X^2 = \sigma_Y^2 \\ H_1: \sigma_X^2 > \sigma_Y^2 \end{cases}$$

Nếu giả thuyết H_0 đúng thì ta có thống kê

$$F = \frac{S_X^2}{S_Y^2} \sim F(n-1, m-1).$$

Thay $S_X^2 = 2,25$; $S_Y^2 = 1,44$, ta được $F_{qs} = \dfrac{40266,67}{37407,41} = 1,076$

Với mức ý nghĩa $\alpha = 5\% = 0,05$ suy ra giá trị tới hạn $f_{0,05}(24; 27) = 1,89$

Miền bác bỏ: $W_\alpha = (1,89; +\infty)$

Do $F_{qs} \notin W_\alpha$ nên chưa đủ cơ sở bác bỏ H_0, vậy với mức ý nghĩa 5%, phương sai như nhau.

4.6. Kiểm định phi tham số

4.6.1. Kiểm định về tính độc lập

4.6.1.1. So sánh bộ số liệu quan sát với bộ số liệu lý thuyết.

Trong trường hợp này, với một bộ số liệu quan sát,

$$N_1,\ N_2,\ ...,\ N_r,$$

ta cần so sánh nó với bộ số liệu lý thuyết

$$N_1^{/},\ N_2^{/},\ ...,\ N_r^{/},$$

trong đó bộ số liệu lý thuyết này được tính theo quy luật phân phối các phạm trù trong tổng thể cho trước.

Ta mô hình kiểm định sau

a. $\begin{cases} H_0\text{: Các bộ số liệu quan sát và lý thuyết giống nhau} \\ H_1\text{ : Các bộ số liệu quan sát và lý thuyết khác nhau} \end{cases}$

$$Q = \sum_{i=1}^{r} \frac{\left(N_i - N_i^{/}\right)^2}{N_i^{/}} = \frac{\left(N_1 - N_1^{/}\right)^2}{N_1^{/}} + \frac{\left(N_2 - N_2^{/}\right)^2}{N_2^{/}} + ... + \frac{\left(N_r - N_r^{/}\right)^2}{N_r^{/}} \qquad (4.35)$$

thì phân phối xác suất của Q cho bởi :

$Q \sim \chi^2(r-1)$, nếu không có tham số nào cần ước lượng trong quá trình tính các số liệu lý thuyết.

$Q \sim \chi^2(r-k-1)$, nếu có k tham số cần ước lượng trong quá trình tính các số liệu lý thuyết. Chẳng hạn, nếu ta cần ước lượng trung bình tổng thể μ, thì $Q \sim \chi^2(r-2)$; nếu ta cần ước lượng cả trung bình lẫn phương sai tổng thể $(\mu;\sigma^2)$ thì $Q \sim \chi^2(r-3)$.

b. Với nguy cơ sai lầm α, hay độ tin cậy $\gamma = 1-\alpha$, cho trước, ta tìm được khoảng tin cậy $[0, C]$ cho Q.

c. Tính giá trị cụ thể của Q.

d. So sánh Q với khoảng tin cậy :

$Q > C$: bác bỏ H_0,

$Q \le C$: chấp nhận H_1.

4.6.1.2. So sánh các bộ số liệu quan sát với nhau.

Trong trường hợp này, ta so sánh k bộ số liệu quan sát

$$N_{1,1},\ N_{1,2},\ ...,\ N_{1,r},$$

$$N_{2,1},\ N_{2,2},\ ...,\ N_{2,r},\ldots,$$

$$N_{k,1},\ N_{k,2},\ ...,\ N_{k,r}.$$

với nhau mà người ta còn gọi là so sánh các số liệu trong một bảng :

	P. trù 1	P. trù 2	...	P. trù r
Bộ 1	$N_{1,1}$	$N_{1,2}$	...	$N_{1,r}$

Bộ 2	$N_{2,1}$	$N_{2,2}$	...	$N_{2,r}$
...				
Bộ k	$N_{k,1}$	$N_{k,2}$	...	$N_{k,r}$.

Người ta quy bài toán so sánh k bộ số liệu này với nhau (mỗi bộ số liệu gồm r phạm trù) bằng cách coi nó như là một bộ số liệu gồm $k \times r$ phạm trù và so sánh nó với một bộ số liệu lý thuyết.

Muốn vậy, ta thành lập các tổng hàng và tổng cột như sau

	P. trù 1	P. trù 2	...	P. trù r	$\sum$
Bộ 1	$N_{1,1}$	$N_{1,2}$	...	$N_{1,r}$	$\mathbf{H}_1$
Bộ 2	$N_{2,1}$	$N_{2,2}$	...	$N_{2,r}$	$\mathbf{H}_2$
...	...	...	...	...	...
Bộ k	$N_{k,1}$	$N_{k,2}$	...	$N_{k,r}$.	$\mathbf{H}_k$
$\sum$	$\mathbf{C}_1$	$\mathbf{C}_2$	...	$\mathbf{C}_r$	**N**

trong đó

$\mathbf{H}_i = \sum_{j=1}^{r} \mathbf{N}_{i,j}$ là tổng số số liệu quan sát của bộ thứ i (hàng thứ i), với $i = 1, 2, ..., k$.

$\mathbf{C}_j = \sum_{i=1}^{k} \mathbf{N}_{i,j}$ là tổng số số liệu quan sát của phạm trù thứ j (cột thứ j), với $j = 1, 2, ..., r$.

$\mathbf{N} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{H}_i = \sum_{\mathbf{i=1}}^{\mathbf{k}} \sum_{j=1}^{r} \mathbf{N}_{i,j} = \sum_{j=1}^{\mathbf{r}} \mathbf{C}_j = \sum_{j=1}^{\mathbf{r}} \sum_{i=1}^{k} \mathbf{N}_{i,j}$ là tổng số toàn bộ số liệu quan sát. Khi đó, nếu k bộ số liệu này là như nhau, thì

tỷ lệ của các phạm trù 1 trong tổng thể là $p_1 = \frac{C_1}{N}$,

tỷ lệ của các phạm trù 2 trong tổng thể là $p_2 = \frac{C_2}{N}$,

...

tỷ lệ của các phạm trù r trong tổng thể là $p_r = \frac{C_r}{N}$.

Từ đó, ta xây dựng được bộ số liệu lý thuyết gồm $k \times r$ số hạng liệt kê trong bảng như sau

	P. trù 1	P. trù 2	...	P. trù r
Bộ 1	$N'_{1,1}$	$N'_{1,2}$	...	$N'_{1,r}$
Bộ 2	$N'_{2,1}$	$N'_{2,2}$	...	$N'_{2,r}$
...				
Bộ k	$N'_{k,1}$	$N'_{k,2}$	...	$N'_{k,r}$.

trong đó

$$N'_{i,j} = p_j \times \mathbf{H}_i = \frac{\mathbf{C}_j \times \mathbf{H}_i}{\mathbf{N}}.$$

Chú ý rằng khi đó, ta dùng thống kê

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{\left(\mathbf{N}_{i,j} - \mathbf{N}'_{i,j}\right)^2}{\mathbf{N}'_{i,j}} \tag{4.36}$$

và vì trong quá trình thành lập bộ số liệu lý thuyết, ta đã ước lượng $k+r-2$ số hạng nên ta có mô hình kiểm định sau

a. $\begin{cases} H_0: \text{Các bộ số liệu quan sát và lý thuyết giống nhau} \\ H_1: \text{Các bộ số liệu quan sát và lý thuyết khác nhau} \end{cases}$

Nếu H_0 đúng thì $Q \sim \chi^2\left[k \times r - (k+r-2) - 1\right] \equiv \chi^2\left[(k-1)\times(r-1)\right]$.

b. Với nguy cơ sai lầm α, hay độ tin cậy $\gamma = 1-\alpha$, cho trước, ta tìm được khoảng tin cậy $[0,C]$ cho Q.

c. Tính giá trị cụ thể của Q.

d. So sánh Q với khoảng tin cậy :

$Q > C$: bác bỏ H_0,

$Q \le C$: chấp nhận H_1 .

Chú ý :

i) Công thức tính Q nêu trên còn có thể viết lại thành

$$Q = \mathbf{N} \times \left(\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{\left(\mathbf{N}_{i,j}\right)^2}{\mathbf{H}_i \times \mathbf{C}_j} - 1\right).$$

ii) Trường hợp ta so sánh hai bảng số liệu nhưng lại so sánh từng cặp số liệu với nhau. Chẳng hạn, với hai bảng số liệu

X_1	X_2	...	X_n
Y_1	Y_2	...	Y_n

ta không phải ta so sánh hai bộ số liệu với nhau mà là so sánh sự sai khác của từng cặp số liệu xem nó có ý nghĩa không. Do vậy, ta xét hiệu số

$$D_i = X_i - Y_i, \text{ với } i = 1, 2, ..., n.$$

Khi đó, giả thiết H_0 : "Hai bộ số liệu giống nhau từng cặp"

được thay bằng

H_0 : "trung bình của bộ số liệu D_i bằng 0",

và ta nhận được bài toán so sánh trung bình (của một bộ số liệu) với một số (số 0).

Tóm lại, ta có mô hình cho bài toán so sánh hai bộ số liệu từng cặp này như sau

Lập hiệu số từng cặp của hai bộ số liệu $D = X - Y$.

$$\begin{cases} H_0: \text{Hai bộ số liệu giống nhau từng cặp} \\ H_1: \text{Hai bộ số liệu khác nhau từng cặp} \end{cases}$$

và ta nhận được giả thiết tương đương

$$\begin{cases} H_0: \mu_D = 0 \\ H_1: \mu_D \neq 0 \end{cases}.$$

Ví dụ 4.16. Có một lô hàng mà người giao hàng cho biết tỷ lệ hỏng 0,10; thứ phẩm 0,30; đạt 0,40; tốt 0,20. Ta kiểm tra một số trường hợp thấy có 25 sản phẩm hỏng; 50 thứ phẩm; 50 sản phẩm đạt; 25 sản phẩm tốt. Hỏi rằng lời người giao hàng nói có đúng không ? (kết luận với $\alpha = 5\%$)

Giải

Ta có bài toán kiểm định

a) $\begin{cases} H_0: \text{người giao hàng nói đúng} \\ H_1: \text{người giao hàng nói không đúng} \end{cases}$

Ta có bảng phân phối tần số quan sát

Hỏng	Thứ phẩm	đạt	tốt
25	50	50	25

Nếu H_0 đúng, thì trên tổng số 150 sản phẩm kiểm tra, ta được bảng tần số lý thuyết

Hỏng	Thứ phẩm	đạt	tốt
$0,1 \times 150 = 15$	$0,3 \times 150 = 45$	$0,4 \times 150 = 60$	$0,2 \times 150 = 30$

và khi đó

$$Q = \sum_{i=1}^{4} \frac{\left(N_i - N_i^{/}\right)^2}{N_i^{/}} \sim \chi^2(3)$$

với N_i là số liệu quan sát và $N_i^{/}$ là số liệu lý thuyết.

b) Với nguy cơ sai lầm $\alpha = 5\% = 0,05,$ ta có $C = \chi^2_{0,05}(3) = 7,815.$

c) Thế các số liệu quan sát và lý thuyết vào biểu thức (1), ta nhận được $Q = 9,7222$

d) Ta có $Q = 9,7222 > C = 7,815$. Do đó, ta từ chối H_0, nghĩa là người giao hàng nói không đúng.

Ví dụ 4.17. Quan sát ngẫu nhiên một số trường hợp trong 3 lô thuốc (rất nhiều), ta ghi nhận được

	Tốt	Tạm dùng	Hỏng
Lô A	125	52	23
Lô B	117	61	22
Lô C	178	97	25

Hỏi rằng chất lượng của 3 lô thuốc có như nhau không ? Kết luận với mức ý nghĩa 5%.

Giải

Ta có

	Tốt	Tạm dùng	Hỏng	Tổng
Lô A	125	52	23	200
Lô B	117	61	22	200
Lô C	178	97	25	300
Tổng	420	210	70	700

Xét bài toán kiểm định

$$\begin{cases} H_0: \text{Chất lượng ba lô như nhau} \\ H_1: \text{Chất lượng ba lô khác nhau} \end{cases}$$

Nếu H_0 đúng thì

$$P(\text{tốt}) = \frac{125 + 117 + 178}{700} = 0,6,$$

$$P(\text{tạm}) = \frac{52 + 61 + 97}{700} = 0,3, \text{ và}$$

$$P(\text{hỏng}) = \frac{23 + 22 + 25}{700} = 0,1.$$

Khi đó, bảng phân phối tần số lý thuyết phải là

	Tốt	Tạm dùng	Hỏng
Lô A	$0,6\times 200=120$	$0,3\times 200=60$	$0,1\times 200=20$
Lô B	$0,6\times 200=120$	$0,3\times 200=60$	$0,1\times 200=20$
Lô C	$0,6\times 300=180$	$0,3\times 300=90$	$0,1\times 300=30$

Độ khác biệt giữa quan sát và lý thuyết là

$$Q=\sum\frac{(N-N')^2}{N'}=\frac{25}{120}+\frac{64}{60}+\frac{9}{20}+\frac{9}{120}+\frac{1}{60}+\frac{4}{20}+\frac{4}{180}+\frac{49}{90}+\frac{25}{30}=3,42.$$

Nếu H_0 đúng thì $Q\sim\chi^2\left[(3-1)(3-1)\right]=\chi^2(4)$, với mức ý nghĩa $\alpha=5\%=0,05$, ta có $C=\chi^2_{0,05}(4)=9,488$. Vì $Q\le C$, ta chấp nhận H_0, nghĩa là 3 lô thuốc như nhau.

Chú ý : Ta có thể thành lập trực tiếp bảng phân phối tần số lý thuyết như sau

	Tốt	Tạm dùng	Hỏng	Tổng
Lô A	$\frac{420\times 200}{700}$	$\frac{210\times 200}{700}$	$\frac{70\times 200}{700}$	*200*
Lô B	$\frac{420\times 200}{700}$	$\frac{210\times 200}{700}$	$\frac{70\times 200}{700}$	*200*
Lô C	$\frac{420\times 300}{700}$	$\frac{210\times 300}{700}$	$\frac{70\times 300}{700}$	*300*
Tổng	*420*	*210*	*70*	700

Hơn nữa, ta có thể dùng trực tiếp công thức

$$Q=700\left(\frac{(125)^2}{200\times 420}+\frac{(52)^2}{200\times 210}+...+\frac{(25)^2}{70\times 300}-1\right)=3,42.$$

Ví dụ 4.18. Trong một công ty, người ta chọn ngẫu nhiên 1000 công nhân và theo dõi số ngày nghỉ của họ trong một năm. Kết quả thu được :

Giới tính Số ngày nghỉ	Nữ	Nam
0 – 5	300	500
5 – 20	80	70
> 20	20	30

Với mức ý nghĩa 1%, hãy kiểm định giả thiết cho rằng sự nghỉ việc không phụ thuộc vào giới tính.

Giải

Ta có bài toán kiểm định

$$\begin{cases} H_0: \text{Sự nghỉ việc không phụ thuộc vào giới tính} \\ H_1: \text{Sự nghỉ việc phụ thuộc vào giới tính} \end{cases}$$

Gọi A là nữ

$$P(A) = \frac{300+80+20}{1000} = 0,4;\ P(\bar{A}) = \frac{500+70+30}{1000} = 0,6.$$

Khi đó, bảng phân phối tần số lý thuyết là

Giới tính Số ngày nghỉ	Nữ	Nam
0 – 5	320	480
5 – 20	60	150
> 20	20	50

Độ khác biệt giữa quan sát và lý thuyết là

$$Q = \sum \frac{(N-N')^2}{N'} = \frac{400}{320} + \frac{400}{480} + \frac{400}{60} + \frac{400}{90} = 13,194.$$

Nếu H_0 đúng thì $Q \sim \chi^2(3-1)(2-1) = \chi^2(2)$. Với mức ý nghĩa $\alpha = 1\% = 0,01$, ta có $C = \chi^2_{0,01}(2) = 9,21$. Vì $Q > C$, nên ta bác bỏ H_0, nghĩa là sự nghỉ việc phụ thuộc vào giới tính.

Ví dụ 4.19. Nghiên cứu ảnh hưởng của hoàn cảnh gia đình đối với tình hình phạm tội của trẻ em vị thành niên, người ta thu được.

Hoàn cảnh gia đình Tình trạng phạm tội	Bố hoặc mẹ đã chết	Bố mẹ ly hôn	Còn cả bố mẹ
Không phạm tội	20	25	13
Phạm tội	29	43	18

Với mức ý nghĩa 5%, có thể kết luận là hoàn cảnh gia đình của trẻ em độc lập với tình trạng phạm tội hay không.

Giải

Gọi X : Bố hoặc mẹ đã chết, Y : Bố mẹ ly hôn, Z : còn cả bố mẹ.

Ta có bài toán kiểm định

$$\begin{cases} H_0: \text{Hoàn cảnh gia đình độc lập với tình trạng xã hội} \\ H_1: \text{Hoàn cảnh gia đình không độc lập với tình trạng xã hội} \end{cases}$$

Nếu H_0 đúng thì

$$P(X) = \frac{20+29}{148} = 0,331,\ P(Y) = \frac{25+43}{148} = 0,459,\ P(Z) = \frac{13+18}{148} = 0,209.$$

Ta có bảng phân phối tần số lý thuyết như sau

Hoàn cảnh gia đình / Tình trạng phạm tội	X	Y	Z
Không phạm tội	19	27	12
Phạm tội	30	41	19

Độ khác biệt giữa quan sát và lý thuyết là

$$Q = \sum \frac{(N - N')^2}{N'} = \frac{1}{19} + \frac{4}{27} + \frac{1}{12} + \frac{1}{30} + \frac{4}{41} + \frac{1}{19} = 0,468.$$

Nếu H_0 đúng thì $Q \sim \chi^2(2-1)(3-1) = \chi^2(2)$, với mức ý nghĩa $\alpha = 5\% = 0,05$, ta có $C = \chi^2_{0,05}(2) = 5,991$. Vì $Q \le C$, nên ta chấp nhận H_0, nghĩa là hoàn cảnh gia đình độc lập với tình trạng phạm tội.

Ví dụ 4.20. Thống kê số tai nạn lao động tại 2 xí nghiệp, ta có các số liệu sau :

Xí nghiệp	Số công nhân	Số tai nạn lao động
A	200	20
B	800	120

Với mức ý nghĩa 5%, hãy kết luận xem chất lượng công tác bảo vệ an toàn lao động tại hai xí nghiệp trên có khác nhau không?

Giải

Ta có bài toán kiểm định

$$\begin{cases} H_0: \text{Chất lượng bảo vệ an toàn của xí nghiệp là như nhau} \\ H_1: \text{Chất lượng bảo vệ an toàn của xí nghiệp là khác nhau} \end{cases}$$

Nếu H_0 đúng thì

$$P(\text{công nhân}) = \frac{200+800}{1140} = 0,8772;\ P(\text{tai nạn}) = \frac{20+120}{1140} = 0,123.$$

Ta có bảng phân phối tần số lý thuyết

Xí nghiệp	Số công nhân	Số tai nạn lao động
A	193	27
B	807	113

Độ khác biệt giữa quan sát và lý thuyết là

$$Q = \sum \frac{\left(\left|N - N'\right| - 0,5\right)^2}{N'} = \frac{42,25}{193} + \frac{42,25}{27} + \frac{42,25}{807} + \frac{42,25}{113} = 2,21.$$

Nếu H_0 đúng thì $Q \sim \chi^2(2-1)(2-1) = \chi^2(1)$, với mức ý nghĩa $\alpha = 5\% = 0,05$, ta có $C = \chi^2_{0,05}(1) = 3,841$. Vì $Q \le C$, nên ta chấp nhận H_0, nghĩa là chất lượng bảo vệ an toàn lao động của hai xí nghiệp là như nhau.

4.6.2. Kiểm định về tính phù hợp (hay về luật phân phối)

Bài toán: Giả sử ta cần kiểm định một biến ngẫu nhiên X xem biến ngẫu nhiên X thuộc phân phối nào như: phân phối Poisson, nhị thức, siêu bội, chuẩn,… Dựa vào thông tin trên mẫu, đưa ra kết luận với mức ý nghĩa α.

Giải quyết bài toán:

Bước 1. Phát biểu bài toán kiểm định

H_0: X có luật phân phối đã cho

H_1: X không có luật phân phối xác suất như lý thuyết H_0 đã nêu.

Bước 2. Tính các xác suất

Gọi $P_j = P(X = x_j)$ là xác suất của X tại giá trị x_j nếu X là biến ngẫu nhiên rời rạc và trong trường hợp X là biến ngẫu nhiên liên tục thì $P_j = P(x_{j-1} \le X \le x_j)$.

Bước 3. Tính giá trị kiểm định

$$Q = \sum_{i=1}^{k} \frac{(n_i - n_i P_i)^2}{n_i P_i}$$ với k là số nhóm tính chất.

Bước 4. Tìm giá trị tới hạn, $C = \chi^2_\alpha(k-1)$ trong bảng Chi bình phương.

Bước 5. Nếu $Q \le C$ thì chấp nhận H_0 và ngược lại.

4.6.2.1. Kiểm định về luật phân phối Poisson, $P(\mu)$

Ví dụ 4.21. Đếm số hồng cầu X trong mỗi ô vuông của một hồng cầu kế, ta ghi được

Lượng hồng cầu	0	1	2	3	4	5	6	7	8	9
Số ô	3	8	12	20	20	14	10	8	4	1

Hỏi X có tuân theo phân phối Poisson hay không? Kết luận với mức ý nghĩa 5%.

Giải

Bước 1. Bài toán kiểm định

H_0: X tuân theo luật phân phối Poisson (với $n = 3$ và μ chưa biết)

H_1 : X không tuân theo luật phân phối Poisson.

Bước 2. Do μ không biết nên ta thay thế μ bằng ước lượng không chệch của nó là $\overline{X}$ ($\overline{X}$ là trung bình số hồng cầu trong mẫu), ta có

$$\mu = \overline{X} = \frac{1}{n}\sum_{i=1}^{10} n_i X_i = 3,99 \approx 4$$

Nếu H_0 đúng thì $X \sim P(4)$, hàm xác suất của X là $f(x) = e^{-x}\dfrac{4^x}{x!}$, $x = 0,1,2,...$

Từ hàm xác suất, ta có thể tính được xác suất mỗi trường hợp xảy ra $P_k = e^{-k}\dfrac{4^k}{k!}$, $k = 0,1,2,...$ rồi suy ra tần số lý thuyết $n_i^{/} = nP_i$.

Bước 3. Tính giá trị kiểm định: $Q = \sum_{i=1}^{k} \dfrac{\left(n_i - n_i P_i\right)^2}{n_i P_i}$

Vì số lần quan sát ở trường hợp $X = 0$ và $X = 8, 9$ quá bé, nên ta sát nhập với các trường hợp bên cạnh

Lượng hồng cầu	≤ 1	2	3	4	5	6	≥ 7
Số ô	11	12	20	20	14	10	13

Ta lập bảng để tính các giá trị như sau:

X	n_i	P_i	$n_i^{/} = nP_i$	$\dfrac{\left(n_i - n_i^{/}\right)^2}{n_i^{/}}$
0 và 1	11	0,09	9	0,4444
2	12	0,15	15	0,6
3	20	0,2	20	0
4	20	0,2	20	0
5	14	0,16	16	0,25
6	10	0,1	10	0
7, 8 và 9	13	0,1	10	0,9
	100			$Q = 2,1944$

Bước 4. Với mức ý nghĩa 5%, ta có $C = \chi^2_{0,05}(7-2) = 11,07$

Bước 5. Ta có $Q \le C$ nên chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, có thể xem X là đại lượng ngẫu nhiên có phân phối Poisson.

4.6.2.2. Kiểm định về luật phân phối nhị thức, $B(n;p)$

Ví dụ 4.22. Sản phẩm sản xuất ra trên một dây chuyền tự động được đóng gói một cách ngẫu nhiên theo quy cách: 3 sản phẩm/hộp. Với mức ý nghĩa 1%, có thể xem số sản phẩm loại I có trong hộp là đại lượng ngẫu nhiên có phân phối nhị thức hay không? Biết rằng kiểm tra ngẫu nhiên 100 hộp người ta thấy có 75 hộp có 3 sản phẩm loại I, 20 hộp có 2 sản phẩm loại I; 5 hộp có 1 sản phẩm loại I và không có hộp nào có số sản phẩm loại I là 0.

Giải

Bước 1. Gọi X là số sản phẩm loại I có trong 1 hộp. Bài toán kiểm định

H_0: X tuân theo luật phân phối nhị thức (với $n=3$ và p chưa biết)

H_1: X không tuân theo luật phân phối nhị thức.

Bước 2. Do p không biết nên ta thay thế p bằng ước lượng không chệch của nó là f (f là tỷ lệ sản phẩm loại I của mẫu), ta có

$$f=\frac{5\cdot 1+20\cdot 2+75\cdot 3}{100\cdot 3}=0,9$$

Vậy $P_j=P\left(X=x_j\right)$ được tính theo công thức:

$$P_j=P\left(X=x_j\right)=C_3^j\left(0,9\right)^j\left(0,1\right)^{3-j}$$

Bước 3. Tính giá trị kiểm định: $Q=\sum_{i=1}^{k}\frac{\left(n_i-n_iP_i\right)^2}{n_iP_i}$

Ta lập bảng để tính các giá trị như sau:

X	n_i	P_i	$n_i^{/}=nP_i$	$\frac{\left(n_i-n_i^{/}\right)^2}{n_i^{/}}$
0	0	0,001	0,1	0,1
1	5	0,027	2,7	1,96
2	20	0,243	24,3	0,76
3	75	0,729	72,9	0,06
Tổng	100			$Q=2,88$

Bước 4. Với mức ý nghĩa 1%, ta có $C=\chi^2_{0,01}\left(4-1\right)=11,344$

Bước 5. Ta có $Q\le C$ nên chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 1%, có thể xem số sản phẩm loại I có trong hộp là đại lượng ngẫu nhiên có phân phối nhị thức.

4.6.2.3. Kiểm định về luật phân phối chuẩn, $N\left(\mu,\sigma^2\right)$

Ví dụ 4.23. Quan sát sức nặng X (kg) của một nhóm người, ta ghi lại được

X	30-40	40-45	45-50	50-55	55-60	60-70
Số người	9	15	24	27	17	8

Hỏi X có phân phối theo luật chuẩn không? Kết luận với mức ý nghĩa 5%.

Giải

Bước 1. Bài toán kiểm định

H_0: X tuân theo luật phân phối chuẩn (μ,σ^2 chưa biết)

H_1: X không tuân theo luật phân phối chuẩn.

Bước 2. Do μ,σ^2 không biết nên ta thay thế μ,σ^2 bằng ước lượng không chệch của nó là $\overline{X}, S_X^2$ ($\overline{X}, S_X^2$ là trung bình và phương sai mẫu), ta có

$$\mu = \overline{X} = 49,715,$$

$$\sigma^2 = S_X^2 = 50,6831 = \left(7,1192\right)^2$$

Vậy $P_j = P\left(x_{j-1} \le X \le x_j\right)$ được tính theo công thức:

$$P_j = P\left(a \le X \le b\right) = \phi_0\left(\frac{b-40,715}{7,1192}\right) - \phi_0\left(\frac{a-49,715}{7,1192}\right)$$

Bước 3. Tính giá trị kiểm định: $Q = \sum_{i=1}^{k}\frac{\left(n_i - n_iP_i\right)^2}{n_iP_i}$

Ta lập bảng để tính các giá trị như sau:

Lớp	n_i	P_i	$n_i^{/} = nP_i$	$\frac{\left(n_i - n_i^{/}\right)^2}{n_i^{/}}$
≤ 40	9	0,0861	8,61	0,0177
40-45	15	0,1678	16,78	0,1888
45-50	24	0,2621	26,21	0,1863
50-55	27	0,2551	25,51	0,087
55-60	17	0,1547	15,47	0,1513
≥ 60	8	0,0743	7,43	0,0437
	100		100	$Q = 0,6748$

Bước 4. Với mức ý nghĩa 5%, ta có $C = \chi^2_{0,01}\left(6-2-1\right) = 7,815$

Bước 5. Ta có $Q \le C$ nên chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, có thể xem X là đại lượng ngẫu nhiên có phân phối chuẩn.

4.6.3. Kiểm định dấu và hạng Wilconxon

Trong phần 4.2, chúng ta đã tìm hiểu về kiểm định giá trị trung bình của một biến ngẫu nhiên X trên một tổng thể dựa vào giả thuyết biến X có luật phân phối chuẩn. Khi giả thuyết này bị vi phạm đồng nghĩa với biến X không tuân theo luật phân phối chuẩn. Do đó, việc kiểm định giả thuyết về giá trị trung bình thông qua việc so sánh $\mu = \mu_0$ sẽ không còn phù hợp nữa. Thay vào đó người ta sẽ lựa chọn giá trị đại diện tốt hơn cho biến X đó là trung vị, $Me(X)$. Trước khi đi vào phương pháp ta định nghĩa hạng (rank) của phần tử.

Giả sử ta có một dãy các số thực được xếp thứ tự tăng dần, trong dãy này không có giá trị nào bằng nhau: $x_1 < x_2 < ... < x_n$

Khi đó : $rank(x_1) = 1, rank(x_2) = 2, ..., rank(x_n) = n$

Nếu các phần tử có giá trị bằng nhau thì hạng của nó là hạng trung bình của các phần tử kế tiếp nhau.

4.6.3.1. Trường hợp mẫu nhỏ $(n < 20)$

Các bước tiến hành kiểm định như sau:

Bước 1. Đặt giả thuyết

- Kiểm định phi tham số hai phía (hai bên)

$$\begin{cases} H_0 : Me(X) = d_0 \\ H_1 : Me(X) \neq d_0 \end{cases}$$

- Kiểm định phi tham số một phía (phía phải)

$$\begin{cases} H_0 : Me(X) = d_0 \\ H_1 : Me(X) > d_0 \end{cases}$$

- Kiểm định phi tham số một phía (phía trái)

$$\begin{cases} H_0 : Me(X) = d_0 \\ H_1 : Me(X) < d_0 \end{cases}$$

Bước 2. Dựa vào mẫu thu thập được, ta lập bảng như sau :

Đặt $d_i = x_i - d_0, i = 1, 2, ..., n$

Giá trị quan sát (x_i)	Độ sai lệch (d_i)	$\lvert d_i \rvert$	Hạng	R^+	R^-

x_1	$x_1 - d_0$	$\lvert d_1 \rvert$			
x_2	$x_2 - d_0$	$\lvert d_2 \rvert$			
…					
x_n	$x_n - d_0$	$\lvert d_n \rvert$			

Trong đó

- $|d_i|$ là giá trị tuyệt đối của độ sai lệch.
- R^+ là hạng của các độ sai lệch với $d_i > 0$.
- R^- là hạng của các độ sai lệch với $d_i < 0$.

Bước 3. Tính giá trị kiểm định Wilconxon

- Nếu kiểm định hai phía thì $W_0 = \min\left\{\sum R^+, \sum R^-\right\}$
- Nếu kiểm định phía phải thì $W_0 = \sum R^+$
- Nếu kiểm định phía trái thì $W_0 = \sum R^-$

Bước 4. Kết luận

Với mức ý nghĩa α cho trước ta tìm được giá trị tới hạn trong bảng dấu hạng Wilconxon như sau: hai phía $W_{\alpha/2}^{n'} = (W_L; W_M)$ và một phía $W_\alpha^{n'} = (W_L; W_M)$ với n' là số phần tử có $d_i \neq 0$

- Đối với hai phía: Nếu $W_0 \notin (W_L; W_M)$ thì bác bỏ giả thuyết H_0.
- Đối với phía phải: Nếu $W_0 > W_M$ thì bác bỏ giả thuyết H_0.
- Đối với phía trái: Nếu $W_0 < W_L$ thì bác bỏ giả thuyết H_0.

Ví dụ 4.24. Khảo sát ngẫu nhiên mức lương (X: triệu đồng) của 10 sinh viên mới tốt nghiệp đại học của trường ta được kết quả được cho như sau:

Sinh viên	1	2	3	4	5	6	7	8	9	10
Mức lương	8,5	9	6	6,5	7	7,5	5,5	8	10	9,5

Với mức ý nghĩa 5%, hãy cho kết luận giả thuyết cho rằng thu nhập sinh viên tốt nghiệp đại học sau hai năm làm việc vượt quá 8 triệu đồng?

Giải

Bước 1. Đặt giả thuyết: H_0: "thu nhập của sinh viên sau hai năm tốt nghiệp đại học không vượt quá 8 triệu đồng"

$$\begin{cases} H_0 : Me(X) = 8 \\ H_1 : Me(X) > 8 \end{cases}$$

Bước 2. Dựa vào mẫu thu thập được, ta lập bảng và xếp hạng như sau:

Đặt $d_i = x_i - 8,\ i = 1, 2, ..., n$

Giá trị quan sát .	Độ sai lệch (d_i)	$\lvert d_i \rvert$	Hạng	R^+	R^-
8,5	0,5	0,5	1,5	1,5	
9	1	1	3,5	3,5	
6	-2	2	7,5		7,5
6,5	-1,5	1,5	5,5		5,5
7	-1	1	3,5		3,5
7,5	-0,5	0,5	1,5		1,5
5,5	-2,5	2,5	9		9
8	0	0			
10	2	2	7,5	7,5	
9,5	1,5	1,5	5,5	5,5	
Tổng				18	27

Bước 3. Tính giá trị kiểm định Wilconxon

Kiểm định phía phải thì $W_0 = \sum R^+ = 18$

Với mức ý nghĩa $\alpha = 0,05$ cho trước ta tìm được giá trị tới hạn trong bảng dấu hạng Wilconxon một phía $W_{\alpha}^{n'} = W_{0,05}^{9} = (8;37)$

Bước 4. So sánh và kết luận

Ta có $8 = W_L \leq W_0 = 18 \leq W_M = 37$ thì chưa đủ cơ sở bác bỏ giả thuyết H_0. Vậy với mức ý nghĩa 5%, thu nhập của sinh viên sau hai năm tốt nghiệp đại học không vượt quá 8 triệu đồng.

4.6.3.2. Trường hợp mẫu lớn $(n > 20)$

Các bước tiến hành kiểm định như sau:

Bước 1. Đặt giả thuyết

- Kiểm định phi tham số hai phía (hai bên)

$$\begin{cases} H_0 : Me(X) = d_0 \\ H_1 : Me(X) \neq d_0 \end{cases}$$

- Kiểm định phi tham số một phía (phía phải)

$$\begin{cases} H_0 : Me(X) = d_0 \\ H_1 : Me(X) > d_0 \end{cases}$$

- Kiểm định phi tham số một phía (phía trái)

$$\begin{cases} H_0 : Me(X) = d_0 \\ H_1 : Me(X) < d_0 \end{cases}$$

Bước 2. Dựa vào mẫu thu thập được, ta lập bảng và xếp hạng như sau:

Đặt $d_i = x_i - d_0$, $i = 1, 2, ..., n$

Giá trị quan sát (x_i)	Độ sai lệch (d_i)	$\|d_i\|$	Hạng	R^+	R^-
x_1	$x_1 - d_0$	$\|d_1\|$			
x_2	$x_2 - d_0$	$\|d_2\|$			
…					
x_n	$x_n - d_0$	$\|d_n\|$			

Trong đó

- $|d_i|$ là giá trị tuyệt đối của độ sai lệch.
- R^+ là hạng của các độ sai lệch với $d_i > 0$.
- R^- là hạng của các độ sai lệch với $d_i < 0$.

Bước 3. Tính giá trị kiểm định Z_0 theo phân phối chuẩn như sau:

- Trung bình : $\mu_W = \dfrac{n'(n'+1)}{4}$
- Phương sai: $\sigma_W^2 = \dfrac{n'(n'+1)(2n'+1)}{24}$

Suy ra

$$Z_0 = \frac{W_0 - \mu_W}{\sigma_W}$$

- n' là số phần tử có $d_i \neq 0$
- Nếu kiểm định hai phía thì $W_0 = \min\left\{\sum R^+, \sum R^-\right\}$
- Nếu kiểm định phía phải thì $W_0 = \sum R^+$

- Nếu kiểm định phía trái thì $W_0 = \sum R^-$

Bước 4. So sánh và kết luận

Với mức ý nghĩa α cho trước ta tìm được giá trị tới hạn trong bảng phân phối chuẩn: hai phía $z_{\alpha/2}$ và một phía z_α

- Đối với hai phía: Nếu $Z_0 > z_{\alpha/2}$ thì bác bỏ giả thuyết H_0.
- Đối với phía phải: Nếu $Z_0 > z_\alpha$ thì bác bỏ giả thuyết H_0.
- Đối với phía trái: Nếu $Z_0 < -z_\alpha$ thì bác bỏ giả thuyết H_0.

Ví dụ 4.25. Khảo sát ngẫu nhiên mức lương (X: triệu đồng) của 24 sinh viên mới tốt nghiệp đại học của trường ta được kết quả được cho như sau:

Sinh viên	1	2	3	4	5	6	7	8	9	10	11	12
Mức lương	8,5	9	6	6,5	7	7,5	5,5	8	10	9,5	5	6,5
Sinh viên	13	14	15	16	17	18	19	20	21	22	23	24
Mức lương	7,2	7,8	8,3	6,8	9,2	8,4	7,6	6	7,5	10,5	11	11,5

Với mức ý nghĩa 5%, hãy cho kết luận giả thuyết cho rằng thu nhập sinh viên tốt nghiệp đại học sau hai năm làm việc vượt quá 7,5 triệu đồng?

Giải

Bước 1. Đặt giả thuyết: H_0: "thu nhập của sinh viên sau hai năm tốt nghiệp đại học không vượt quá 7,5 triệu đồng"

$$\begin{cases} H_0 : Me(X) = 7,5 \\ H_1 : Me(X) > 7,5 \end{cases}$$

Bước 2. Dựa vào mẫu thu thập được, ta lập bảng và xếp hạng như sau:

Đặt $d_i = x_i - 7,5,\ i = 1,2,...,24$

Giá trị quan sát (x_i)	Độ sai lệch (d_i)	$\lvert d_i \rvert$	Hạng	R^+	R^-
8,5	1	1	10	10	
9	1,5	1,5	13	13	
6	-1,5	1,5	13		13
6,5	-1	1	10		10
7	-0,5	0,5	4,5		4,5
7,5	0	0			
5,5	-2	2	16,5		16,5

8	0,5	0,5	4,5	4,5	
10	2,5	2,5	18,5	18,5	
9,5	2	2	16,5	16,5	
5	-2,5	2,5	18,5		18,5
6,5	-1	1	10		10
7,2	-0,3	0,3	2,5		2,5
7,8	0,3	0,3	2,5	2,5	
8,3	0,8	0,8	7	7	
6,8	-0,7	0,7	6		6
9,2	1,7	1,7	15	15	
8,4	0,9	0,9	8	8	
7,6	0,1	0,1	1	1	
6	-1,5	1,5	13		13
7,5	0	0			
10,5	3	3	20	20	
11	3,5	3,5	21	21	
11,5	4	4	22	22	
Tổng				159	94

Bước 3. Tính giá trị kiểm định Wilconxon

Kiểm định phía phải thì $W_0 = \sum R^+ = 159$

- Trung bình : $\mu_W = \dfrac{n'\left(n'+1\right)}{4} = \dfrac{22(22+1)}{4} = 126,5$

- Phương sai: $\sigma_W^2 = \dfrac{n'\left(n'+1\right)\left(2n'+1\right)}{24} = \dfrac{22(22+1)(44+1)}{24} = 948,75$

Suy ra

$$Z_0 = \frac{W_0 - \mu_W}{\sigma_W} = \frac{159 - 126,5}{\sqrt{948,75}} = 1,06$$

Với mức ý nghĩa $\alpha = 0,05$ cho trước ta tìm được giá trị tới hạn trong bảng phân phối Gauss: $z_\alpha = z_{0,05} = 1,64$

Bước 4. So sánh và kết luận

Ta có $Z_0 = 1,06 \le z_{0,05} = 1,64$ thì chưa đủ cơ sở bác bỏ giả thuyết H_0. Vậy với mức ý nghĩa 5%, thu nhập của sinh viên sau hai năm tốt nghiệp đại học không vượt quá 7,5 triệu đồng.

4.6.3.3. Trường hợp mẫu phối hợp từng cặp và mẫu nhỏ $(n < 20)$

Các bước tiến hành kiểm định như sau:

Bước 1. Đặt giả thuyết

- Kiểm định phi tham số hai phía

$$\begin{cases} H_0 : Me(X_d) = d_0 \\ H_1 : Me(X_d) \ne d_0 \end{cases}$$

- Kiểm định phi tham số phía phải

$$\begin{cases} H_0 : Me(X_d) = d_0 \\ H_1 : Me(X_d) > d_0 \end{cases}$$

- Kiểm định phi tham số phía trái

$$\begin{cases} H_0 : Me(X_d) = d_0 \\ H_1 : Me(X_d) < d_0 \end{cases}$$

Với

$Me(X_d) = d_1 - d_2$

d_1 là giá trị trung vị của mẫu thứ 1

d_2 là giá trị trung vị của mẫu thứ 2

d_0 là một hằng số (thông thường thì $d_0 = 0$)

Bước 2. Dựa vào mẫu thu thập được, ta lập bảng và xếp hạng như sau:

Đặt $d_i = x_i - y_i - d_0$, $i = 1,2,...,n$

Số thứ tự	Mẫu phối hợp từng cặp		Độ sai lệch (d_i)	$\lvert d_i \rvert$	Hạng	R^+	R^-
	Mẫu 1	Mẫu 2					
1	x_1	y_1	$x_1 - y_1 - d_0$	$\lvert d_1 \rvert$			
2	x_2	y_2	$x_2 - y_2 - d_0$	$\lvert d_2 \rvert$			
…	…	…					
n	x_n	y_n	$x_n - y_n - d_0$	$\lvert d_n \rvert$			

Trong đó

- $|d_i|$ là giá trị tuyệt đối của độ sai lệch.
- R^+ là hạng của các độ sai lệch với $d_i > 0$.
- R^- là hạng của các độ sai lệch với $d_i < 0$.

Bước 3. Tính giá trị kiểm định Wilconxon

- Nếu kiểm định hai phía thì $W_0 = \min\left\{\sum R^+, \sum R^-\right\}$
- Nếu kiểm định phía phải thì $W_0 = \sum R^+$
- Nếu kiểm định phía trái thì $W_0 = \sum R^-$

Bước 4. So sánh và kết luận

Với mức ý nghĩa α cho trước ta tìm được giá trị tới hạn trong bảng dấu hạng Wilconxon như sau: hai phía $W_{\alpha/2}^{n'} = (W_L; W_M)$ và một phía $W_{\alpha}^{n'} = (W_L; W_M)$ với n' là số phần tử có $d_i \neq 0$

- Đối với hai phía: Nếu $W_0 \notin (W_L; W_M)$ thì bác bỏ giả thuyết H_0.
- Đối với phía phải: Nếu $W_0 > W_M$ thì bác bỏ giả thuyết H_0.
- Đối với phía trái: Nếu $W_0 < W_L$ thì bác bỏ giả thuyết H_0.

Lưu ý : Trong trường hợp mẫu phối hợp mẫu lớn $(n > 20)$. Tính giá trị kiểm định W_0 ở bước 3 được thay bằng Z_0 theo phân phối chuẩn như sau:

$$\mu_W = \frac{n'(n'+1)}{4} \text{ và } \sigma_W^2 = \frac{n'(n'+1)(2n'+1)}{24}$$

Suy ra

$$Z_0 = \frac{W_0 - \mu_W}{\sigma_W}$$

- n' là số phần tử có $d_i \neq 0$
- Nếu kiểm định hai phía thì $W_0 = \min\left\{\sum R^+, \sum R^-\right\}$
- Nếu kiểm định phía phải thì $W_0 = \sum R^+$
- Nếu kiểm định phía trái thì $W_0 = \sum R^-$

và bước 4. Với mức ý nghĩa α cho trước ta tìm được giá trị tới hạn trong bảng phân phối chuẩn, $z_{\alpha/2}$ và một phía z_{α}

- Đối với hai phía: Nếu $|Z_0| > z_{\alpha/2}$ thì bác bỏ giả thuyết H_0.
- Đối với phía phải: Nếu $Z_0 > z_{\alpha}$ thì bác bỏ giả thuyết H_0.
- Đối với phía trái: Nếu $Z_0 < -z_{\alpha}$ thì bác bỏ giả thuyết H_0.

Ví dụ 4.26. Chọn ngẫu nhiên 9 khách hàng và yêu cầu họ cho biết sở thích về hai sản phẩm A, B thông qua thang điểm từ 1 đến 5. Ta có kết quả như sau:

Khách hàng	1	2	3	4	5	6	7	8	9
Sản phẩm A	4	5	2	3	3	1	3	2	2
Sản phẩm B	3	5	5	2	5	5	3	5	5

Giả thuyết rằng không có xu hướng nghiêng về loại sản phẩm nào trong sở thích của hai loại sản phẩm A, B. Với mức ý nghĩa 5%, hãy cho biết có sự khác biệt về sở thích hai loại sản phẩm hay không?

Giải

Đây là mẫu thu thập dưới dạng phối hợp từng cặp vì đánh giá sở thích hai loại sản phẩm A, B không thể cho khách hàng 1 dùng thử sản phẩm A và khách hàng 2 dùng thử sản phẩm B rồi đưa ra đánh giá sản phẩm nào ưa thích hơn vì khách hàng 1 và 2 khác nhau hoàn toàn về mức độ ưa thích trong quá trình sử dụng sản phẩm A hay B.

Bước 1. Đặt giả thuyết

Vì đây là bài toán chỉ yêu cầu kiểm định xem có sự khác nhau về mức độ ưa thích của hai sản phẩm A và B nên $d_0 = 0$ trong trường hợp này ta có bài toán kiểm định

$$\begin{cases} H_0 : Me_d = 0 \\ H_1 : Me_d \neq 0 \end{cases}$$

Với

$Me_d = d_A - d_B$

d_A là giá trị trung vị của sản phẩm A

d_B là giá trị trung vị của sản phẩm B

Bước 2. Dựa vào mẫu lập bảng và xếp hạng

Khách hàng	1	2	3	4	5	6	7	8	9	Tổng
Sản phẩm A	4	5	2	3	3	1	3	2	2	
Sản phẩm B	3	5	5	2	5	5	3	5	5	

d_i	1	0	-3	1	-2	-4	0	-3	-3	
$\lvert d_i \rvert$	0	0	1	1	2	3	3	3	4	
			1	2	3	4	5	6	7	
R^+			1,5	1,5						3
R^-					3	5	5	5	7	25

Bước 3. Tính giá trị kiểm định Wilconxon vì đây là kiểm định hai phía nê

$$W_0 = \min\left\{\sum R^+, \sum R^-\right\} = \min\{3, 25\} = 3$$

Bước 4. So sánh kết luận

Với mức ý nghĩa 5%, ta tìm được giá trị tới hạn trong bảng dấu và hạng Wilconxon như sau $w_{\alpha/2}^{n'} = w_{0,025}^{7} = (2;26)$. Ta có $W_0 = 3 \in W_{0,025}^{7} = (2;26)$, chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, không có sự khác biệt về việc ưu thích hai loại sản phẩm A, B.

4.6.4. Kiểm định tổng và hạng Wilconxon

Để so sánh sự khác biệt hoặc hơn kém trên hai mẫu độc lập trong trường hợp các giả thuyết về tổng thể như phải có phân phối chuẩn, phương sai bằng nhau cho cỡ mẫu nhỏ, …không thỏa mãn thì ta có thể sử dụng kiểm định tổng và hạng Wilconxon để so sánh với ý tưởng chính là sử dụng trung vị thay thế cho giá trị trung bình. Các bước thực hiện được tiến hành như sau:

Bước 1. Đặt giả thuyết

Kiểm định hai phía (hai bên)

$$\begin{cases} H_0 : d_1 = d_2 \\ H_1 : d_1 \neq d_2 \end{cases}$$

Kiểm định một phía (phía phải)

$$\begin{cases} H_0 : d_1 = d_2 \\ H_1 : d_1 > d_2 \end{cases}$$

Kiểm định một phía (phía trái)

$$\begin{cases} H_0 : d_1 = d_2 \\ H_1 : d_1 < d_2 \end{cases}$$

Trong đó

d_1 là giá trị trung vị của mẫu thứ 1

d_2 là giá trị trung vị của mẫu thứ 2

Bước 2. Dựa vào mẫu thu thập được, ta lập bảng và xếp hạng như sau

Số thứ tự	Mẫu 1	Mẫu 2	Mẫu kết hợp	Hạng mẫu kết hợp	Hạng mẫu 1	Hạng mẫu 2
1	x_1	y_1	x_1			
2	x_2	y_2	x_2			
…	…	…..	…			
n	x_n	y_n	x_n			
			y_1			
			y_2			
			….			
			y_n			
Tổng						

Trong đó

- Mẫu kết hợp được sắp xếp từ nhỏ tới lớn dựa vào giá trị thu thập được từ 2 mẫu. Giá trị nhỏ nhất có hạng bằng 1.
- Hạng của mẫu 1 và mẫu 2 sẽ được tính theo nguyên tắc từ hạng của mẫu kết hợp nhưng ứng với giá trị của nó.

Bước 3. Tính giá trị kiểm định Wilconxon bằng tổng hạng của mẫu có số phần tử ít hơn. Nếu hai mẫu có số phần tử bằng nhau thì giá trị kiểm định tính trên mẫu nào cũng được. Với mức ý nghĩa α cho trước ta tìm được giá trị tới hạn trong bảng tổng và hạng Wilconxon như sau: hai phía $W_{\alpha/2}^{n_1,n_2} = (W_L; W_M)$ và một phía $W_{\alpha}^{n_1,n_2} = (W_L; W_M)$

Bước 4. So sánh và kết luận

- Đối với hai phía: Nếu $T_0 \notin (W_L; W_M)$ thì bác bỏ giả thuyết H_0.
- Đối với phía phải: Nếu $T_0 > W_M$ thì bác bỏ giả thuyết H_0.
- Đối với phía trái: Nếu $T_0 < W_L$ thì bác bỏ giả thuyết H_0.

Lưu ý : Trong trường hợp cả hai mẫu độc lập lớn $(n_1, n_2 > 20)$. Tính giá trị kiểm định T_0 ở bước 3 thay bằng Z_0 theo phân phối chuẩn như sau:

$$\mu_{T_0} = \frac{n_1(n+1)}{4} \text{ và } \sigma_{T_0}^2 = \frac{n_1 n_2 (n+1)}{12}$$

Suy ra

$$Z_0 = \frac{T_0 - \mu_{T_0}}{\sigma_{T_0}}$$

và bước 4. Với mức ý nghĩa α cho trước ta tìm được giá trị tới hạn trong bảng phân phối chuẩn, $z_{\alpha/2}$ và một phía z_α

- Đối với hai phía: Nếu $|Z_0| > z_{\alpha/2}$ thì bác bỏ giả thuyết H_0.
- Đối với phía phải: Nếu $Z_0 > z_\alpha$ thì bác bỏ giả thuyết H_0.
- Đối với phía trái: Nếu $Z_0 < -z_\alpha$ thì bác bỏ giả thuyết H_0.

Ví dụ 4.27. Để kiểm định xem việc trưng bày hàng hóa có tác động đến doanh số bán hay không, người ta chọn ngẫu nhiên hai mẫu, mẫu thứ nhất gồm 10 cửa hàng trưng bày bình thường; mẫu thứ hai gồm 10 cửa hàng trưng bày đặc biệt. Sau đó quan sát doanh số bán của các cửa hàng này (doanh số bán, X đơn vị tính triệu đồng: trưng bày bình thường; doanh số bán, Y đơn vị tính triệu đồng: trưng bày đặc biệt) ta thu thập được bảng số liệu như sau:

X	20	33	50	60	30	40	62	80	54	61
Y	50	70	74	55	65	80	64	90	75	85

Với mức ý nghĩa 5%, hãy cho biết có hay không sự khác biệt về doanh số bán giữa hai cách trưng bày trên? Giả sử doanh số bán không có phân phối chuẩn.

Giải

Bước 1. Đặt giả thuyết

Bài toán kiểm định hai phía như sau

$$\begin{cases} H_0 : d_1 = d_2 \\ H_1 : d_1 \neq d_2 \end{cases}$$

Trong đó

d_1 là doanh số bán theo trung vị của cửa hàng trưng bày bình thường

d_2 là doanh số bán theo trung vị của cửa hàng trưng bày đặc biệt

Bước 2. Dựa vào mẫu thu thập được, ta lập bảng và xếp hạng như sau

Số thứ tự	X	Y	Mẫu kết hợp	Hạng mẫu kết hợp	Hạng mẫu 1	Hạng mẫu 2
1	20	50	20	1	1	5,5
2	33	70	30	2	3	14

3	50	74	33	3	5,5	15
4	60	55	40	4	9	8
5	30	65	50	5,5	2	13
6	40	80	50	5,5	4	17,5
7	62	64	54	7	11	12
8	80	90	55	8	17,5	20
9	54	75	60	9	7	16
10	61	85	61	10	10	19
			62	11		
			64	12		
			65	13		
			70	14		
			74	15		
			75	16		
			80	17,5		
			80	17,5		
			85	19		
			90	20		
Tổng					70	140

Bước 3. Tính giá trị kiểm định Wilconxon $T_0 = 70$

Lưu ý: Do hai mẫu có số phần tử khảo sát bằng nhau nên ta chọn tổng hạng của mẫu nào cũng được

Với mức ý nghĩa 5%, ta tìm được giá trị tới hạn trong bảng tổng và hạng Wilconxon như sau $W_{\alpha/2}^{n_1,n_2} = w_{0,025}^{10,10} = (78;132)$.

Bước 4. So sánh và kết luận

Ta có $T_0 = 70 \notin W_{0,025}^{10,10} = (78;132)$, nên bác bỏ H_0. Vậy với mức ý nghĩa 5%, có sự khác biệt về doanh số bán ứng với hai cách trưng bày.

4.6.5. Kiểm định Kruskal – Wallis

Chúng ta sẽ thực hiện bài toán kiểm định về sự bằng nhau của k trung bình tổng thể. Chọn k mẫu ngẫu nhiên độc lập có $n_1, n_2, ... n_k$ quan sát, gọi $n = \sum n_i$. Xếp hạng tất cả các quan sát theo thứ tự tăng dần, những giá trị bằng nhau sẽ nhận hạng trung bình. Gọi $R_1, R_2, ..., R_k$ là tổng hạng của từng mẫu.

Bước 1. Đặt giả thuyết

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \exists \mu_i \neq \mu_j, i \neq j \end{cases}$$

Trong đó

Giả thuyết H_0 là hàm ý không có sự khác biệt về mặt trung bình giữa các nhóm tính chất của biến nguyên nhân dựa trên giá trị thu nhận của biến kết quả.

Đối thuyết H_1 hàm ý tồn tại ít nhất hai nhóm tính chất của biến nguyên nhân có sự khác biệt về mặt trung bình.

Bước 2. Tính giá trị kiểm định

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

Bước 3. Qui tắc quyết định

+) Nếu $Q > \chi_\alpha^2 (k-1)$ thì bác bỏ H_0.

+) Nếu $Q \leq \chi_\alpha^2 (k-1)$ thì chấp nhận H_0.

Trong trường hợp giả thuyết H_0 bị bác bỏ thì tức là tồn tại ít nhất một cặp có giá trị trung bình khác nhau, ta có tiến hành kiểm định như sau:

Bước 4. Xác định số cặp cần kiểm định thông qua công thức: C_k^2

Đặt giả thuyết cần kiểm định

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases} \text{với } i, j = 1, 2, ..., k;\ i \neq j$$

Bước 5. Tính chênh lệch về hạng trung bình giữa các nhóm:

$$D_{ij} = \left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right|$$

Bước 6. Tính giá trị kiểm định

$$C_{ij} = \sqrt{\chi_\alpha^2 (k-1) \left(\frac{n(n+1)}{12} \right) \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Bước 7. So sánh và kết luận

Nếu $D_{ij} > C_{ij}, (i \neq j)$ thì ta bác bỏ H_0 và ngược lại.

Ví dụ 4.28. Một nhà nghiên cứu muốn xem xét tổng giá trị sản phẩm sản xuất của 3 ngành A, B, C có giống nhau không, người ta chọn một số xí nghiệp hoạt động trong ngành này có bảng số liệu sau:

Ngành A	1,38	1,55	1,9	2	1,22	2,11	1,98	1,61
Ngành B	2,33	2,5	2,79	3,01	1,99	2,45		
Ngành C	1,06	1,37	1,09	1,65	1,44	1,11		

Có kết luận gì với mức ý nghĩa 5%.

Giải

Gọi μ_1 là giá trị sản phẩm sản xuất trung bình của ngành A

μ_2 là giá trị sản phẩm sản xuất trung bình của ngành B

μ_3 là giá trị sản phẩm sản xuất trung bình của ngành C

Bước 1. Đặt giả thuyết

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \\ H_1 : \exists \mu_i \neq \mu_j,\ i \neq j \end{cases}$$

Bước 2. Tính giá trị kiểm định

Ngành A	1,38	1,55	1,9	2	1,22	2,11	1,98	1,61	Tổng
Ngành B	2,33	2,5	2,79	3,01	1,99	2,45			
Ngành C	1,06	1,37	1,09	1,65	1,44	1,11			
Rank(A)	4	6	8	9	11	12	14	15	79
Rank(B)	13	16	17	18	19	20			103
Rank(C)	1	2	3	5	7	10			28

Ta có

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{20 \times 21} \left(\frac{79^2}{8} + \frac{103^2}{6} + \frac{28^2}{6} \right) - 3 \times 21 = 13,542$$

Bước 3. Qui tắc bác bỏ

Với mức ý nghĩa $\alpha = 5\%$, ta có tìm được giá trị tới hạn trong bảng chi bình phương như sau, $\chi^2_{0,05}(2) = 5,991$

Từ bước 2, ta có $Q = 13,54 > \chi^2_{0,05}(2) = 5,991$. Bác bỏ H_0. Nghĩa là tổng giá trị sản phẩm sản xuất trung bình của các ngành là không giống nhau (có ít nhất hai ngành có giá trị sản phẩm sản xuất khác nhau).

Bước 4. Xác định và đặt giả thuyết như sau

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}; \begin{cases} H_0: \mu_1 = \mu_3 \\ H_1: \mu_1 \neq \mu_3 \end{cases}; \begin{cases} H_0: \mu_2 = \mu_3 \\ H_1: \mu_2 \neq \mu 3 \end{cases}$$

Bước 5. Tính chênh lệch về hạng trung bình giữa các nhóm:

Ta có

$$D_{12} = \left| \frac{R_1}{n_1} - \frac{R_2}{n_2} \right| = \left| \frac{79}{8} - \frac{103}{6} \right| = 7,292$$

$$D_{13} = \left| \frac{R_1}{n_1} - \frac{R_3}{n_3} \right| = \left| \frac{79}{8} - \frac{28}{6} \right| = 5,21$$

$$D_{23} = \left| \frac{R_2}{n_2} - \frac{R_3}{n_3} \right| = \left| \frac{103}{6} - \frac{28}{6} \right| = 12,5$$

Bước 6. Tính giá trị kiểm định

$$C_{12} = \sqrt{\chi^2_\alpha (k-1) \left(\frac{n(n+1)}{12} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{5,991 \frac{20 \times 21}{12} \left(\frac{1}{8} + \frac{1}{6} \right)} = 7,82$$

$$C_{13} = \sqrt{\chi^2_\alpha (k-1) \left(\frac{n(n+1)}{12} \right) \left(\frac{1}{n_1} + \frac{1}{n_3} \right)} = \sqrt{5,991 \frac{20 \times 21}{12} \left(\frac{1}{8} + \frac{1}{6} \right)} = 7,82$$

$$C_{23} = \sqrt{\chi^2_\alpha (k-1) \left(\frac{n(n+1)}{12} \right) \left(\frac{1}{n_2} + \frac{1}{n_3} \right)} = \sqrt{5,991 \frac{20 \times 21}{12} \left(\frac{1}{6} + \frac{1}{6} \right)} = 8,36$$

Bước 7. So sánh và kết luận

- Ta có $D_{12} < C_{12}$ nên chưa đủ cơ sở bác bỏ H_0. Kết luận: Tổng giá trị sản phẩm sản xuất ngành A và B không có sự khác biệt.
- Ta có $D_{13} < C_{13}$ nên chưa đủ cơ sở bác bỏ H_0. Kết luận: Tổng giá trị sản phẩm sản xuất ngành A và C không có sự khác biệt.
- Ta có $D_{23} > C_{23}$ nên ta bác bỏ H_0. Kết luận: Tổng giá trị sản phẩm sản xuất ngành B và C có sự khác biệt.

4.7. Tóm tắt chương 4

A. Kiểm định tham số

1. Kiểm định giả thuyết về giá trị trung bình, nếu biết σ_0^2

i) Đối xứng (hai phía)

Cặp giả thuyết thống kê : $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_o \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Z = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0} \sim N(0;1)$

Với mức ý nghĩa α, ta có $\phi_0(C) = \dfrac{1-\alpha}{2}$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ Z = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0};\ |Z| > C \right\}$$

ii) Một phía (Phía phải)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu > \mu_o \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Z = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0} \sim N(0;1)$

Với mức ý nghĩa α, ta có $\phi_0(C) = \dfrac{1-2\alpha}{2}$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ Z = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0}; Z > C \right\}$$

iii) Một phía (Phía trái)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu < \mu_o \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Z = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0} \sim N(0;1)$

Với mức ý nghĩa α, ta có $\phi_0(C) = \dfrac{1-2\alpha}{2}$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ Z = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{\sigma_0}; Z < -C \right\}$$

Nếu $Z_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.

Nếu $Z_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

2. Kiểm định giả thuyết về giá trị trung bình, nếu chưa biết σ^2

i) Đối xứng (hai phía)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu \neq \mu_o \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $T = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X} \sim St\left(n-1\right)$

Với mức ý nghĩa α, ta có $C = t_{\alpha/2}\left(n-1\right)$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X};\ \left|T\right| > C \right\}$$

ii) Một phía (Phía phải)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu > \mu_o \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $T = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X} \sim St\left(n-1\right)$

Với mức ý nghĩa α, ta có $C = t_{\alpha}\left(n-1\right)$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X};\ T > C \right\}$$

iii) Một phía (Phía trái)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \mu = \mu_0 \\ H_1:\ \mu < \mu_o \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $T = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X} \sim St\left(n-1\right)$

Với mức ý nghĩa α, ta có $C = t_{\alpha}\left(n-1\right)$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ T = \frac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S_X};\ T < -C \right\}$$

Nếu $T_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.

Nếu $T_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

3. Kiểm định giả thuyết về tỷ lệ

i) Đối xứng (hai phía)

Cặp giả thuyết thống kê : $\begin{cases} H_0: \ p = p_0 \\ H_1: \ p \neq p_0 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Z = \dfrac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0;1)$

Với mức ý nghĩa α, ta có $\phi_0(C) = \dfrac{1-\alpha}{2}$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}; \ |Z| > C \right\}$$

ii) Một phía (Phía phải)

Cặp giả thuyết thống kê : $\begin{cases} H_0: \ p = p_0 \\ H_1: \ p > p_0 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Z = \dfrac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0;1)$

Với mức ý nghĩa α, ta có $\phi_0(C) = \dfrac{1-2\alpha}{2}$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}; \ Z > C \right\}$$

iii) Một phía (Phía trái)

Cặp giả thuyết thống kê : $\begin{cases} H_0: \ p = p_0 \\ H_1: \ p < p_0 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Z = \dfrac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \sim N(0;1)$

Với mức ý nghĩa α, ta có $\phi_0(C) = \dfrac{1-2\alpha}{2}$ ta thu được miền bác bỏ:

$$W_\alpha = \left\{ Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1-p_0)}}; \ Z < -C \right\}$$

Nếu $Z_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.

Nếu $Z_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

4. Kiểm định giả thuyết về phương sai

i) Đối xứng (hai phía)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \sigma^2 = \sigma_0^2 \\ H_1:\ \sigma^2 \neq \sigma_0^2 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có : $Q = \dfrac{(n-1)S_X^2}{\sigma_0^2} \sim \chi^2(n-1)$.

Với mức ý nghĩa α, ta thu được miền bác bỏ :

$$W_\alpha = \left\{ Q = \frac{(n-1)S_X^2}{\sigma_0^2};\ \chi^2 < \chi^2_{1-\alpha/2}(n-1) \text{ hoac } \chi^2 > \chi^2_{\alpha/2}(n-1) \right\}$$

ii) Một phía (Phía phải)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \sigma^2 = \sigma_0^2 \\ H_1:\ \sigma^2 > \sigma_0^2 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Q = \dfrac{(n-1)S_X^2}{\sigma_0^2} \sim \chi^2(n-1)$.

Với mức ý nghĩa α, ta thu được miền bác bỏ:

$$W_\alpha = \left\{ Q = \frac{(n-1)S_X^2}{\sigma_0^2};\ \chi^2 > \chi^2_{\alpha}(n-1) \right\}$$

iii) Một phía (Phía trái)

Cặp giả thuyết thống kê : $\begin{cases} H_0:\ \sigma^2 = \sigma_0^2 \\ H_1:\ \sigma^2 < \sigma_0^2 \end{cases}$

Nếu giả thuyết H_0 đúng thì ta có: $Q = \dfrac{(n-1)S_X^2}{\sigma_0^2} \sim \chi^2(n-1)$.

Với mức ý nghĩa α, ta thu được miền bác bỏ :

$$W_\alpha = \left\{ Q = \frac{(n-1)S_X^2}{\sigma_0^2};\ \chi^2 < \chi^2_{1-\alpha}(n-1) \right\}$$

- Nếu $Q_{qs} \in W_\alpha$ thì bác bỏ H_0, thừa nhận H_1.
- Nếu $Q_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0.

5. So sánh hai trung bình μ_X và μ_Y của hai tổng thể nếu chưa biết $\sigma_X^2;\ \sigma_Y^2$.

6. So sánh hai tỷ lệ p_X và p_Y của hai tổng thể.

7. So sánh hai phương sai σ_X^2 và σ_Y^2 của hai tổng thể.

B. Kiểm định phi tham số

1. Kiểm định về tính độc lập.

2. Kiểm định về tính phù hợp (hay về luật phân phối).

3. Kiểm định dấu và hạng Wilconxon.

4. Kiểm định tổng và hạng Wilconxon.

5. Kiểm định Kruskal – Wallis.

4.8. Bài tập

Bài số 1. Giám đốc một xí nghiệp cho biết lương trung bình của 1 công nhân thuộc xí nghiệp là 7,6 triệu đồng/tháng. Chọn ngẫu nhiên 36 công nhân thấy lương trung bình là 7 triệu đồng/tháng, với độ lệch chuẩn $\sigma = 8$. Lời báo cáo của giám đốc có tin cậy được không, với mức có ý nghĩa là 5%.

Đáp số: $Z = -0,45$, *bác bỏ.*

Bài số 2. Khối lượng các bao gạo là biến ngẫu nhiên có phân phối chuẩn $N(50;\, 0,01)$. Có nhiều ý kiến khách hàng phản ánh là khối lượng bị thiếu. Một nhóm thanh tra đã cân ngẫu nhiên 25 bao gạo trong kho, kết quả như sau :

Khối lượng bao gạo (kg)	48-48,5	48,5-49	49-49,5	49,5-50	50-50,5
Số bao	2	5	10	6	2

Hãy xem ý kiến khách hàng có đúng không? Với mức ý nghĩa 5%.

Đáp số: $Z = -36,5$, *bác bỏ.*

Bài số 3. Trong điều kiện chăn nuôi bình thường, lượng sữa trung bình của 1 con bò là 14kg/ngày. Nghi ngờ điều kiện chăn nuôi kém đi làm cho lượng sữa giảm xuống, người ta điều tra ngẫu nhiên 25 con và tính được lượng sữa trung bình của 1 con trong 1 ngày là 12,5 và độ lệch tiêu chuẩn 2,5. Với mức ý nghĩa 5%. Hãy kết luận điều nghi ngờ nói trên. Giả thiết lượng sữa bò là 1 biến ngẫu nhiên có phân phối chuẩn.

Đáp số: $T = -3$, *bác bỏ.*

Bài số 4. Một cửa hàng thực phẩm nhận thấy thời gian vừa qua trung bình một khách hàng mua 25 ngàn đồng thực phẩm trong ngày. Nay cửa hàng chọn ngẫu nhiên 15 khách hàng thấy trung bình một khách hàng mua 24 ngàn đồng trong ngày và độ lệch chuẩn mẫu hiệu chỉnh là 2 ngàn đồng.

Với mức ý nghĩa là 5%, thử xem có phải sức mua của khách hàng hiện nay có thực sự giảm sút hay không.

Đáp số: $T = -1,9365$, *bác bỏ.*

Bài số 5. Điều tra một mẫu gồm 100 gia đình ở vùng nông thôn người ta thu được kết quả về chi tiêu trung bình hàng tháng của các gia đình đó là 3,455 triệu đồng với độ lệch chuẩn là 0,3 triệu đồng. Với mức ý nghĩa 5% có thể cho rằng chi tiêu trung bình hàng tháng của các gia đình ít hơn 3,5 triệu hay không. Giả thiết mức chi tiêu có phân phối chuẩn.

Đáp số: $T = 1,5$*, chấp nhận giả thuyết.*

Bài số 6. Khối lượng trung bình khi xuất chuồng ở một trại chăn nuôi trước là 3,3 kg/con. Năm nay người ta sử dụng một loại thức ăn mới, cân thử 15 con khi xuất chuồng ta được các số liệu như sau:

3,25; 2,50; 4,00; 3,75; 3,80; 3,90; 4,02; 3,60; 3,80; 3,20; 3,82; 3,40; 3,75; 4,00; 3,50

Giả thiết khối lượng gà là đại lượng ngẫu nhiên có phân phối theo quy luật chuẩn.

a) Với mức ý nghĩa 5%. Hãy cho kết luận về tác dụng của loại thức ăn này ?

b) Nếu trại chăn nuôi báo cáo khối lượng trung bình khi xuất chuồng là 3,5 kg/con thì có chấp nhận được không, với mức ý nghĩa 5%.

Đáp số: a) $T = 3,0534$*; b)* $T = 1,141$.

Bài số 7. Một máy sản xuất tự động với tỷ lệ chính phẩm 98%. Sau một thời gian hoạt động, người ta nghi ngờ tỷ lệ trên đã bị giảm. Kiểm tra ngẫu nhiên 500 sản phẩm thấy có 28 phế phẩm, với mức ý nghĩa 5%, hãy kiểm tra xem chất lượng làm việc của máy có còn được như trước hay không?

Đáp số: $Z = -5,75$*, bác bỏ.*

Bài số 8. Theo một nguồn tin thì tỉ lệ hộ dân thích xem dân ca trên Tivi là 80%. Thăm dò 36 hộ dân thấy có 25 hộ thích xem dân ca. Với mức có ý nghĩa là 5%. Kiểm định xem nguồn tin này có đáng tin cậy không?

Đáp số: $Z = -1,583$*, chấp nhận.*

Bài số 9. Tỷ lệ phế phẩm của một nhà máy trước đây là 5%. Năm nay nhà máy áp dụng một biện pháp kỹ thuật mới. Để nghiên cứu tác dụng của biện pháp kỹ thuật mới, người ta lấy một mẫu gồm 800 sản phẩm để kiểm tra và thấy có 24 phế phẩm.

a) Với mức ý nghĩa 1%. Hãy cho kết luận về biện pháp kỹ thuật mới này ?

b) Nếu nhà máy báo cáo tỷ lệ phế phẩm sau khi áp dụng biện pháp kỹ thuật mới là 2% thì có chấp nhận được không, với mức ý nghĩa 5%.

Đáp số: a) $Z = -2,596$*, bác bỏ; b)* $Z = 2,02$*, bác bỏ.*

Bài số 10. Nếu máy đóng bao hoạt động bình thường thì khối lượng của một loại sản phẩm sản xuất ra là đại lượng ngẫu nhiên có phân phối theo quy luật chuẩn $N(60;0,04)$. Kiểm tra khối lượng của một số sản phẩm do máy sản xuất, ta được kết quả :

60; 60,2; 70; 60,8; 50,6 ;50,8; 50,9; 60,1; 50,3; 60,5; 60,1; 60,2; 60,3; 50,8; 60; 70

a) Với mức ý nghĩa 5%, hãy cho biết máy đóng bao hoạt động có bình thường hay không?

b) Ước lượng khối lượng trung bình của loại sản phẩm này hiện nay với độ tin cậy 95%.

Đáp số: a) $Z=-30,6$, *bác bỏ; b)* $\mu\in[58,377;58,573]$.

Bài số 11. Trồng cùng một giống lúa trên hai thửa ruộng như nhau và bón hai loại phân khác nhau. Đến ngày thu hoạch ta có kết quả như sau : Thửa thứ nhất lấy mẫu 1000 bông lúa thấy số hạt trung bình của mỗi bông $\overline{X}=70$ hạt và $S_X=10$. Thửa thứ hai lấy mẫu 500 bông thấy số hạt trung bình mỗi bông $\overline{Y}=72$ hạt và $S_Y=20$. Hỏi số hạt trung bình mỗi bông lúa của hai thủa ruộng có như nhau hay không, với mức ý nghĩa 5%?

Đáp số: $T=-2,11$, *bác bỏ.*

Bài số 12. Để so sánh khối lượng trung bình của trẻ sơ sinh ở thành thị và nông thôn, người ta thử cân khối lượng của 10000 cháu và thu được kết quả sau đây :

Vùng	Số cháu được cân	Khối lượng trung bình	Độ lệch chuẩn mẫu
Nông thôn	8000	3,0kg	0,3kg
Thành thị	2000	3,2kg	0,2kg

Với mức ý nghĩa 5%, có thể coi khối lượng trung bình của trẻ sơ sinh ở thành thị và ở nông thôn là như nhau được hay không? (Giả thiết khối lượng trẻ sơ sinh là biến ngẫu nhiên có phân phối chuẩn).

Đáp số: $T=-35,78$, *bác bỏ giả thuyết.*

Bài số 13. Trong thập niên 80, khối lượng trung bình của thanh niên là 50kg. Nay để xác định lại khối lượng ấy, người ta chọn ngẫu nhiên 100 thanh niên đo khối lượng trung bình là 52kg và phương sai mẫu hiệu chỉnh $S^2=(10kg)^2$. Thử xem khối lượng thanh niên hiện nay phải chăng có thay đổi, với mức có ý nghĩa là 1%.

Đáp số: $T=2$, *chấp nhận giả thiết*

Bài số 14. Giả sử ta muốn xác định xem hiệu quả của chế độ ăn kiêng đối với việc giảm khối lượng như thế nào. 20 người quá béo đã thực hiện chế độ ăn kiêng. Khối lượng của từng người trước khi ăn kiêng (X kg) và sau khi ăn kiêng (Y kg) được cho như sau:

X	80	78	85	70	90	78	92	88	75	
Y	75	77	80	70	84	74	85	82	80	
X	75	63	72	89	76	77	71	83	78	82	90
Y	65	62	71	83	72	82	71	79	76	83	81

Dùng tiêu chuẩn phi tham số kiểm tra xem chế độ ăn kiêng có tác dụng làm giảm khối lượng hay không, với mức ý nghĩa 5%.

Đáp số: $T = -3,3$, *bác bỏ giả thuyết.*

Bài số 15. Dùng 3 phương án xử lý hạt giống kết quả cho như sau :

Kết quả	Phương án I	Phương án II	Phương án III
Số hạt mọc	360	603	490
Số hạt không mọc	40	97	180

Các phương án xử lý có tác dụng như nhau đối với tỷ lệ nảy mầm hay không, với mức ý nghĩa 5%.

Đáp số: $Q = 1831,394$, *bác bỏ giả thuyết.*

Bài số 16. Theo dõi sự phụ thuộc giữa màu mắt và màu tóc ở 124 phụ nữ ở một nước Châu Âu ta có kết quả sau :

Màu tóc / Màu mắt	Vàng nâu	Nâu	Đen	Vàng hoe
Xanh	25	9	3	7
Xám	13	17	10	7
Nâu mực	7	13	8	5

Với mức ý nghĩa 5%, hãy kiểm tra giả thiết cho rằng màu của tóc và màu của mắt độc lập với nhau.

Đáp số: $Q = 15,07$, *bácbỏ giả thuyết.*

Bài số 17. Để xác định thời vụ phun thuốc diệt sâu có lợi nhất, tổ bảo vệ cây trồng đã theo dõi các lứa sâu trong từng thời kỳ và đếm số sâu non mới nở bắt được. Kết quả ghi ở bảng sau

Thời kỳ theo dõi	Tháng 1	Tháng 2	Tháng 3	Tháng 4	Tháng 5
Số sâu non mới	62	28	70	75	15

nở bắt được					
Tổng số sâu non bắt được	488	392	280	515	185

Tỷ lệ sâu non mới nở trong các thời kỳ quan sát khác nhau có ý nghĩa hay không, với mức ý nghĩa 5%.

Đáp số: $Q = 37{,}28$*, bácbỏ giả thuyết.*

Bài số 18. Một nhà máy có 3 phân xưởng cùng sản xuất một loại sản phẩm. Chất lượng sản phẩm được chia thành 3 loại. Kiểm tra, phân loại ngẫu nhiên một số sản phẩm từ lô sản phẩm của 3 phân xưởng ta có số liệu sau :

Phân xưởng / Chất lượng	PX I	PX II	PX III
Loại I	70	80	60
Loại II	25	20	15
Loại III	5	10	5

Với mức ý nghĩa 5%, có thể kết luận chất lượng sản phẩm phụ thuộc vào nơi làm ra chúng hay không?

Đáp số: $Q = 2{,}872$*, chấp nhận giả thuyết.*

Bài số 19. Sản phẩm được sản xuất ra trên một dây chuyền tự động được đóng gói một cách ngẫu nhiên theo qui cách: 3 sản phẩm/hộp. Tiến hành kiểm tra 200 hộp ta được kết quả

Số sản phẩm loại I có trong hộp	0	1	2	3
Số hộp	6	14	110	70

Với mức ý nghĩa 5%, có thể xem số sản phẩm loại I có trong hộp là đại lượng ngẫu nhiên có quy luật phân phối nhị thức không?

Đáp số: $Q = 18{,}88$*, bác bỏ giả thuyết.*

Bài số 20. Một nhà máy sản xuất máy in nói rằng số lỗi in trong 1 cuốn sách dày 300 trang của máy in là 1 đại lượng ngẫu nhiên có quy luật phân phối Poisson với tham số $\mu = 4{,}7$. Kiểm tra 300 trang sách in của 50 máy in cùng loại, ta được

Số lỗi	0	1	2	3	4	5	6	7	8	≥ 9
Số máy	1	1	8	6	13	10	5	5	1	0

Với mức ý nghĩa 1%, hỏi lời tuyên bố của nhà sản xuất có đúng không?

Đáp số: $Q = 9{,}2785$*, chấp nhận giả thuyết.*

Bài số 21. Số con của 2000 phụ nữ thủ đô dưới 25 tuổi cho ở bảng sau :

Số con X	0	1	2	3	4
Số phụ nữ	1090	650	220	30	10

Với mức ý nghĩa 5% có thể xem X tuân theo luật Poisson hay không?

Đáp số: $Q = 0,6247$*, chấp nhận giả thuyết.*

Bài số 22. Kiểm tra 200 thùng một loại đồ hộp, người ta thu được số liệu sau

Số hộp bị hỏng/thùng	0	1	2	3	4
Số thùng	116	56	22	4	2

Với mức ý nghĩa 5%, chứng tỏ rằng số hộp bị hỏng của một thùng là biến ngẫu nhiên tuân theo qui luật Poisson?

Đáp số: $Q = 2,685$*, chấp nhận giả thuyết.*

Bài số 23. Số tai nạn giao thông xảy ra mỗi ngày ở 1 thành phố quan sát được

Số tai nạn	0	1	2	3	4	5	6	7	8
Số ngày	10	32	46	35	20	9	2	1	1

Với mức ý nghĩa 1%, xét xem số tai nạn giao thông có quy luật Poisson?

Đáp số: $Q = 2,3582$*, chấp nhận giả thuyết.*

Bài số 24. Năng suất lúa (X) thử nghiệm trên 100 lô đất cho kết quả

Năng suất (tấn/ha)	8-9	9-10	10-11	11-12	12-13	13-14	14-15
Số trường hợp	8	15	21	23	16	9	8

Với mức ý nghĩa 10%, xét xem X có phân phối chuẩn không?

Đáp số: $Q = 2,2784$*, chấp nhận giả thuyết.*

Bài số 25. Để nghiên cứu về mức độ yêu thích của khách hàng đối với hai mẫu quảng cáo A và B. Tiến hành nghiên cứu trên 1 nhóm khách hàng được cho xem 2 mẫu quảng cáo và đánh giá 2 mẫu trên thang điểm 10. Kết quả đánh giá như sau:

KH	1	2	3	4	5	6	7	8	9	10	11	12	13
Mẫu A	6	6	5	7	8	7	5	10	6	8	8	9	8
Mẫu B	7	9	6	6	10	9	8	9	8	9	10	8	8

Kiểm định xem có sự khác biệt mức độ yêu thích 2 mẫu quảng cáo hay không, với mức ý nghĩa 5%.

Đáp số $T = 10,5.$

Bài số 26. Để nghiên cứu về mức độ yêu thích các môn học định lượng giữa 2 nhóm nam và nữ. Tiến hành nghiên cứu trên 1 nhóm nam và 1 nhóm nữ cho biết mức độ yêu thích của họ đối với các môn học định lượng trên thang điểm 10 (1 là hoàn toàn không thích, 10 là hoàn toàn thích). Kết quả nghiên cứu như sau:

Nam	6	6	10	7	8	7	9	10	6	8	8	9		
Nữ	7	9	5	6	5	9	7	9	8	4	8	8	7	10

Kiểm định xem có sự khác biệt mức độ yêu thích các môn học định lượng giữa 2 nhóm giới tính hay không, với mức ý nghĩa 5%.

Đáp số $W = 0,4471$. *chấp nhận.*

Bài số 27. Để kiểm tra xem có mối liên hệ giữa giới tính và sự yêu thích các thể loại phim. Khảo sát 500 người dân và kết quả được tổng hợp trong bảng chéo như sau:

		Giới tính		
		Nam	Nữ	Tổng
Thể loại phim yêu thích nhất	Phim hoạt hình	20	40	60
	Phim hành động	110	60	170
	Phim tình cảm	50	120	170
	Phim khoa học viễn tưởng	80	20	100
Tổng		260	240	500

Hãy thực hiện kiểm định mối liên hệ giữa giới tính và sự yêu thích các thể loại phim với mức ý nghĩa 5%.

Đáp số $Q = 85,533$ *bác bỏ* H_0.

Bài số 28. Để nghiên cứu về sự khác biệt trong chi tiêu cho việc mua sắm quần áo cho bản thân giữa nam và nữ, người ta chọn khảo sát 1 nhóm phụ nữ và 1 nhóm nam giới về số tiền chi tiêu cho việc mua sắm quần áo trong 1 năm cho bản thân và kết quả khảo sát như sau: (đơn vị triệu đồng):

Nam	3	1,5	1	4	5	1,8	5	4	4,5	4	2,5
Nữ	1,5	4,5	5	2,4	2	3,5	2	3	1,5	5	2,5

Kiểm định xem có sự khác biệt trong chi tiêu cho việc mua sắm quần áo cho bản thân giữa nam và nữ, với mức ý nghĩa 5%.

Đáp số: $W = 12,256$, *bác bỏ.*

Bài số 29. Công ty vừa tổ chức buổi tập huấn cho nhân viên phòng sales, đây là kết quả đánh giá 10 nhân viên của khách hàng trước và sau khi tập huấn:

Nhân viên	1	2	3	4	5	6	7	8	9	10
Trước	85	60	80	70	95	75	70	80	90	80
Sau	80	70	80	100	80	80	95	95	90	85

Kiểm định xem có sự khác kết quả đánh giá 10 nhân viên của khách hàng trước và sau khi tập huấn, với mức ý nghĩa 10%.

Đáp số: $T = 1,655$, *chấp nhận.*

Bài số 30. Để kiểm tra xem có mối liên hệ giữa có đi làm thêm và kết quả học tập của sinh viên. Khảo sát 500 sinh viên và kết quả được tổng hợp trong bảng chéo như sau:

		Đi làm thêm		
		Có đi làm thêm	Không đi làm thêm	Tổng
Xếp loại kết quả học tập	Trung bình	20	30	50
	Khá	110	150	260
	Giỏi	60	100	160
	Xuất sắc	10	20	30
Tổng		200	300	500

Hãy thực hiện kiểm định mối liên hệ giữa việc đi làm thêm và kết quả học tập của sinh viên với mức ý nghĩa 5%.

Đáp số $Q = 1,55$ *chấp nhận* H_0.

Bài số 31. Để nghiên cứu về sự khác biệt trong chi tiêu giữa nhóm có và không có người yêu, người ta chọn khảo sát 1 nhóm có người yêu và không có người yêu về số tiền chi tiêu hàng tháng và kết quả khảo sát như sau: (đơn vị triệu đồng):

Có NY	3	1,5	2,4	3	2	1,8	2,4	4	4,5	4	2,5
Không NY	3,4	4,5	5	2,4	2	3,5	2	3	1,5	5	2,5

Kiểm định xem có sự khác biệt trong chi tiêu giữa nhóm có và không có người yêu, với mức ý nghĩa 5%.

Đáp số: $W = 0,4312$, *chấp nhận.*

Bài số 32. Để kiểm tra xem có mối liên hệ giữa trình độ học vấn và chủ đề quan tâm khi đọc báo của người dân. Khảo sát 500 người dân và kết quả được tổng hợp trong bảng chéo như sau:

		Trình độ học vấn		
		Tốt nghiệp 12	Cao đẳng/ĐH	Sau đại học
Các chủ đề	Chính trị	10	15	20
	Xã hội	50	150	80
	Giải trí	110	35	30

Hãy thực hiện kiểm định mối liên hệ giữa giữa trình độ học vấn và chủ đề quan tâm khi đọc báo của người dân 5%.

Đáp số: $Q = 108{,}931$, *bác bỏ.*

4.9. Tài liệu tham khảo

[1] Nguyễn Cao Văn, Trần Thái Ninh, Ngô Văn Thứ, Lý thuyết xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[2] Nguyễn Cao Văn, Trần Thái Ninh, Nguyễn Thế Hệ, Bài tập xác suất và thống kê toán, NXB Đại học Kinh tế Quốc dân, 2012.

[3] Phạm Văn Chững, Lê Thanh Hoa, Nguyễn Đình Uông, Thống kê ứng dụng, NXB Đại học Quốc gia Thành Phố Hồ Chí Minh, 2016.

[4] Hà văn Sơn, Giáo trình Lý thuyết Thống kê, ứng dụng trong Quản trị và kinh tế.

[5] Anderson, Sweeney, and William [2010], Statistics for Business and Economics, South-Western Cengage Learning (11th Edition).

[6] Michael Barrow, Statistics for Economics, Accounting and Business Studies-Prentice Hall, 2006.

[7] Newbold Paul - Statistics for Bussiness and Economics, 5th edition - Prentice Hall, 2005.

Thuật ngữ chính chương 4

Tiếng Anh	Tiếng Việt
Alternative Hypothese	Giả thuyết thay thế
Analysis tool	Công cụ phân tích
Critical value	Giá trị quan trọng
Data analysis	Phân tích dữ liệu
Key formulas	Công thức quan trọng
Kruskal-Wallis test	Kiểm tra Kruskal-Wallis
First sample	Mẫu thứ nhất
Hypothese	Giả thuyết
Research hypothesis	Giả thuyết nghiên cứu
Mann-Whitney-Wilcoxon test	Kiểm định Mann-Whitney-Wilcoxon
Methods	Các phương pháp
Null hypothesis	Giả thuyết không
Normal approximation	Xấp xỉ phân phối chuẩn
Nonparametric method	Phương pháp phi tham số
Level of significance	Mức ý nghĩa
One-tailed test	Kiểm định 1 phía
P-value approach	Cách tiếp cận giá trị xác suất
Population median	Trung vị tổng thể
Parametric method	Phương pháp tham số
Type I and type II errors	Sai lầm loại 1 và sai lầm loại 2
Two-tailed test	Kiểm định 2 phía
Test median	Kiểm định trung vị
Total number of observations	Tổng số các quan sát
Sign test	Kiểm định dấu
Second sample	Mẫu thứ hai
Steps of hypothesis testing	Các bước tiến hành kiểm định
Spearman rank correlation	Tương quan hạng Spearman
Sum of the ranks	Tổng số các hạng
Wilcoxon signed-rank test	Kiểm định dấu hạng Wilcoxon

Chương 5 **PHÂN TÍCH PHƯƠNG SAI**

Mục tiêu chương 5

Chương này giúp sinh viên:

- Nắm được thế nào là phân tích phương sai và khi nào sử dụng phân tích phương sai.
- Hiểu và áp dụng được phân tích phương sai một yếu tố.
- Hiểu và áp dụng được phân tích phương sai hai yếu tố.

Trong chương 4 chúng ta đã kiểm định sự khác nhau về trung bình hai tổng thể có thể dựa vào hai mẫu khảo sát ngẫu nhiên. Tuy nhiên, trên thực tế thì đôi khi chúng ta cần so sánh sự bằng nhau về giá trị trung bình nhiều hơn hai tổng thể. Phân tích phương sai có thể giải quyết vấn đề này.

5.1. Phân tích phương sai một yếu tố

Giả sử biến ngẫu nhiên X chịu tác động của nhân tố A gồm k mức nhân tố $X_1, X_2, ..., X_k$ với X_j có phân phối chuẩn, $N\left(\mu_j, \sigma_j^2\right)$ có mẫu điều tra như sau:

	X_1	X_2	$\cdots$	X_k
1	X_{11}	X_{12}	$\cdots$	X_{1k}
2	X_{21}	X_{22}	$\cdots$	X_{2k}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n_i	$X_{n_{i1}}$	$X_{n_{i2}}$	$\cdots$	$X_{n_{ik}}$
Giá trị trung bình	$\bar{X}_1$	$\bar{X}_2$	$\cdots$	$\bar{X}_k$

Trong đó:

- k là số mẫu khảo sát;
- n_j là cỡ mẫu của nhóm thứ j;
- X_{ij} là giá trị của phần tử thứ i trong mẫu thứ j;
- $\bar{X}_j$ là trung bình mẫu của mẫu thứ j;
- $\bar{X}$ là trung bình mẫu chung của toàn bộ các phần tử trong k mẫu.

Ta xây dựng các bước để kiểm định cặp giả thiết như sau

Bước 1. Đặt giả thuyết

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : H_0 \text{ sai} \end{cases}$$

Trong đó μ_i là giá trị trung bình tổng thể thứ $i\ (i = 1, 2, ..., k)$

Bước 2. Tính trung bình mẫu

- Tổng số quan sát :

$$n = \sum_{j=1}^{k} n_j$$

- Trung bình mẫu nhóm $j\ (j = 1, 2, ..., k)$ (trung bình cột) theo công thức:

$$\overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij} = \frac{T_j}{n_j} \quad \text{với } T_j = \sum_{i=1}^{n_j} X_{ij}$$

- Trung bình mẫu chung của k nhóm theo công thức

$$\overline{X} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}}{\sum_{j=1}^{k} n_j} = \frac{T}{n} \quad \text{với } T = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} = \sum_{j=1}^{k} T_j$$

Bước 3. Tính tổng bình phương độ lệch

- Phương sai có hiệu chỉnh nhóm $j\ (j = 1, 2, ..., k)$ (phương sai có hiệu chỉnh cột)

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(X_{ij} - \overline{X}_j\right)^2 = \frac{T_j}{n_j}$$

- Tổng bình phương các độ lệch của toàn bộ các phần tử được tính theo công thức

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(X_{ij} - \overline{X}\right)^2$$

- Tổng bình phương các độ lệch riêng của các nhóm so với $\overline{X}$ được tính theo công thức

$$SSA = \sum_{j=1}^{k} n_j \left(\overline{X}_j - \overline{X}\right)^2$$

- Tổng bình phương các độ lệch trong nội bộ của nhóm được tính theo công thức

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(X_{ij} - \overline{X}_j\right)^2$$

- Mối quan hệ giữa các bình phương độ lệch, từ SST ta có

$$\mathrm{SST}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij}-\overline{X}_j+\overline{X}_j-\overline{X}\right)^2$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij}-\overline{X}_j\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{X}_j-\overline{X}\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}2\left(X_{ij}-\overline{X}_j\right)\left(\overline{X}_j-\overline{X}\right)$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij}-\overline{X}_j\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{X}_j-\overline{X}\right)^2=\mathrm{SSE}+\mathrm{SSA}$$

Vì

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}2\left(X_{ij}-\overline{X}_j\right)\left(\overline{X}_j-\overline{X}\right)=2\sum_{j=1}^{k}\left(\overline{X}_j-\overline{X}\right)\sum_{i=1}^{n_j}\left(X_{ij}-\overline{X}_j\right)$$

$$=2\sum_{j=1}^{k}\left(\overline{X}_j-\overline{X}\right)\left[\sum_{i=1}^{n_j}X_{ij}-\sum_{i=1}^{n_j}\overline{X}_j\right]$$

$$=2\sum_{j=1}^{k}\left(\overline{X}_j-\overline{X}\right)\left[n_j\overline{X}_j-n_j\overline{X}_j\right]=0$$

Chú ý: SST thể hiện sự biến thiên của hiện tượng nghiên cứu; SSA thể hiện sự biến thiên do yếu tố cột tạo ra; SSE thể hiện sự biến thiên do các yếu tố khác. Như vậy hiện tượng nghiên cứu phụ thuộc vào hai phần: do yếu tố đang xem xét và những yếu tố tác động. Nếu như sự tác động của yếu tố đang nghiên cứu cũng như các yếu tố khác tác động thì ta có thể kết luận hiện tượng nghiên cứu không phụ thuộc vào yếu tố đang xem xét. Điều này dẫn đến trung bình theo nhóm bằng nhau.

Bước 4. Tính phương sai

- Phương sai trong nội bộ nhóm theo công thức

$$\mathrm{MSE}=\frac{\mathrm{SSE}}{n-k}$$

- Phương sai giữa các nhóm theo công thức

$$\mathrm{MSA}=\frac{\mathrm{SSA}}{k-1}$$

Bước 5. Tính giá trị kiểm định

$$F=\frac{\mathrm{MSA}}{\mathrm{MSE}}$$

Với mức ý nghĩa α cho trước, ta tìm được giá trị tới hạn trong bảng Fisher

$$C=f_\alpha\left(k-1,n-k\right)$$

+) Nếu $F>f_\alpha\left(k-1,n-k\right)$ thì bác bỏ H_0.

+) Nếu $F \le f_\alpha(k-1, n-k)$ thì chưa đủ cơ sở bác bỏ H_0.

Ta có thể tóm tắt các bước kiểm định trên trong bảng sau:

Phân tích phương sai 1 yếu tố				
	Tổng các chênh lệch bình phương	**Bậc tự do**	**Phương sai**	**Tỷ số (F)**
Giữa các nhóm	SSA	$k-1$	MSA	$\frac{MSA}{MSE}$
Nội bộ nhóm	SSE	$n-k$	MSE	
Tổng	SST	$n-1$		

Ví dụ 5.1. Một nghiên cứu được thực hiện nhằm xem xét năng suất lúa trung bình của 3 giống lúa có bằng nhau hay không? Kết quả thu thập qua 4 năm như sau:

Năm	Giống A	Giống B	Giống C
1	65	69	75
2	74	72	70
3	64	68	78
4	83	78	76

Hãy kết luận với mức ý nghĩa 5%.

Giải

Bước 1. Đặt giả thuyết

Gọi μ_1, μ_2, μ_3 là năng suất lúa trung bình tương ứng của các giống lúa A, B, C. Ta có

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: H_0 \text{ sai} \end{cases}$$

Bước 2. Tính trung bình mẫu

- Trung bình mẫu của từng giống lúa:

Giống lúa A: $\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{i1} = 71,5$

Giống lúa B: $\overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{i2} = 71,75$

Giống lúa C: $\overline{X}_3 = \frac{1}{n_3}\sum_{i=1}^{n_3} X_{i3} = 74,75$

- Trung bình mẫu chung của 3 giống lúa theo công thức

$$\overline{X} = \frac{\sum_{j=1}^{k} n_j \overline{X}_j}{\sum_{j=1}^{k} n_j} = \frac{218}{3} \approx 72,667$$

Bước 3. Tính tổng bình phương độ lệch

- Tổng bình phương độ lệch giữa các giống lúa

$$SSA = \sum_{j=1}^{k} n_j \left(\overline{X}_j - \overline{X}\right)^2 = \frac{157}{6} \approx 26,167$$

- Tổng bình phương độ lệch trong nội bộ các giống lúa

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_{ij} - \overline{X}_j\right)^2 = 237 + 60,75 + 34,75 = 332,5$$

- Tổng bình phương độ lệch toàn bộ các các giống lúa

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_{ij} - \overline{X}\right)^2 = \frac{1076}{3} \approx 358,667$$

Bước 4. Tính phương sai

- Phương sai trong nội bộ giống lúa

$$MSE = \frac{SSE}{n-k} = \frac{332,5}{12-3} = \frac{665}{18} \approx 36,9444$$

- Phương sai giữa các giống lúa

$$MSA = \frac{SSA}{k-1} = \frac{26,167}{3-1} = \frac{157}{12} \approx 13,0833$$

Bước 5. Tính giá trị kiểm định

$$F = \frac{MSA}{MSE} = \frac{13,0833}{36,9444} = 0,3541$$

Với mức ý nghĩa $\alpha = 0,05$ tra bảng Fisher ta có giá trị tới hạn $C = f_{0,05}\left(2,9\right) = 4,26$

Do $F < C$ chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, năng suất trung bình của ba giống lúa là như nhau.

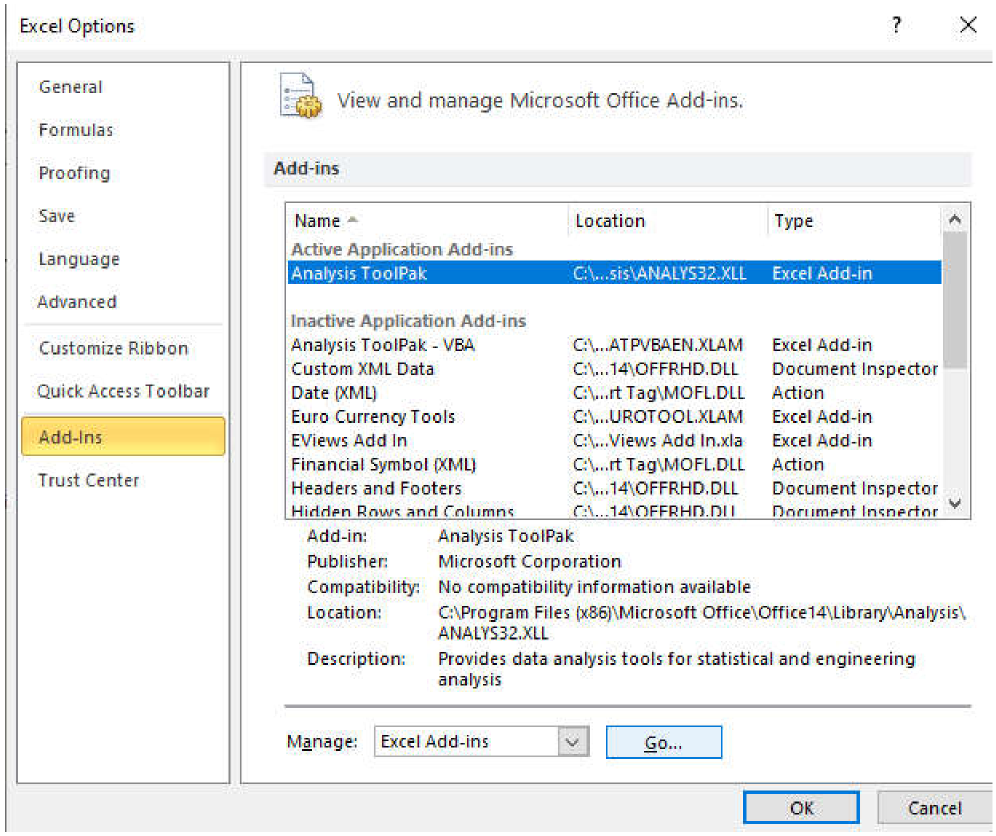
Dùng Excel phân tích phương sai

Kích hoạt Data Analysis trong Excel

Data Analysis đã từng được tích hợp trong Excel 2003. Tuy nhiên, các phiên bản Excel mới không tích hợp công cụ này mà ẩn trong phần Add-in.

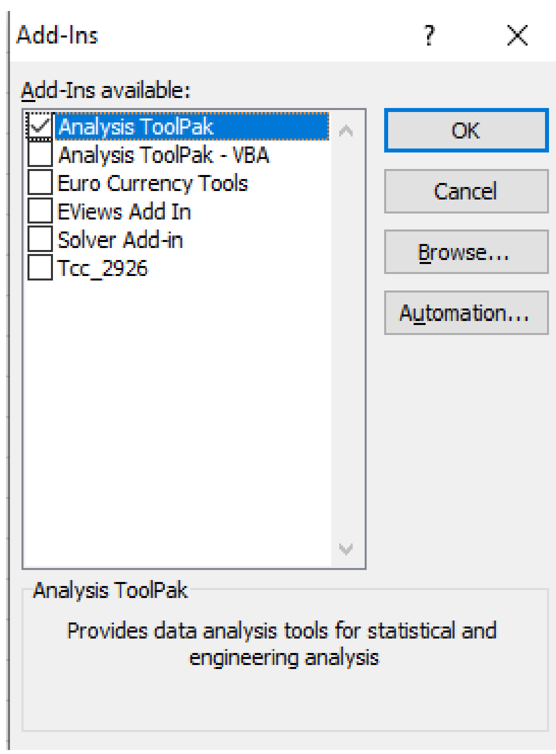
Bước 1: Chọn **"File"** => chọn **"Options"**

Bước 2: Vào mục **"Add-Ins"** => chọn **"Analysis ToolPak"** sau đó bấm vào**"Go"**



Bước 3: Sau khi bấm **"Go"** thì sẽ hiện ra 1 giao diện cửa sổ **"Add-Ins".**

Ban chọn **"Analysis ToolPak"** rồi bấm **"Ok"**

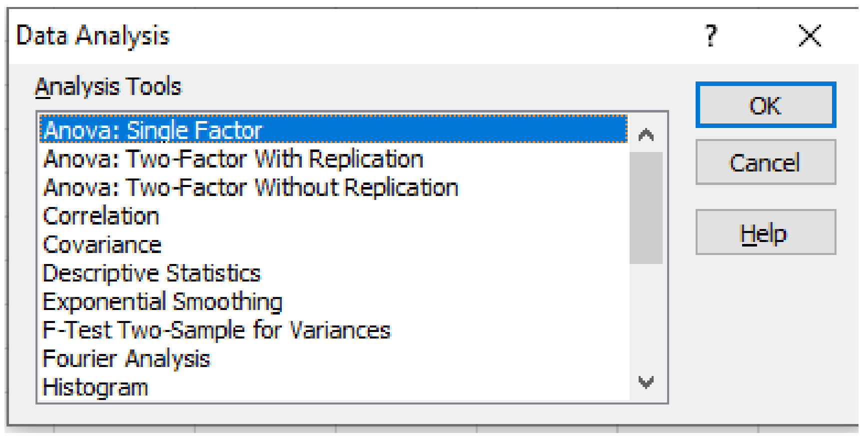


Với ví dụ trên, ta thực hiện như sau

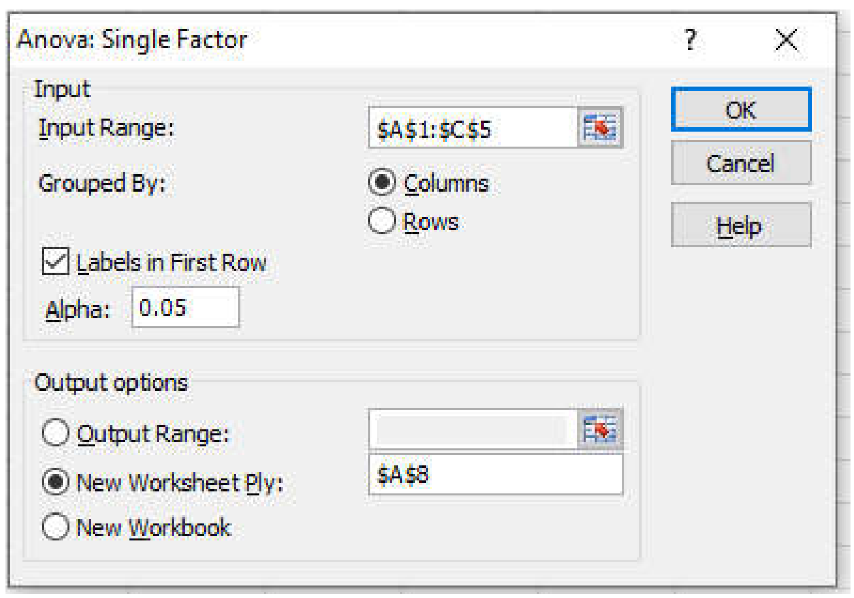
Bước 1. Nhập dữ liệu theo cột

	A	B	C
1	Giống A	Giống B	Giống C
2	65	69	75
3	74	72	70
4	64	68	78
5	83	78	76

Bước 2. Từ màn hình Excel chọn Data analysis sau đó chọn : **Anova: Single Factor**



Nhấn **OK** màn hình xuất hiện như sau



Trong đó

+) Input Range: chọn vùng dữ liệu (quét bảng dữ liệu)

+) Columns: theo cột

+) Labels in First Row: Tên các giống lúa A, B, C

+) Alpha: Mức ý nghĩa 5%

+) New Worsheet ply: Chọn vùng xuất kết quả là A8

Bước 3. Nhấn **Ok**, ta được kết quả

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Giống A	4	286	71.5	79		
Giống B	4	287	71.75	20.25		
Giống C	4	299	74.75	11.58333		
ANOVA						
ce of Varic	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between (	26.16667	2	13.08333	0.354135	0.711136	4.256495
Within Gr	332.5	9	36.94444			
Total	358.6667	11				

5.2. Phân tích phương sai hai yếu tố

Phân tích phương sai hai chiều là xét đến hai yếu tố ảnh hưởng đến hiện tượng nghiên cứu

5.2.1. Phân tích phương sai hai yếu tố không lặp

Trong trường hợp này tương ứng với sự tác động của yếu tố cột và yếu tố hàng chúng ta chỉ chọn một quan sát. Đây là trường hợp mở rộng của phân tích phương sai một yếu tố, có nghĩa là ta vừa kiểm định giả thuyết trung bình theo cột bằng nhau vừa kiểm định trung bình theo hàng bằng nhau.

Yếu tố thứ hai (hàng)	**Yếu tố thứ nhất (cột)**			
	1	**2**		**c**
1	X_{11}	X_{12}	…	X_{1k}
2	X_{21}	X_{22}	…	X_{2k}
…	…	…	…	…
h	X_{h1}	X_{h2}	…	X_{hk}

Bước 1. Xây dựng cặp giả thuyết

Giả thuyết H_0 :

- Trung bình của tổng thể theo yếu tố hàng bằng nhau

- Trung bình của tổng thể theo yếu tố cột bằng nhau

Đối thuyết H_1 : Không xảy ra ít nhất một trong hai giả thuyết trên của H_0

Bước 2. Tính các tổng theo bảng sau

	1	**2**		**c**	$T_{i*} = \sum_j X_{ij}$	$\sum_j X_{ij}^2$
1	X_{11}	X_{12}	$\cdots$	X_{1c}	T_{1*}	$\sum_j X_{1j}^2$
2	X_{21}	X_{22}	$\cdots$	X_{2c}	T_{2*}	$\sum_j X_{2j}^2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
h	X_{h1}	X_{h2}	$\cdots$	X_{hc}	T_{h*}	$\sum_j X_{hj}^2$
$T_{*j} = \sum_i X_{ij}$	T_{*1}	T_{*2}	$\cdots$	T_{*c}	$T = \sum_{i,j} X_{ij}$	
$\sum_i X_{ij}^2$	$\sum_i X_{i1}^2$	$\sum_i X_{i2}^2$	$\cdots$	$\sum_i X_{ic}^2$		$\sum_{i,j} X_{ij}^2$

Bảng ANOVA

Nguồn	**SS**	**df**	**MS**	**F**
Yếu tố A	$SSA = \frac{1}{c}\sum_{i=1}^{h} T_{i*}^2 - \frac{T^2}{h \cdot c}$	$h-1$	$MSA = \frac{SSA}{h-1}$	$F_A = \frac{MSA}{MSE}$
Yếu tố B	$SSB = \frac{1}{h}\sum_{j=1}^{c} T_{*j}^2 - \frac{T^2}{h \cdot c}$	$c-1$	$MSB = \frac{SSB}{c-1}$	$F_B = \frac{MSB}{MSE}$
Sai số	$SSE = SST - SSA - SSB$	$(h-1)(c-1)$	$MSE = \frac{SSE}{(h-1)(c-1)}$	
Tổng	$SST = \sum_{i,j} X_{ij}^2 - \frac{T^2}{h \cdot c}$	$h \cdot c - 1$		

Bước 3. So sánh và kết luận

- Nếu $F_A > f_\alpha\left[(h-1),(h-1)(c-1)\right]$ thì ta bác bỏ yếu tố A (hàng).

- Nếu $F_B > f_\alpha\left[(c-1),(h-1)(c-1)\right]$ thì ta bác bỏ yếu tố B (cột).

Ví dụ 5.2. Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa và năng suất lúa. Kết quả thu thập như sau:

Loại phân bón	Giống A	Giống B	Giống C
1	65	69	75
2	74	72	70
3	64	68	78
4	83	78	76

Hãy kết luận với mức ý nghĩa 5%.

Giải

Bước 1. Xây dựng cặp giả thuyết

Giả thuyết H_0 :

- Năng suất trung bình không phụ thuộc vào phân bón
- Năng suất trung bình không phụ thuộc vào giống lúa

Đối thuyết H_1 : Không xảy ra ít nhất một trong hai giả thuyết trên của H_0

Bước 2. Lập bảng tính các tổng

Loại phân bón	Giống A	Giống B	Giống C	$T_{i*} = \sum_j X_{ij}$	$\sum_j X_{ij}^2$
1	65	69	75	209	14611
2	74	72	70	216	15560
3	64	68	78	210	14804
4	83	78	76	237	18749
$T_{*j} = \sum_i X_{ij}$	286	287	299	872	
$\sum_i X_{ij}^2$	20686	20653	22385		63724

- Tổng bình phương độ lệch toàn bộ

$$SST = \sum_{i,j} X_{ij}^2 - \frac{T^2}{h \cdot c} = 63724 - \frac{872^2}{12} = \frac{1076}{3} = 358,67$$

- Tổng bình phương độ lệch của từng yếu tố hàng:

$$SSA = \frac{1}{c}\sum_{i=1}^{h} T_{i*}^2 - \frac{T^2}{h \cdot c} = \frac{1}{3}\left(209^2 + 216^2 + 210^2 + 237^2\right) - \frac{872^2}{4 \times 3} = 170$$

- Tổng bình phương độ lệch của từng yếu tố cột:

$$\mathrm{SSB} = \frac{1}{h}\sum_{j=1}^{c} T_{*j}^2 - \frac{T^2}{h \cdot c} = \frac{1}{4}\left(286^2 + 287^2 + 299^2\right) - \frac{872^2}{4 \times 3} = \frac{157}{6} = 26,167$$

- Tổng bình phương độ lệch của sai số:

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSA} - \mathrm{SSB} = 162,5$$

Bước 3. Tính phương sai

- Phương sai giữa các nhóm yếu tố hàng:

$$\mathrm{MSA} = \frac{\mathrm{SSA}}{h-1} = \frac{170}{4-1} \approx 56,6667.$$

- Phương sai giữa các nhóm yếu tố cột:

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{c-1} = \frac{26,1667}{3-1} = 13,0833.$$

- Phương sai phần dư:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{(h-1)(k-1)} = \frac{162,5}{(4-1)(3-1)} \approx 27,0833.$$

Bước 4. Tính giá trị kiểm định

- Kiểm định theo hàng:

$$F_A = \frac{\mathrm{MSA}}{\mathrm{MSE}} = \frac{56,667}{27,0833} = 2,0923.$$

- Kiểm định theo cột:

$$F_B = \frac{\mathrm{MSB}}{\mathrm{MSE}} = \frac{13,0833}{27,0833} \approx 0,4831.$$

Bước 5. So sánh và kết luận

Với mức ý nghĩa 5%, giá trị tới hạn theo yếu tố hàng: $C_1 = f_{0,05}\left(3,6\right) = 4,76$; giá trị tới hạn theo cột: $C_2 = f_{0,05}\left(2,6\right) = 5,14$.

Ta có $F_A \le C_1$ chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, năng suất không phụ thuộc vào phân bón.

Ta có $F_B \le C_2$ chưa đủ cơ sở bác bỏ H_0. Vậy với mức ý nghĩa 5%, năng suất không phụ thuộc vào giống lúa.

Ví dụ 5.3. Trong một đề tài nghiên cứu sự phụ thuộc giữa mức lương với kết quả học tập và môi trường học, người ta khảo sát ngẫu nhiên một số sinh viên mới tốt nghiệp ở 3 trường đại học và thu thập được bảng kết quả sau:

Kết quả học tập	Trường đại học		
	A	B	C
Trung bình	5	6	7
Trung bình khá	6	7	9
Khá	7	8	11
Giỏi	9	10	14
Xuất sắc	13	15	20

Dựa vào kết quả được tính sẵn ở bảng sau:

	Tổng các chênh lệch bình phương	Bậc tự do	Phương sai	Tỷ số F
Kết quả học tập	185,07	…	…	43,375
Môi trường học tập	…	…	…	21,938
Phần dư	8,5333	8	1,0667	
Tổng	240,4	14		

a) Hãy điền vào các chỗ còn thiếu trong bảng

b) Với mức ý nghĩa 5%, hãy kiểm định giả thuyết cho rằng không có sự phụ thuộc của tiền lương vào kết quả học tập và môi trường học. Nếu có thì sự phụ thuộc xảy ra giữa mức lương và kết quả học tập hay môi trường học?

Giải

a) Hãy điền vào các chỗ còn thiếu trong bảng

Ta có

$$h = 5, c = 3,\ SSA = 185,07;\ SSE = 8,5333;\ SST = 240,4$$

Suy ra

$$SSB = 240,4 - 185,07 - 8,5333 = 46,8$$

$$MSA = \frac{SSA}{h-1} = \frac{185,07}{5-1} \approx 46,2675.$$

$$MSB = \frac{SSB}{c-1} = \frac{46,8}{3-1} = 23,4.$$

Ta có bảng kết quả

	Tổng các chênh lệch bình phương	Bậc tự do	Phương sai	Tỷ số F
Kết quả học tập	185,07	4	46,267	43,375
Môi trường học tập	46,8	2	23,4	21,938
Phần dư	8,5333	8	1,0667	
Tổng	240,4	14		

b) Dựa vào kết quả câu a, ta xét hai trường hợp ứng với cặp giả thuyết như sau:

Trường hợp 1.

H_0 : Kết quả học tập không có sự khác biệt về trung bình dựa trên giá trị thu nhập của mức lương.

H_1 : Kết quả học tập có sự khác biệt về trung bình dựa trên giá trị thu nhập của mức lương.

Với mức ý nghĩa 5%, tra bảng Fisher ta có giá trị tới hạn theo kết quả học tập:

$$C_1 = f_{0,05}(4,8) = 3,84$$

Ta có $F_1 = 43,375 > C_1 = 3,84$ nên bác bỏ H_0. Vậy với mức ý nghĩa 5%, kết quả học tập có sự khác biệt về trung bình dựa trên giá trị thu nhập của mức lương.

Trường hợp 2.

H_0 : Môi trường học tập không có sự khác biệt về trung bình dựa trên giá trị thu nhập của mức lương.

H_1 : Môi trường học tập có sự khác biệt về trung bình dựa trên giá trị thu nhập của mức lương.

Với mức ý nghĩa 5%, tra bảng Fisher ta có giá trị tới hạn theo kết quả học tập

$$C_2 = f_{0,05}(2,8) = 4,46$$

Ta có $F_2 = 21,938 > C_2 = 4,46$ nên bác bỏ H_0. Vậy với mức ý nghĩa 5%, môi trường học tập có sự khác biệt về trung bình dựa trên giá trị thu nhập của mức lương.

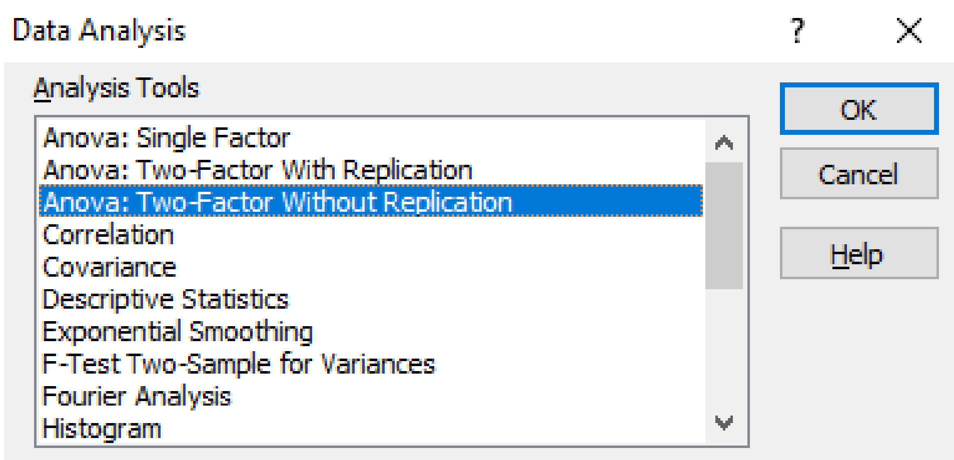
Dùng Excel phân tích phương sai hai yếu tố không lặp

Với ví dụ 2 ở trên, ta thực hiện các bước sau trên Excel như sau:

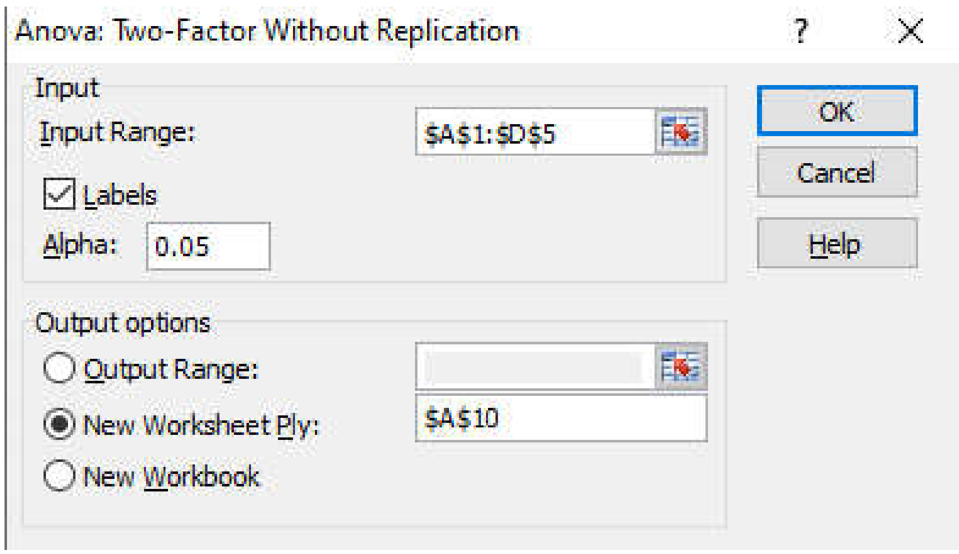
Bước 1. Nhập dữ liệu theo như sau

Loại phân bón	Giống A	Giống B	Giống C
1	65	69	75
2	74	72	70
3	64	68	78
4	83	78	76

Bước 2. Từ màn hình Excel chọn Data analysis sau đó chọn : **Anova: Two- Factor Without Replicaion**



Nhấn **OK** màn hình xuất hiện như sau



Trong đó

+) Input Range: chọn vùng dữ liệu (quét bảng dữ liệu)

+) Labels in First Row: Tên các giống lúa A, B, C và các loại phân bón 1, 2, 3, 4

+) Alpha: Mức ý nghĩa 5%

+) New Worsheet ply: Chọn vùng xuất kết quả là A10

Bước 3. Nhấn **Ok**, ta được kết quả

Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
1	3	209	69.66667	25.33333
2	3	216	72	4
3	3	210	70	52
4	3	237	79	13
Giống A	4	286	71.5	79
Giống B	4	287	71.75	20.25
Giống C	4	299	74.75	11.58333

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	170	3	56.66667	2.092308	0.202691	4.757063
Columns	26.16667	2	13.08333	0.483077	0.638961	5.143253
Error	162.5	6	27.08333			
Total	358.6667	11				

5.2.2. Phân tích phương sai hai yếu tố có lặp

Trong trường hợp tương ứng với mỗi yếu tố cột và yếu tố hàng ta có thể chọn nhiều quan sát. Trong phần này, ngoài việc kiểm định về trung bình theo cột bằng nhau, trung bình theo hàng bằng nhau mà chúng ta còn xem xét sự tương tác giữa yếu tố hàng và yếu tố cột có ảnh hưởng đến hiện tượng nghiên cứu hay không. Ta có bảng kết hợp hai tiêu thức như sau:

Yếu tố thứ hai (hàng)	Yếu tố thứ nhất (cột)			
	1	2		c
1	X_{111}	X_{121}	…	X_{1c1}
	X_{112}	X_{122}	…	X_{1c2}
	…	…	…	…
	X_{11r}	X_{12r}	…	X_{1cr}
2	X_{211}	X_{221}	…	X_{2c1}
	X_{212}	X_{222}	…	X_{2c2}
	…	…	…	…
	X_{21r}	X_{22r}	…	X_{2cr}
…	…	…	…	…
h	X_{h11}	X_{h21}	…	X_{hc1}

	X_{h12}	X_{h22}	…	X_{hc2}
	…	…	…	…
	X_{h1r}	X_{h2r}	…	X_{hcr}

Bước 1. Xây dựng cặp giả thuyết

Giả thuyết H_0 :

- Trung bình của tổng thể theo yếu tố hàng bằng nhau
- Trung bình của tổng thể theo yếu tố cột bằng nhau
- Không có sự tương tác giữa yếu tố hàng và yếu tố cột

Đối thuyết H_1 : Không xảy ra ít nhất một trong ba giả thuyết trên của H_0

Bước 2. Lập bảng tính các tổng

Yếu tố thứ hai (hàng)	**Yếu tố thứ nhất (cột)**				
	1	**2**		**c**	T_{i**}
1	X_{111}	X_{121}	…	X_{1c1}	$T_{1**} = \sum_{j,k} X_{1jk}$
	X_{112}	X_{122}	…	X_{1c2}	
	…	…	…	…	
	X_{11r}	X_{12r}	…	X_{1cr}	
2	X_{211}	X_{221}	…	X_{2c1}	$T_{2**} = \sum_{j,k} X_{2jk}$
	X_{212}	X_{222}	…	X_{2c2}	
	…	…	…	…	
	X_{21r}	X_{22r}	…	X_{2cr}	
…	…	…	…	…	…
h	X_{h11}	X_{h21}	…	X_{hc1}	$T_{h**} = \sum_{j,k} X_{hjk}$
	X_{h12}	X_{h22}	…	X_{hc2}	
	…	…	…	…	
	X_{h1r}	X_{h2r}	…	X_{hcr}	
T_{*j*}	$T_{*1*} = \sum_{i,k} X_{i1k}$	$T_{*2*} = \sum_{i,k} X_{i2k}$		$T_{*c*} = \sum_{i,k} X_{ick}$	$T = \sum_{i,j,k} X_{ijk}$

- Tổng bình phương độ lệch toàn bộ

$$SST = \sum_{i,j,k}\left(X_{ijk} - \overline{X}\right)^2 = \sum_{i,j,k} X_{ijk}^2 - \frac{T^2}{hcr}$$

- Tổng bình phương độ lệch của từng yếu tố hàng:

$$SSA = cr\sum_{i}\left(\overline{X}_{i**} - \overline{X}\right)^2 = \frac{1}{cr}\sum_{i} T_{i**}^2 - \frac{T^2}{hcr}$$

- Tổng bình phương độ lệch của từng yếu tố cột:

$$SSB = hr\sum_{j}\left(\overline{X}_{*j*} - \overline{X}\right)^2 = \frac{1}{hr}\sum_{j} T_{*j*}^2 - \frac{T^2}{hcr}$$

- Tổng bình phương độ lệch của yếu tố tương tác :

$$SSAB = r\sum_{i,j}\left(\overline{X}_{ij*} - \overline{X}_{i**} + \overline{X}_{*j*} - \overline{X}\right)^2$$

$$= \frac{1}{r}\sum_{r} T_{ij*}^2 - \frac{1}{cr}\sum_{i} T_{i**}^2 - \frac{1}{hr}\sum_{j} T_{*j*}^2 + \frac{T^2}{hcr}$$

trong đó

$$T_{ij*}^2 = \left(\sum_{k} X_{ijk}\right)^2$$

- Tổng bình phương độ lệch của sai số:

$$SSE = SST - SSA - SSB - SSAB = \sum_{i,j,k} X_{ijk}^2 - \frac{1}{r}\sum_{i,j} X_{ij*}^2$$

Bảng ANOVA

Nguồn	**SS**	**df**	**MS**	**F**
Yếu tố A	SSA	$h-1$	$MSA = \frac{SSA}{h-1}$	$F_A = \frac{MSA}{MSE}$
Yếu tố B	SSB	$c-1$	$MSB = \frac{SSB}{c-1}$	$F_B = \frac{MSB}{MSE}$
Tương tác AB	SSAB	$(h-1)(c-1)$	$MSAB = \frac{SSAB}{(h-1)(c-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Sai số	SSE	$hc(r-1)$	$MSE = \frac{SSE}{hc(r-1)}$	
Tổng	SST	$hcr-1$		

Bước 3. So sánh và kết luận

- Nếu $F_A > f_\alpha\left[(h-1), hc(r-1)\right]$ thì ta bác bỏ yếu tố A (hàng).

- Nếu $F_B > f_\alpha\left[(c-1), hc(r-1)\right]$ thì ta bác bỏ yếu tố B (cột).

- Nếu $F_{AB} > f_\alpha\left[(h-1)(c-1), hc(r-1)\right]$ thì ta bác bỏ yếu tố tương tác.

Ví dụ 5.4. Điều tra mức tăng trưởng chiều cao của một loại cây trồng theo loại đất trồng và phân bón ta có kết quả sau :

Loại đất / Phân bón	1	2	3
A	5,5	4,5	3,5
	5,5	4,5	4,0
	6,0	4,0	3,0
B	5,6	5,0	4,0
	7,0	5,5	5,0
	7,0	5,0	4,5

Hỏi có sự khác nhau của mức tăng trưởng chiều cao theo loại đất và loại phân bón không? Hãy kết luận với mức ý nghĩa 5%.

Giải

Bước 1. Xây dựng cặp giả thuyết

Giả thuyết H_0 :

- Mức tăng trưởng chiều cao trung bình theo các mức phân bón bằng nhau
- Mức tăng trưởng chiều cao trung bình theo các loại đất bằng nhau
- Không có sự tương tác giữa mức phân bón và loại đất

Đối thuyết H_1 : Không xảy ra ít nhất một trong ba giả thuyết trên của H_0

Bước 2. Lập bảng tính các tổng

Loại đất / Phân bón	1	2	3	T_{i**}
A	5,5	4,5	3,5	$T_{1**} = 40,5$
	5,5	4,5	4,0	
	6,0	4,0	3,0	
B	5,6	5,0	4,0	$T_{2**} = 48,6$
	7,0	5,5	5,0	
	7,0	5,0	4,5	

T_{*j*}	$T_{*1*} = 36,6$	$T_{*2*} = 28,5$	$T_{*3*} = 24$	$T = 89,1$

Ta có

$$\sum_{i,j} T_{ij*}^2 = \sum_{i,j}\left(\sum_k X_{ijk}\right)^2 = 17^2 + 13^2 + 10,5^2 + 19,6^2 + 15,5^2 + 13,5^2 = 1374,91$$

- Tổng bình phương độ lệch toàn bộ

$$SST = \sum_{i,j,k} X_{ijk}^2 - \frac{T^2}{hcr} = 461,11 - \frac{89,1^2}{18} = 20,065$$

- Tổng bình phương độ lệch của từng yếu tố hàng:

$$SSA = \frac{1}{cr}\sum T_{i**}^2 - \frac{T^2}{hcr} = \frac{1}{9}\left(40,5^2 + 48,6^2\right) - \frac{89,1^2}{18} = 3,645$$

- Tổng bình phương độ lệch của từng yếu tố cột:

$$SSB = \frac{1}{hr}\sum T_{*j*}^2 - \frac{T^2}{hcr} = \frac{1}{6}\left(36,6^2 + 28,5^2 + 24^2\right) - \frac{89,1^2}{18} = 13,59$$

- Tổng bình phương độ lệch của yếu tố tương tác :

$$SSAB = \frac{1}{r}\sum T_{ij*}^2 - \frac{1}{cr}\sum_i T_{i**}^2 - \frac{1}{hr}\sum_j T_{*j*}^2 + \frac{T^2}{hcr}$$

$$= \frac{1}{3}\cdot 1374,91 - 444,69 - 454,635 + \frac{89,1^2}{18} = 0,0233$$

- Tổng bình phương độ lệch của sai số:

$$SSE = SST - SSA - SSB - SSAB = 2,8067$$

Bước 3. Tính phương sai

- Phương sai giữa các nhóm yếu tố hàng:

$$MSA = \frac{SSA}{h-1} = \frac{3,645}{2-1} = 3,645.$$

- Phương sai giữa các nhóm yếu tố cột:

$$MSB = \frac{SSB}{c-1} = \frac{13,59}{3-1} = 6,795$$

- Phương sai giữa các nhóm yếu tố tương tác:

$$MSAB = \frac{SSAB}{(h-1)(c-1)} = \frac{0,0233}{(2-1)(3-1)} = 0,012$$

- Phương sai phần dư:

$$\text{MSE} = \frac{\text{SSE}}{hc(r-1)} = \frac{2,8067}{12} = 0,234$$

Bước 4. Tính giá trị kiểm định

- Kiểm định theo hàng: $F_A = \frac{\text{MSA}}{\text{MSE}} = \frac{3,645}{0,234} = 15,577.$

- Kiểm định theo cột: $F_B = \frac{\text{MSB}}{\text{MSE}} = \frac{6,795}{0,234} = 29,0385.$

- Kiểm định theo yếu tố tương tác: $F_{AB} = \frac{\text{MSAB}}{\text{MSE}} = \frac{0,012}{0,234} \approx 0,0513.$

Bước 5. So sánh và kết luận

Với mức ý nghĩa 5%, giá trị tới hạn theo yếu tố hàng: $C_1 = f_{0,05}(1,12) = 4,75;$ giá trị tới hạn theo yếu tố cột: $C_2 = f_{0,05}(2,12) = 3,89;$ giá trị tới hạn theo yếu tố tương tác: $C_3 = f_{0,05}(2,12) = 3,89.$

Ta có $F_A > C_1$ bác bỏ yếu tố hàng. Vậy với mức ý nghĩa 5%, mức tăng trưởng chiều cao trung bình theo các mức phân bón không bằng nhau.

Ta có $F_B > C_2$ bác bỏ yếu tố cột. Vậy với mức ý nghĩa 5%, mức tăng trưởng chiều cao trung bình theo các loại đất không bằng nhau.

Ta có $F_{AB} \le C_3$ chưa đủ cơ sở bác bỏ yếu tố tương tác . Vậy với mức ý nghĩa 5%, không có sự tương tác giữa mức phân bón và loại đất.

Ví dụ 5.5. Trong một đề tài nghiên cứu sự phụ thuộc giữa mức lương với kết quả học tập và môi trường học, người ta khảo sát ngẫu nhiên một số sinh viên mới tốt nghiệp ở 3 trường đại học và thu thập được bảng kết quả sau:

Kết quả học tập	Trường đại học		
	A	B	C
Trung bình	4	6	7
	5	7	8
	6	5	7
	5	6	9
Trung bình khá	6	7	9
	5	7	8
	6	6	7

	5	6	6
Khá	7	8	11
	6	7	10
	8	7	9
	7	9	8
Giỏi	9	10	14
	7	11	16
	8	12	18
	10	12	20

Dựa vào kết quả được tính sẵn ở bảng sau:

Phân tích phương sai hai yếu tố				
	Tổng bình phương các chênh lệch	Bậc tự do	Phương sai	Tỷ số F
Kết quả học tập	277,06	…	…	…
Môi trương học	127,79	…	…	…
Tương tác	59,375	6	9,8958	6,9512
Phần dư	51,25	36	1,4236	
Tổng	515,48	47		

a) Hãy điền vào các chỗ còn thiếu trong bảng

b) Với mức ý nghĩa 5%, hãy kiểm định giả thuyết sau:

Giả thuyết 1: Không có sự phụ thuộc giữa mức lương vào kết quả học tập và môi trường học tập.

Giả thuyết 2: Không có sự phụ thuộc giữa môi trường học tập và kết quả học tập.

Giải

a) Hãy điền vào các chỗ còn thiếu trong bảng

Ta có $h = 4, c = 3$

$SSA = 277,06; SSB = 127,79;\ SSAB = 59,375;\ SSE = 51,25;\ SST = 515,48$

Suy ra

$$MSA = \frac{SSA}{h-1} = \frac{277,06}{4-1} \approx 92,3533.$$

$$MSB = \frac{SSB}{c-1} = \frac{127,79}{3-1} = 63,895.$$

$$MSAB = \frac{SSAB}{(h-1)(c-1)} = \frac{59,375}{6} = 9,896.$$

$$F_A = \frac{SSA}{SSE} = \frac{92,3533}{1,4236} = 64,873$$

$$F_B = \frac{SSB}{SSE} = \frac{63,896}{1,4236} = 44,883$$

Ta có bảng kết quả

Phân tích phương sai hai yếu tố				
	Tổng bình phương các chênh lệch	Bậc tự do	Phương sai	Tỷ số F
Kết quả học tập	277,06	3	92,353	64,873
Môi trương học	127,79	2	63,895	44,883
Tương tác	59,375	6	9,896	6,9512
Phần dư	51,25	36	1,4236	
Tổng	515,48	47		

b) Dựa vào kết quả câu a, ta xét hai trường hợp ứng với cặp giả thuyết như sau:

Giả thuyết 1. Không có sự phụ thuộc giữa mức lương vào kết quả học tập và môi trường học tập.

Với mức ý nghĩa 5%, ứng với kết quả câu a tra bảng Fisher ta có giá trị tới hạn theo kết quả học tập:

$$C_1 = f_{0,05}(3,36) = 2,87$$

Ta có $F_A = 64,873 > C_1 = 2,87$ nên bác bỏ giả thuyết 1. Vậy với mức ý nghĩa 5%, có sự phụ thuộc giữa mức lương vào kết quả học tập.

Với mức ý nghĩa 5%, ứng với kết quả câu a tra bảng Fisher ta có giá trị tới hạn theo môi trường học tập:

$$C_2 = f_{0,05}(2,36) = 3,26$$

Ta có $F_B = 44,883 > C_2 = 4,46$ nên bác bỏ giả thuyết 1. Vậy với mức ý nghĩa 5%, có sự phụ thuộc giữa mức lương vào môi trường học tập.

Giả thuyết 2. Không có sự phụ thuộc giữa môi trường học tập và kết quả học tập

Với mức ý nghĩa 5%, ứng với kết quả câu a tra bảng Fisher ta có giá trị tới hạn :

$$C_3 = f_{0,05}(6,36) = 3,26$$

Ta có $F_{AB} = 6,9512 > C_3 = 2,364$ nên bác bỏ giả thuyết 2. Vậy với mức ý nghĩa 5%, có sự phụ thuộc giữa môi trường học tập và kết quả học tập.

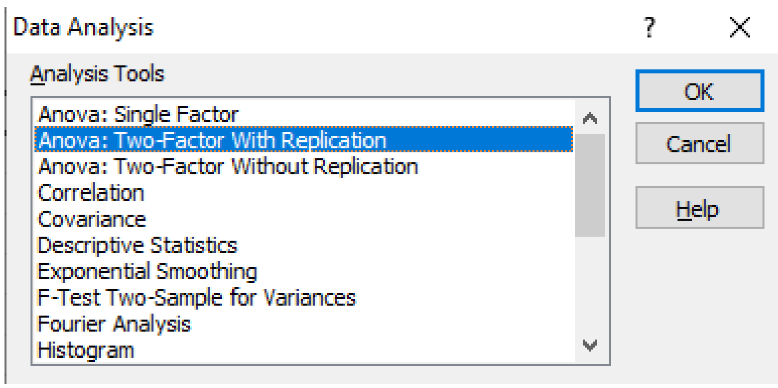
Dùng Excel phân tích phương sai hai yếu tố có lặp

Với ví dụ 5.4 trên, ta thực hiện trên Excel như sau:

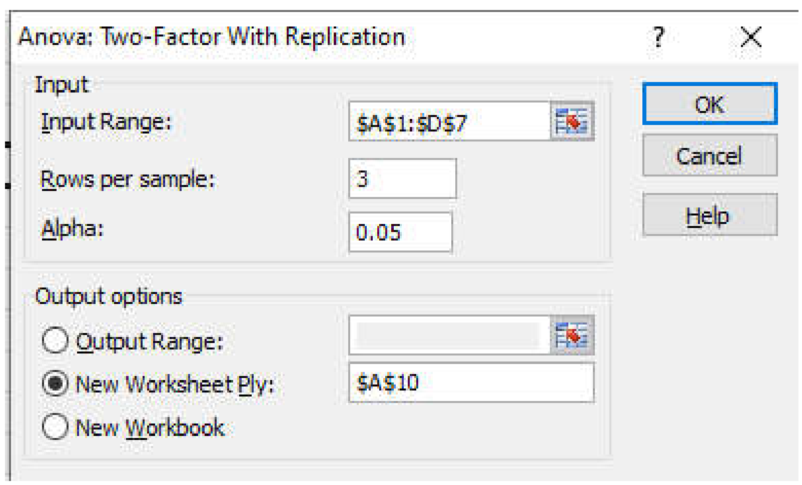
Bước 1. Nhập dữ liệu theo như sau

	1	2	3
	5.5	4.5	3.5
A	5.5	4.5	4
	6	4	3
	5.6	5	4
B	7	5.5	5
	7	5	4.5

Bước 2. Từ màn hình Excel chọn Data analysis sau đó chọn : **Anova: Two- Factor With Replicaion**



Nhấn **OK** màn hình xuất hiện như sau



Trong đó

+) Input Range: chọn vùng dữ liệu (quét bảng dữ liệu)

+) Row per sample: Số lần lặp: 3

+) Alpha: Mức ý nghĩa 5%

+) New Worsheet ply: Chọn vùng xuất kết quả là A10

Bước 3. Nhấn **Ok**, ta được kết quả

Anova: Two-Factor With Replication						
SUMMARY	1	2	3	Total		
A						
Count	3	3	3	9		
Sum	17	13	10.5	40.5		
Average	5.666667	4.333333	3.5	4.5		
Variance	0.083333	0.083333	0.25	1		
B						
Count	3	3	3	9		
Sum	19.6	15.5	13.5	48.6		
Average	6.533333	5.166667	4.5	5.4		
Variance	0.653333	0.083333	0.25	1.0525		
Total						
Count	6	6	6			
Sum	36.6	28.5	24			
Average	6.1	4.75	4			
Variance	0.52	0.275	0.5			
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	3.645	1	3.645	15.58432	0.001936	4.747225
Columns	13.59	2	6.795	29.05226	2.52E-05	3.885294
Interaction	0.023333	2	0.011667	0.049881	0.951539	3.885294
Within	2.806667	12	0.233889			
Total	20.065	17				

5.3. Tóm tắt chương 5

1. Phân tích phương sai một yếu tố

Bước 1. Đặt giả thuyết

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: H_0 \text{ sai} \end{cases}$$

Bước 2. Tính trung bình mẫu

- Tổng số quan sát : $n = \sum_{j=1}^{k} n_j$

- Trung bình mẫu nhóm j $(j = 1,2,...,k)$: $\bar{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$

- Trung bình mẫu chung của k nhóm: $\bar{X} = \dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}}{\sum_{j=1}^{k} n_j}$

Bước 3. Tính tổng bình phương độ lệch:

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}\right)^2 \text{ ; } SSA = \sum_{j=1}^{k} n_j\left(\overline{X}_j - \overline{X}\right)^2 \text{ ; } SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2$$

Bước 4. Tính phương sai: $MSE = \dfrac{SSE}{n-k}$; $MSA = \dfrac{SSA}{k-1}$

Bước 5. Tính giá trị kiểm định: $F = \dfrac{MSA}{MSE}$

Với mức ý nghĩa α cho trước, ta tìm được giá trị tới hạn trong bảng Fisher

$$C = f_\alpha\left(k-1, n-k\right)$$

+) Nếu $F > f_\alpha\left(k-1, n-k\right)$ thì bác bỏ H_0.

+) Nếu $F \le f_\alpha\left(k-1, n-k\right)$ thì chưa đủ cơ sở bác bỏ H_0.

Ta có thể tóm tắt các bước kiểm định trên trong bảng sau:

Phân tích phương sai 1 yếu tố				
	Tổng các chênh lệch bình phương	**Bậc tự do**	**Phương sai**	**Tỷ số (F)**
Giữa các nhóm	SSA	$k-1$	MSA	$\dfrac{MSA}{MSE}$
Nội bộ nhóm	SSE	$n-k$	MSE	
Tổng	SST	$n-1$		

2. Phân tích phương sai hai yếu tố có lặp

Bước 1. Xây dựng cặp giả thuyết

Giả thuyết H_0 :

- Trung bình của tổng thể theo yếu tố hàng bằng nhau
- Trung bình của tổng thể theo yếu tố cột bằng nhau
- Không có sự tương tác giữa yếu tố hàng và yếu tố cột

Đối thuyết H_1 : Không xảy ra ít nhất một trong ba giả thuyết trên của H_0

Bước 2. Tính các tổng

$$SST = \sum_{i,j,k}\left(X_{ijk} - \overline{X}\right)^2 \text{ ; } SSA = cr\sum_{i}\left(\overline{X}_{i**} - \overline{X}\right)^2$$

$$SSB = hr\sum_{j}\left(\overline{X}_{*j*} - \overline{X}\right)^2 \text{ ; } SSAB = r\sum_{i,j}\left(\overline{X}_{ij*} - \overline{X}_{i**} + \overline{X}_{*j*} - \overline{X}\right)^2$$

$$SSE = SST - SSA - SSB - SSAB = \sum_{i,j,k} X_{ijk}^2 - \frac{1}{r}\sum_{i,j} X_{ij*}^2$$

Bảng ANOVA

Nguồn	**SS**	**df**	**MS**	**F**
Yếu tố A	SSA	$h-1$	$MSA = \frac{SSA}{h-1}$	$F_A = \frac{MSA}{MSE}$
Yếu tố B	SSB	$c-1$	$MSB = \frac{SSB}{c-1}$	$F_B = \frac{MSB}{MSE}$
Tương tác AB	SSAB	$(h-1)(c-1)$	$MSAB = \frac{SSAB}{(h-1)(c-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Sai số	SSE	$hc(r-1)$	$MSE = \frac{SSE}{hc(r-1)}$	
Tổng	SST	$hcr-1$		

Bước 3. So sánh và kết luận

- Nếu $F_A > f_\alpha\left[(h-1), hc(r-1)\right]$ thì ta bác bỏ yếu tố A (hàng).

- Nếu $F_B > f_\alpha\left[(c-1), hc(r-1)\right]$ thì ta bác bỏ yếu tố B (cột).

- Nếu $F_{AB} > f_\alpha\left[(h-1)(c-1), hc(r-1)\right]$ thì ta bác bỏ yếu tố tương tác.

5.4. Bài tập

Bài số 1. Trong một chủ đề nghiên cứu, người ta muốn tìm hiểu xem doanh số bán hàng và vị trí cửa hàng trong cùng một chuỗi cửa hàng có sự phụ thuộc vào nhau hay không. Người ta hiến hành khảo sát ngẫu nhiên trên một số cửa hàng và doanh số của các cửa hàng này thì thu được bảng số liệu sau:

	Cửa hàng		
Tháng	A	B	C
1	22,2	24,6	22,7
4	19,9	23,1	21,9
6	20,3	22	23,3
8	21,4	23,5	24,1
9	21,2	23,6	22,1
11	21	22,1	23,4
12	20,3	23,5	

Giả định rằng doanh số bán hàng tuân theo luật phân phối chuẩn, phương sai về sự biến động doanh số giữa các cửa hàng là bằng nhau.

Dựa vào bảng kết quả trên, ta có bảng kết quả tính sẵn như sau

Phân tích phương sai một yếu tố (ANOVA)					
	Tổng các chênh lệch bình phương	Bậc tự do	Phương sai	Tỷ số F	P_value
Giữa các nhóm	….	2	……	……	0,0001
Nội bộ nhóm	7,934	……	0,57		
Tổng	28,1212	16			

1. Hãy điền các giá trị còn thiếu trong bảng trên.
2. Với mức ý nghĩa 5%, kiểm định xem doanh số bán hàng giữa các cửa hàng có sự khác biệt không?

Đáp số: 1. $SSA = 20,188;\ n - k = 14;\ MSA = 10,094;\ F = 17,812$.

2. Có sự khác biệt doanh số bán hàng của cửa hàng.

Bài số 2. Một nhà sản xuất nước giải khát đang xem xét ba màu lon cho một loại nước ngọt: đỏ, vàng, xanh có ảnh hưởng đến doanh thu như thế nào. 16 cửa hàng được chọn ra để gởi các lon nước ngọt đến bán. Những lon màu đỏ được gởi đến 6 cửa hàng. Những lon màu vàng được gởi đến 5 cửa hàng khác và các lon màu xanh cũng được gởi đến 5 cửa hàng còn lại. Sau một vài ngày, nhà sản xuất kiểm tra ở các cửa hàng thì doanh số bán nước ngọt được cho trong bảng số liệu sau:

	Màu lon nước ngọt		
Doanh số (ngàn đồng)	Đỏ	Vàng	Xanh
Cửa hàng 1	43	52	61
Cửa hàng 2	52	37	29
Cửa hàng 3	59	38	38
Cửa hàng 4	76	64	53
Cửa hàng 5	61	74	79
Cửa hàng 6	81		

Giả định rằng doanh số bán hàng tuân theo luật phân phối chuẩn, phương sai về sự biến động doanh số giữa các màu là bằng nhau.

Dựa vào bảng kết quả trên, ta có bảng kết quả tính sẵn như sau

Phân tích phương sai một yếu tố (ANOVA)

	Tổng các chênh lệch bình phương	Bậc tự do	Phương sai	Tỷ số F	P_value
Giữa các nhóm	340,9375	….	170,4688	……	0,556
Nội bộ nhóm	…..	13	……		
Tổng	3948,938	15			

1. Hãy điền các giá trị còn thiếu trong bảng trêm.

2. Với mức ý nghĩa 5%, kiểm định giả thuyết cho rằng không có sự khác biệt về doanh số bán hàng của ba màu nước ngọt này?

Đáp số: 1 $SSE = 3608$; $k - 1 = 2$; $MSE = 277,5385$; $F = 0,61422$;

2.Doanh số bán hàng của ba màu nước ngọt không có sự khác biệt.

Bài số 3. Để đánh giá chất lượng của 3 loại cà phê mới được tung ra thị trường, người ta tiến hành khảo sát trên 1 nhóm khách hàng (đánh giá trên thang điểm 10) và kết quả đánh giá 3 loại cà phê được trình bày như sau:

STT	Mẫu A	Mẫu B	Mẫu C
1	5	8	5
2	6	9	6
3	7	10	5
4	5	7	4
5	9	8	5
6	6	6	6
7	6	9	
8	5		

Dùng phân tích phương sai. Với mức ý nghĩa 5%, có thể nói có sự khác biệt về chất lượng của 3 loại cà phê này hay không?

Đáp số: Ba loại cà phê có sự khác biệt về chất lượng.

Bài số 4. Để đánh giá hoạt động kinh doanh của 3 cửa hàng. Ghi nhận doanh thu trong 7 ngày tại 3 cửa hàng như sau (đơn vị triệu đồng):

STT	CH X	CH Y	CH Z
1	6	10	5
2	10	9	6
3	5	10	15
4	15	12	4

5	6	9	10
6	9	6	6
7	10	5	5

Dùng phân tích phương sai. Với mức ý nghĩa 5%, có thể nói doanh thu ở 3 cửa hàng là khác nhau hay không?

Đáp số: Doanh thu ở 3 cửa hàng không có sự khác biệt.

Bài số 5. Để đánh giá chất lượng của 3 loại cà phê mới được tung ra thị trường, người ta tiến hành khảo sát trên 1 nhóm khách hàng và kết quả đánh giá 3 loại cà phê được trình bày như sau:

STT	Mẫu A	Mẫu B	Mẫu C
1	9	8	5
2	8	9	6
3	6	5	5
4	7	7	4
5	10	8	5
6	6	6	6
7	5		

Dùng phân tích phương sai

1. Tính giá trị kiểm định F
2. Với mức ý nghĩa 5%, có thể nói có sự khác biệt về chất lượng của 3 loại cà phê này hay không? nếu có thì nêu rõ.

Đáp số: 1. $F = 4,2389$; *2. Ba loại cà phê có sự khác biệt về chất lượng.*

Bài số 6. Để đánh giá chất lượng của 3 loại trà giúp cải thiện giấc ngủ cho người lớn tuổi vừa mới tung ra thị trường, người ta tiến hành khảo sát trên 3 nhóm người lớn tuổi đã thử nghiệm 3 loại trà trên, đánh giá của 3 nhóm người này đối với 3 loại trà được ghi nhận như sau:

STT	Mẫu A	Mẫu B	Mẫu C
1	9	5	8
2	7	7	4
3	8	4	5
4	7	8	5
5	8	6	7

6	9	7	8
7	8	7	7
8	8		

Dùng phân tích phương sai

1. Tính giá trị kiểm định F
2. Với mức ý nghĩa 5%, có thể nói có sự khác biệt về đánh giá của 3 loại trà này hay không?

Đáp số: 1. $F = 4,606$ *2. ba loại trà có sự khác biệt về đánh giá.*

Bài số 7. Bốn chuyên gia tài chính được yêu cầu dự đoán về tốc độ tăng trưởng (%) trong 5 năm tới của 5 công ty trong một ngành được cho như sau:

Công ty	Chuyên gia			
	A	B	C	D
1	8	12	8	14
2	13	10	9	11
3	11	9	12	12
4	9	13	11	13
5	12	11	10	10

Có thể cho rằng dự đoán tốc độ tăng trưởng trung bình là như nhau cho cả 5 công ty hay không. Hãy kết luận với mức ý nghĩa 5%.

Đáp số $F = 1,2$ *tốc độ tăng trưởng trung bình là như nhau.*

Bài số 8. Nghiên cứu về hiệu quả của ba loại thuốc A, B, C dùng điều trị chứng suy nhược thần kinh. 12 người bệnh được chia thành 4 nhóm theo mức độ bệnh 1, 2, 3, 4; trong mỗi nhóm chia ra để cùng dùng trong 3 loại thuốc trên. Sau một tuần điều trị, kết quả đánh giá bằng thang điểm như sau:

Mức độ bệnh / Thuốc	1	2	3	4
A	25	40	25	30
B	30	25	25	30
C	25	20	20	25

Hãy đánh giá hiệu quả của các loại thuốc A, B, C có khác nhau hay không? Kết luận với mức ý nghĩa 1%.

Đáp số: $F = 1,33$, *hiệu quả của các loại thuốc A, B, C có như nhau.*

Bài số 9. Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa và năng suất. Năng suất lúa được ghi nhận từ các thực nghiệm như sau :

Giống lúa / Phân bón	A	B	C
1	65	69	75
	68	71	75
	63	67	78
2	74	72	70
	78	69	69
	76	70	65
3	64	68	78
	72	73	82
	65	74	80
4	83	77	76
	82	78	77
	84	75	75

Hãy cho biết sự ảnh hưởng của phân bón, giống lúa lên năng suất? Hãy kết luận với mức ý nghĩa 5%.

Đáp số:

$F_1 = 15,64$, *có sự phụ thuộc giữa phân bón và năng suất.*

$F_2 = 13,42$, *có sự phụ thuộc giữa giống lúa và năng suất.*

Bài số 10. Một nghiên cứu sản lượng bông lúa (tạ/ha) theo mật độ trồng và mức phân bón được ghi nhận từ các thực nghiệm như sau :

Mức phân bón / Mật độ trồng	B1	B2	B3	B4
A1	16	19	19	20
	15	20	21	24
	21	23	22	21
	16	19	20	18
A2	17	19	20	20
	15	18	21	21

	16	18	22	22
	19	20	23	19
A3	18	20	22	25
	20	23	18	22
	19	21	23	21
	17	22	21	23

Hãy cho biết có sự khác nhau của sản lượng bông lúa theo mật độ trồng, theo mức phân bón ? Hãy kết luận với mức ý nghĩa 5%.

Đáp số:

$F_1 = 2,79$*, không có sự phụ thuộc giữa sản lượng và mật độ trồng.*

$F_2 = 1,56$*, không sản lượng và mức phân bón.*

5.5. Tài liệu tham khảo

[1] Phạm Văn Chững, Lê Thanh Hoa, Nguyễn Đình Uông, Thống kê ứng dụng, NXB Đại học Quốc gia Thành Phố Hồ Chí Minh, 2016.

[2] Hà văn Sơn, Giáo trình Lý thuyết Thống kê, ứng dụng trong Quản trị và kinh tế.

[3] S.P. Gordon, Contemporary Statistics, Mc Graw – Hill, Inc. 1994sworth.

[4] Anderson, Sweeney, and William [2010], Statistics for Business and Economics, South-Western Cengage Learning (11^{th} Edition).

[5] Michael Barrow, Statistics for Economics, Accounting and Business Studies- Prentice Hall, 2006.

[6] Newbold Paul - Statistics for Bussiness and Economics, 5th edition - Prentice Hall, 2005.

Thuật ngữ chính chương 5	
Tiếng Anh	**Tiếng Việt**
ANOVA table	Bảng phân tích phương sai
Analysis of Variance	Phân tích phương sai
Assumptions for analysis of variance	Các giả định để phân tích phương sai
Completely randomized design	Thiết kế hoàn toàn ngẫu nhiên
Comparisonwise type I error rate	So sánh tỷ lệ lỗi loại 1
Degrees of freedom	Bậc tự do
Different levels of a factor	Các cấp độ khác nhau của một yếu tố
Experimentwise type I error rate	Thử nghiệm tỷ lệ lỗi loại 1
Equality of k population means	Sự bằng nhau của k trung bình tổng thể
Experimental units	Đơn vị thử nghiệm
Factors	Các nhân tố
Interaction	Sự tương tác
Independent variable	Biến độc lập
Mean square due to error	Trung bình bình phương các sai số
Mean square due to treatments	Trung bình bình phương các nhóm
Multiple comparison procedures	Nhiều phương pháp so sánh
Randomized block design	Thiết kế khối ngẫu nhiên
Replication	Sự lặp lại
Response variable	Biến độc lập
Overall sample mean	Trung bình mẫu chung
Total sum of squares	Tổng bình phương
Sample mean for treatment j	Trung bình mẫu cho mẫu thứ j
Sample variance for treatment j	Phương sai mẫu cho mẫu thứ j
Single factor experiment	Thử nghiệm một yếu tố
Sum of squares due to error	Tổng bình phương các sai số
Sum of squares due to treatments	Tổng bình phương các nhóm
Sum of Squares for Factor A	Tổng bình phương yếu tố A
Sum of squares for interaction	Tổng bình phương yếu tố tương tác
Supplementary exercise	Bài tập bổ trợ (bài tập bổ sung)

Chương 6 **PHÂN TÍCH DÃY SỐ THỜI GIAN**

Mục tiêu chương 6

Chương này giúp sinh viên:

- Nắm được khái niệm và phân loại dãy số thời gian.
- Biết và tính được các chỉ tiêu phân tích dãy số thời gian.
- Phương pháp tìm hàm xu thế.
- Vận dụng dãy số thời gian để dự báo bằng 3 phương pháp: Tốc độ phát triển bình quân, lượng tăng giảm tuyệt đối bình quân và hàm xu thế.

6.1. Dãy số thời gian

Để phân tích biến động của hiện tượng qua thời gian, ta dùng phương pháp phân tích dãy số thời gian. Trong phương pháp này các giá trị quan sát không độc lập với nhau, ngược lại sự phụ thuộc của các giá trị quan sát trong dãy số là đặc điểm, cơ sở cho việc xây dựng các phương pháp nghiên cứu và dự đoán dãy số thời gian. Các phương pháp dự đoán định lượng có thể được phân chia thành hai loại: phân tích các mức độ qua thời gian và phân tích liên hệ nguyên nhân – kết quả. Phương pháp dự đoán bằng phân tích các mức độ qua thời gian liên quan đến việc tính toán các giá trị tương lai của yếu tố nghiên cứu dựa trên toàn bộ các quan sát có được ở quá khứ và hiện tại. Phân tích liên hệ nhân quả liên quan đến việc xác định các yếu tố ảnh hưởng đến yếu tố ta muốn dự đoán, như phân tích hồi quy bội để xem GDP phụ thuộc vào lượng đầu tư trong nước, lượng đầu tư nước ngoài, dân số...

Phân tích các mức độ qua thời gian được dựa trên giả định cơ bản là các yếu tố ảnh hưởng đến biến động của hiện tượng trong quá khứ và hiện tại sẽ còn tiếp tục tồn tại với cùng tính chất, đặc điểm, cường độ như vậy đối với biến động của hiện tượng trong tương lai. Do đó mục tiêu chính của phân tích dãy số thời gian là nhận ra và tách riêng các yếu tố ảnh hưởng này phục vụ cho mục đích dự đoán cũng như cho việc kiểm soát và hoạch định trong quản lý.

6.1.1. Khái niệm và phân loại

6.1.1.1. Khái niệm

Dãy số thời gian là một phương pháp phân tích thống kê được sử dụng khá phổ biến, nhằm nghiên cứu các đặc điểm, bản chất xu hướng và tính quy luật về sự phát triển của hiện tượng thường xuyên biến động theo thời gian. Dãy số thời gian là dãy các trị số của một chỉ tiêu thống kê được sắp xếp theo thứ tự thời gian.

Một dãy số thời gian có dạng tổng quát như sau:

t_i	t_1	t_2	t_3	$\cdots$	t_n
y_i	y_1	y_2	y_3	$\cdots$	y_n

trong đó

t_i $(i = 1, 2, ..., n)$: thời gian thứ i.

y_i $(i = 1, 2, ..., n)$: giá trị của chỉ tiêu ứng với thời gian thứ i.

Ví dụ 6.1. Thống kê kết quả tiêu thụ sản phẩm X, Y, Z trong 5 năm tại khu vực Thành phố Hồ Chí Minh (Đơn vị tính: tấn).

Năm	2016	2015	2014	2013	2012
X	90	95	92	87	89
Y	78	76	75	70	69
Z	80	82	85	82	80

Qua dãy số thời gian về mức độ tiêu thụ sản phẩm của khách hàng, doanh nghiệp có thể nghiên cứu các đặc điểm và sự biến động của sản phẩm. Qua đó có thể nắm được xu hướng và tính quy luật của sự phát triển đồng thời dự đoán được các mức độ phát triển trong tương lai.

6.1.1.2. Phân loại

Căn cứ vào đặc điểm về thời gian của dãy số người ta thường chia dãy số thời gian thành 2 loại như sau:

Có 3 loại dãy số thời gian:

- Dãy số tuyệt đối: Khi các mức độ của dãy số là số tuyệt đối. Trong đó dãy số tuyệt đối lại được chia thành 2 loại là dãy số tuyệt đối thời kỳ và dãy số tuyệt đối thời điểm.

- Dãy số tương đối: Khi các mức độ của dãy số là số tương đối. Ví dụ: Tốc độ phát triển của doanh nghiệp qua các năm.

- Dãy số bình quân: Khi các mức độ của dãy số là các số bình quân. Ví dụ: năng suất lao động bình quân của công nhân của một phân xưởng được tổng hợp qua các tháng.

a. Dãy số thời kỳ: là dãy số biểu hiện sự biến động của hiện tượng nghiên cứu qua từng thời kỳ. Các mức độ trong dãy số thời kỳ có thể cộng với nhau qua thời gian, để phản ánh mặt lượng của hiện tượng nghiên cứu trong một thời kỳ dài hơn.

Ví dụ 6.2. Thống kê giá trị xuất khẩu, nhập khẩu hàng hóa của Việt Nam giai đoạn từ năm 2000 đến năm 2015 như sau

Năm	Xuất khẩu (triệu USD)	Nhập khẩu (triệu USD)
2000	14,449	15,635
2001	15,027	16,162
2002	16,706	19,733
2003	20,176	25,227
2004	26,504	31,954
2005	32,442	36,978
2006	39,826	44,891
2007	48,561	62,682
2008	62,685	80,714
2009	57,096	69,949
2010	72,237	84,839
2011	96,906	106,750
2012	114,529	113,780
2013	132,033	132,033
2014	150,217	147,852
2015	162,017	165,570

b. Dãy số thời điểm: là dãy số biểu hiện sự biến động của hiện tượng nghiên cứu qua các thời điểm nhất định. Các mức độ trong dãy số thời điểm không thể cộng lại theo thời gian vì con số cộng này không có ý nghĩa kinh tế.

Ví dụ 6.3. Thống kê sản lượng tồn kho của xí nghiệp X. Kiểm kê vào ngày 1 hàng tháng như sau:

Ngày	1/1	1/2	1/3	1/4	1/5
Sản lượng tồn kho (tấn)	380	395	350	420	442

6.1.2. Các chỉ tiêu phân tích dãy số thời gian

6.1.2.1. Mức độ trung bình theo thời gian

Là chỉ tiêu phản ánh mức độ điển hình, chung nhất của hiện tượng trong thời gian nghiên cứu.

Giả sử ta có dãy số thời gian: $y_1, y_2, ..., y_n$

Gọi $\overline{y}$: là mức độ trung bình của dãy

a. Dãy số thời kỳ: Đối với dãy số thời kỳ, mức độ bình quân theo thời gian được tính theo công thức:

$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6.1}$$

Ví dụ 6.4. Tình hình doanh thu bán hàng của công ty VLC trong 7 tháng đầu năm 2012 được cho trong bảng số liệu sau :

Doanh thu bán hàng	400	460	480	520	560	596	642

Tính doanh thu bình quân bán ra trong tháng của công ty VLC.

Giải

Để tính doanh thu bình quân bán ra trong tháng của công ty VLC, áp dụng công thức (6.1), ta có

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{400 + 460 + 480 + 520 + 560 + 596 + 642}{7} = 522,57.$$

b. Dãy số thời điểm: có 2 trường hợp

- Đối với dãy số thời điểm có các khoảng cách thời gian bằng nhau, mức độ bình quân được tính theo công thức sau:

$$\overline{y} = \frac{\frac{1}{2}y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{y_i + y_{i+1}}{2}\right) \tag{6.2}$$

- Đối với dãy số thời điểm có các khoảng cách thời gian không bằng nhau thì mức độ bình quân theo thời gian được tính theo công thức

$$\overline{y} = \frac{t_1 y_1 + t_2 y_2 + \cdots + t_n y_n}{t_1 + t_2 + \cdots + t_n} = \frac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i} \tag{6.3}$$

trong đó

y_i $(i = 1, 2, ..., n)$: mức độ thứ i trong dãy số,

t_i $(i = 1, 2, ..., n)$: độ dài thời gian tương ứng với mức độ thứ i.

Ví dụ 6.5. Có số liệu về hàng tồn kho của một xí nghiệp trong quý I như sau :

Ngày	1/1	1/2	1/3	1/4
Mức tồn kho (tấn)	500	800	200	600

Tính mức độ tồn kho bình quân quý I của xí nghiệp.

Giải

Đây là số liệu dãy số tuyệt đối thời điểm, do đó mức độ tồn kho bình quân quý I từ 1/1 đến 31/3

Áp dụng công thức (6.2), ta có

$$\bar{y} = \frac{\frac{1}{2}.500 + 800 + 200 + \frac{1}{2}.600}{4-1} = 516,67.$$

Ví dụ 6.6. Có số liệu về số công nhân tuyển dụng thêm một nhà máy như sau : Từ đầu tháng 1 đến 15/1 có số công nhân tuyển dụng thêm là 50, từ 16/1 đến 10/2 là 58 công nhân, ngày 11/2 đến 5/3 bổ sung thêm 55 công nhân, và đợt tuyển cuối cùng trong quý 1 từ 6/3 đến 31/ 3 là 48 công nhân. Xác định số công nhân tuyển thêm bình quân trong quý I (1/1 đến 31/3).

Giải

Lập bảng : Số công nhân tuyển dụng thêm của một nhà máy trong quý I

Ngày	Số công nhân tuyển	Số ngày (khoảng cách thời gian)
1/1 – 15/1	50	15
16/1 – 10/2	58	26
11/2 – 5/3	55	23
6/3 – 31/3	48	26

Đây là số liệu dãy số tuyệt đối thời kỳ, để tính số công nhân tuyển thêm bình quân trong quý I, áp dụng công thức (6.3), ta có

$$\bar{y} = \frac{\sum_{i=1}^{4} t_i y_i}{\sum_{i=1}^{4} t_i} = \frac{50\times 15 + 58\times 26 + 55\times 23 + 48\times 26}{15+26+23+26} = 53,01.$$

6.1.2.2. Lượng tăng (giảm) tuyệt đối

Lượng tăng (giảm) tuyệt đối là chỉ tiêu phản ánh sự biến động về mức độ tuyệt đối của hiện tượng giữa hai thời gian. Tuỳ theo mục đích nghiên cứu, ta có thể chọn gốc so sánh khác nhau, khi đó có các chỉ tiêu lượng tăng (giảm) tuyệt đối khác nhau.

Tùy theo mục đích nghiên cứu được chia thành 3 loại như sau:

a. Lượng tăng (giảm) tuyệt đối liên hoàn là chỉ tiêu phản ánh biến động về mức độ tuyệt đối của hiện tượng giữa hai thời gian liền nhau và được tính theo công thức:

$$\delta_i = y_i - y_{i-1},\ i = 2,3,...,n. \tag{6.4}$$

b. Lượng tăng (giảm) tuyệt đối định gốc là chỉ tiêu phản ánh sự biến động về mức độ tuyệt đối của hiện tượng trong những khoảng thời gian dài và thường lấy mức độ đầu tiên làm gốc cố định. Công thức tính.

$$\Delta_i = y_i - y_1,\ i = 2,3,...,n. \tag{6.5}$$

với y_1 : kỳ được chọn làm gốc.

Mối liên hệ giữa lượng tăng (giảm) tuyệt đối liên hoàn và lượng tăng (giảm) tuyệt đối định gốc:

$$\sum_{i=2}^{n} \delta_i = \Delta_n \tag{6.6}$$

c. Lượng tăng (giảm) tuyệt đối bình quân thể hiện một cách chung nhất lượng tăng (giảm) tính trung bình cho cả thời kỳ nghiên cứu, là số trung bình cộng của các lượng tăng giảm tuyệt đối liên hoàn.

$$\bar{\delta} = \frac{1}{n-1}\sum_{i=2}^{n} \delta_i = \frac{\Delta_n}{n-1} = \frac{y_n - y_1}{n-1}. \tag{6.7}$$

Chỉ tiêu này chỉ có ý nghĩa khi các lượng tăng (giảm) tuyệt đối liên hoàn xấp xỉ nhau, nghĩa là trong suốt thời kỳ nghiên cứu, hiện tượng tăng (giảm) với một lượng tương đối đều.

Ví dụ 6.7. Tình hình doanh thu bán hàng của công ty X trong 7 tháng đầu năm 2018 được cho trong bảng số liệu sau :

Tháng	1	2	3	4	5	6	7
Doanh thu bán hàng	390	440	470	500	560	600	640

Tính lượng tăng (giảm) tuyệt đối liên hoàn, lượng tăng (giảm) tuyệt đối định gốc và lượng tăng (giảm) tuyệt đối bình quân của doanh thu bán hàng trong tháng của công ty X.

Giải

Gọi y_i là doanh thu bán tháng thứ i hàng của công ty X

Tính lượng tăng (giảm) tuyệt đối liên hoàn, lượng tăng (giảm) tuyệt đối định gốc về doanh thu bán hàng của công ty X, ta lập bảng sau :

Tháng	1	2	3	4	5	6	7
y_i	390	440	470	500	560	600	640
$\delta_i = y_i - y_{i-1}$		50	30	30	60	40	40

$\Delta_i = y_i - y_1$		50	80	110	170	210	250

Lượng tăng (giảm) tuyệt đối bình quân của doanh thu bán hàng:

$$\bar{\delta} = \frac{y_n - y_1}{n-1} = \frac{640-390}{7-1} = \frac{125}{3}.$$

6.1.2.3. Tốc độ phát triển

Tốc độ phát triển là chỉ tiêu phản ánh xu hướng và tốc độ biến động của hiện tượng nghiên cứu qua thời gian, được tính bằng cách chia mức độ của hiện tượng ở kỳ nghiên cứu cho mức độ của hiện tượng ở kỳ gốc. Tuy nhiên, tuỳ theo mục đích nghiên cứu, có thể chọn kỳ gốc khác nhau, khi đó ta có các chỉ tiêu tốc độ phát triển khác nhau như sau:

a. Tốc độ phát triển liên hoàn: là chỉ tiêu phản ánh xu hướng và tốc độ biến động của hiện tượng giữa hai thời gian liền nhau và được tính theo công thức:

$$t_i = \frac{y_i}{y_{i-1}}, \quad i = 2,3,...,n. \tag{6.8}$$

b. Tốc độ phát triển định gốc: là chỉ tiêu phản ánh tốc độ và xu hướng biến động của hiện tượng ở những khoảng thời gian dài, được tính bằng cách so sánh mức độ của hiện tượng ở kỳ nghiên cứu với mức độ ở kỳ được chọn làm gốc so sánh cố định (thường chọn là kỳ đầu tiên) theo công thức:

$$T_i = \frac{y_i}{y_1}, i = 2,3,...,n. \tag{6.9}$$

Mối liên hệ giữa tốc độ phát triển liên hoàn và tốc độ phát triển định gốc

+) Tích các tốc độ phát triển liên hoàn bằng tốc độ phát triển gốc

$$\prod_{i=2}^{n} t_i = T_n \tag{6.10}$$

+) Tỷ số giữa hai tốc độ phát triển định gốc liền nhau trong dãy số bằng tốc độ phát triển liên hoàn.

$$\frac{T_i}{T_{i-1}} = t_i \tag{6.11}$$

c. Tốc độ phát triển bình quân là chỉ tiêu thể hiện nhịp độ phát triển đại diện của hiện tượng trong suốt thời kỳ nghiên cứu.

$$\bar{t} = \sqrt[n-1]{\prod_{i=2}^{n} t_i} = \sqrt[n-1]{\frac{y_n}{y_1}} \tag{6.12}$$

Chỉ tiêu này có ý nghĩa khi các tốc độ phát triển liên hoàn xấp xỉ nhau nghĩa là trong suốt kỳ nghiên cứu hiện tượng phát triển với một tốc độ tương đối đều.

Ví dụ 6.8. Tình hình doanh thu của công ty ABC được cho trong bảng số liệu sau :

Năm	2012	2013	2014	2015	2016	2017	2018
Doanh thu	35	38	45	48	52	58	65

Tính tốc độ phát triển liên hoàn, tốc độ phát triển định gốc và tốc độ phát triển bình quân về doanh thu của công ty ABC.

Giải

Gọi y_i là doanh thu năm thứ i hàng của công ty ABC

Tính tốc độ phát triển liên hoàn và tốc độ phát triển định gốc về doanh thu của công ty ABC, ta lập bảng sau :

Năm	2012	2013	2014	2015	2016	2017	2018
y_i	35	38	45	48	52	58	65
$t_i = \dfrac{y_i}{y_{i-1}}$		1,0857	1,1842	1,0667	1,0833	1,1154	1,1207
$T_i = \dfrac{y_i}{y_1}$		1,08576	1,2857	1,3714	1,4857	1,6571	1,8571

Tốc độ phát triển trung bình về doanh thu của công ty ABC.

$$\overline{t} = \sqrt[n-1]{\frac{y_n}{y_1}} = \sqrt[7-1]{\frac{65}{35}} = 1,1087.$$

6.1.2.4. Tốc độ tăng (giảm)

Tốc độ tăng (giảm) là chỉ tiêu phản ánh nhịp độ tăng (giảm) tương đối giữa các mức độ của hiện tượng qua thời gian. Nghĩa là, qua một hoặc một số đơn vị thời gian, hiện tượng đã tăng (giảm) bao nhiêu lần hoặc bao nhiêu phần trăm. Tuỳ theo mục đích nghiên cứu, có thể chọn kỳ gốc so sánh khác nhau, khi đó ta có các tốc độ tăng (giảm) sau:

a. Tốc độ tăng giảm liên hoàn: là chỉ tiêu phản ánh nhịp độ tăng (giảm) tương đối của hiện tượng giữa hai thời gian liền nhau và được tính theo công thức:

$$a_i = \frac{y_i - y_{i-1}}{y_{i-1}} = \frac{\delta_i}{y_{i-1}} = t_i - 1,\ i = 2,3,...,n. \tag{6.13}$$

b. Tốc độ tăng giảm định gốc: là chỉ tiêu phản ánh nhịp độ tăng (giảm) tương đối của hiện tượng giữa hai thời gian dài và thường lấy mức độ đầu tiên làm gốc cố định và được tính theo công thức:

$$A_i = \frac{y_i - y_1}{y_1} = \frac{\Delta_i}{y_1} = T_i - 1,\ i = 2,3,...,n. \tag{6.14}$$

c. Tốc độ tăng giảm bình quân: là chỉ tiêu phản ánh nhịp độ tăng (giảm) đại diện cho các tốc độ tăng (giảm) liên hoàn và được tính theo công thức

$$\overline{a} = \overline{t} - 1. \tag{6.15}$$

Ví dụ 6.9. Tiếp ví dụ 6.8. Tính tốc độ tăng giảm liên hoàn, tốc độ tăng giảm định gốc và tốc độ tăng giảm bình quân về doanh thu của công ty ABC.

Giải

Tính tốc độ tăng giảm liên hoàn và tốc độ tăng giảm định gốc về doanh thu của công ty ABC, ta lập bảng sau :

Năm	2012	2013	2014	2015	2016	2017	2018
y_i	35	38	45	48	52	58	65
$a_i = t_i - 1$		0,0857	0,1842	0,0667	0,0833	0,1154	0,1207
$A_i = T_i - 1$		0,08576	0,2857	0,3714	0,4857	0,6571	0,8571

Tốc độ tăng giảm bình quân về doanh thu của công ty ABC.

$$\overline{a} = \overline{t} - 1 = 0{,}1087.$$

6.1.2.5. Giá trị tuyệt đối của 1% của tốc độ tăng (giảm) liên hoàn

Giá trị tuyệt đối 1% của tốc độ tăng (giảm) liên hoàn là chỉ tiêu phản ánh cứ 1% của tốc độ tăng (giảm) liên hoàn thì tương ứng hiện tượng nghiên cứu tăng thêm (hoặc giảm đi) một lượng tuyệt đối cụ thể là bao nhiêu. Công thức tính:

$$g_i = \frac{\delta_i}{a_i(\%)} = \frac{y_i - y_{i-1}}{\frac{y_i - y_{i-1}}{y_{i-1}}100} = \frac{y_{i-1}}{100}. \tag{6.16}$$

Ví dụ 6.10. Tiếp ví dụ 6.8. Tính giá trị tuyệt đối 1% của tốc độ tăng (giảm) liên hoàn về doanh thu của công ty ABC.

Giải

Tính giá trị tuyệt đối 1% của tốc độ tăng (giảm) liên hoàn về doanh thu của công ty ABC, ta lập bảng sau :

Năm	2012	2013	2014	2015	2016	2017	2018
y_i	35	38	45	48	52	58	65
$g_i = \frac{y_{i-1}}{100}$		0,35	0,38	0,45	0,48	0,52	0,58

6.2. Hàm xu thế

Tiến trình thể hiện xu hướng bằng phương pháp hàm số được thực hiện qua 3 bước: nhận ra mô hình, lựa chọn mô hình và điều chỉnh mô hình.

Nội dung cơ bản của phương pháp hàm xu thế là khái quát hóa chiều hướng biến động của hiện tượng nghiên cứu bằng một hàm số toán học, nhằm mô tả một cách sát nhất, gần đúng nhất biến động thực tế của hiện tượng. Cụ thể là, thông qua việc xem xét đồ thị biến động thực tế của hiện tượng kết hợp với kinh nghiệm, sự hiểu biết thực tế về hiện tượng, ta chọn một hàm số có tính chất lý thuyết để thể hiện một cách tốt nhất xu hướng phát triển của hiện tượng.

Dưới đây là một số hàm thường được sử dụng

6.2.1. Hàm xu thế tuyến tính

Hàm xu thế tuyến tính được sử dụng khi các lượng tăng (hoặc giảm) tuyệt đối liên hoàn xấp xỉ bằng nhau. Dạng của hàm xu thế tuyến tính là

$$\hat{y}_t = a_0 + a_1 t \tag{6.17}$$

trong đó:

+) $\hat{y}_t$: Giá trị của hiện tượng tại thời gian t xác định bằng hàm số tuyến tính,

+) t : Thứ tự thời gian trong dãy số,

+) a_0, a_1 : Các tham số quy định vị trí của đường thẳng.

Theo phương pháp bình phương bé nhất, $\hat{y}_t$ "thích hợp nhất" đối với dãy số thực tế khi:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - a_0 - a_1 t_i\right)^2 \Rightarrow \min \tag{6.18}$$

Từ điều kiện này ta có hệ hai phương trình

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n} t_i + a_1\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i t_i \end{cases} \tag{6.19}$$

Giải hệ phương trình này ta tìm được a_0, a_1 cho hàm tuyến tính.

Trong thực tế, do t là thứ tự thời gian trong dãy số, nên ta có thể tìm a_0, a_1 đơn giản bằng cách đánh số thứ tự sao cho $\sum_{i=1}^{n} t_i = 0$.

+) Nếu thứ tự thời gian là một số lẻ, thì lấy thời gian ở vị trí giữa bằng 0, các thời gian đứng trước lần lượt là -1, -2, -3,… và thời gian đứng sau lần lượt là 1, 2, 3…

+) Nếu thứ tự thời gian là một số chẵn, thì lấy hai thời gian ở vị trí giữa bằng -1 và 1, các thời gian đứng trước lần lượt là -3, -5, -7,… và thời gian đứng sau lần lượt là 3, 5, 7…

Khi đó hệ phương trình trở nên đơn giản hơn

$$\begin{cases} na_0 = \sum_{i=1}^{n} y_i \\ a_1\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i t_i \end{cases} \Leftrightarrow \begin{cases} a_0 = \dfrac{1}{n}\sum_{i=1}^{n} y_i \\ a_1 = \dfrac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i^2} \end{cases} \quad (6.20)$$

Sau đây ta xét ví dụ về hàm xu thế tuyến tính để biểu hiện xu thế phát triển vốn đầu tư (triệu đồng) của Việt Nam giai đoạn 2000-2007.

Ví dụ 6.11. Số liệu thống kê vốn đầu tư khu vực kinh tế nhà nước của Việt Nam giai đoạn 2000 đến 2010 như sau: (đơn vị tính: tỷ đồng)

Năm	Vốn đầu tư	Năm	Vốn đầu tư
2001	101973	2006	185102
2002	114738	2007	197989
2003	126558	2008	209031
2004	139831	2009	287534
2005	161635	2010	316285

(nguồn: tổng cục thống kê)

Tìm hàm xu thế tuyến tính phản ánh xu hướng biến động của vốn đầu tư theo thời gian.

Giải

Gọi y vốn đầu tư, t là thời gian. Ta lập bảng như sau:

STT	t_i	y_i	t_i^2	$t_i y_i$
1	1	101973	1	101973
2	2	114738	4	229476
3	3	126558	9	379674
4	4	139831	16	559324
5	5	161635	25	808175

6	6	185102	36	1110612
7	7	197989	49	1385923
8	8	209031	64	1672248
9	9	287534	81	2587806
10	10	316285	100	3162850
Tổng	55	1840676	385	11998061

Thay các tổng vào (6.19), ta có hệ phương trình

$$\begin{cases} 10a_0 + 55a_1 = 1840676 \\ 55a_0 + 385a_1 = 11998061 \end{cases} \Leftrightarrow \begin{cases} a_0 = 59111,4 \\ a_1 = 22719,3091 \end{cases}$$

Vậy hàm xu thế cần tìm là

$$\hat{y} = 599111,4 + 22719,3091 \cdot t$$

Hàm xu thế trên cho biết vốn đầu tư khu vực kinh tế nhà nước giai đoạn 2001 – 2010 trung bình tăng mỗi năm là 22719,3 tỷ đồng.

6.2.2. Hàm số bậc 2 (Phương trình Parabol bậc 2)

Hàm xu thế parabol được sử dụng khi các sai phân bậc hai của dãy số xấp xỉ bằng nhau. Dạng tổng quát của hàm xu thế parabol như sau:

$$\hat{y}_t = a_0 + a_1 t + a_2 t^2 \qquad (6.21)$$

Các tham số a_0, a_1, a_2 có thể xác định thông qua hệ phương trình

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} t_i + a_2\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n} t_i + a_1\sum_{i=1}^{n} t_i^2 + a_2\sum_{i=1}^{n} t_i^3 = \sum_{i=1}^{n} y_i t_i \\ a_0\sum_{i=1}^{n} t_i^2 + a_1\sum_{i=1}^{n} t_i^3 + a_2\sum_{i=1}^{n} t_i^4 = \sum_{i=1}^{n} y_i t_i^2 \end{cases} \qquad (6.22)$$

Giải hệ phương trình trên ta tìm được a_0, a_1, a_2. Vì t là thứ tự thời gian nên ta cũng có thể tìm a_0, a_1, a_2 nhanh chóng bằng cách đánh số thứ tự sao cho $\sum_{i=1}^{n} t_i = 0$ và vì vậy $\sum_{i=1}^{n} t_i^3 = 0$ và hệ phương trình trở nên đơn giản hơn để xác định a_0, a_1, a_2 như sau:

$$\begin{cases} na_0 + a_2\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i \\ a_1\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i t_i \\ a_0\sum_{i=1}^{n} t_i^2 + a_2\sum_{i=1}^{n} t_i^4 = \sum_{i=1}^{n} y_i t_i^2 \end{cases} \qquad (6.23)$$

Ví dụ 6.12. Giả sử ta có tài liệu về sản lượng mặt hàng X (triệu mét) của một công ty qua các năm như sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014
Sản lượng	4,25	5,25	5,75	6,95	9,75	10,75	12,5	15,6

Tìm hàm xu thế parabol phản ánh xu hướng biến động của sản lượng theo thời gian.

Giải

Gọi y là sản lượng mặt hàng X, t là thời gian. Ta lập bảng như sau:

STT	t	t^2	t^3	t^4	$y \cdot t$	$y \cdot t^2$	y
1	1	1	1	1	4,25	4,25	4,25
2	2	40	8	16	10,50	21,00	5,25
3	3	9	27	81	17,25	51,75	5,75
4	4	16	64	256	27,8	111,20	6,95
5	5	25	125	625	48,75	243,75	9,75
6	6	36	216	1296	64,50	387,00	10,75
7	7	49	343	2401	87,50	612,50	12,50
8	8	64	512	4096	124,80	998,40	15,60
Tổng	36	204	1296	8772	385,35	2429,85	70,80

Thay các tổng vào (6.22), ta có hệ phương trình

$$\begin{cases} 8a_0 + 36a_1 + 204a_2 = 70{,}8 \\ 36a_0 + 204a_1 + 1296a_2 = 385{,}35 \\ 204a_0 + 1296a_1 + 8772a_2 = 2429{,}85 \end{cases} \Leftrightarrow \begin{cases} a_0 = 3{,}8143 \\ a_1 = 0{,}3196 \\ a_2 = 0{,}1411 \end{cases}$$

Vậy hàm xu thế cần tìm là

$$\hat{y} = 3{,}8143 + 0{,}3196 \cdot t + 0{,}1411 \cdot t^2.$$

6.2.3. Hàm số mũ

Hàm xu thế mũ được sử dụng khi các tốc độ phát triển liên hoàn xấp xỉ bằng nhau. Dạng tổng quát của hàm xu thế mũ là

Hàm số có dạng :

$$\hat{y}_t = a_0 a_1^t \tag{6.24}$$

trong đó

+) a_0 : Điểm gốc của phương trình hồi quy,

+) a_1 : Tốc độ phát triển trung bình theo đơn vị thời gian (số lần).

Lấy logarit cơ số e hai vế của phương trình (6.24) và áp dụng (6.19), ta có hệ phương trình

$$\begin{cases} n\ln(a_0) + \ln(a_1)\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \ln(y_i) \\ \ln(a_0)\sum_{i=1}^{n} t_i + \ln(a_1)\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} t_i \ln(y_i) \end{cases} \tag{6.25}$$

Để đơn giản ta cũng có thể đánh số thứ tự sao cho $\sum_{i=1}^{n} t_i = 0$.

Ví dụ 6.13. Giả sử ta có tài liệu về giá trị sản xuất (GTSX) của một công ty qua các năm được cho trong bảng sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GTSX	2,7	3,5	4,5	5,8	7,4	9,5	11,5	13,6	15,7	18,9

Tìm hàm xu thế mũ phản ánh xu hướng biến động của GTSX theo thời gian.

Giải

Gọi y là giá trị sản xuất của công ty, t là thời gian. Ta lập bảng như sau:

STT	t	t^2	log(y)	$t\cdot\log(y)$
1	1	1	0,9933	0,9933
2	2	4	1,2528	2,5055
3	3	9	1,5041	4,5122
4	4	16	1,7579	7,0314
5	5	25	2,0015	10,0074
6	6	36	2,2513	13,5078
7	7	49	2,4423	17,0964
8	8	64	2,6101	20,88056

9	9	81	2,7537	24,7830
10	10	100	2,9392	29,3916
Tổng	55	385	20,506	103,71

Thay các tổng vào (6.25), ta có hệ phương trình

$$\begin{cases} 10\ln(a_0) & + & 55\ln(a_1) & = & 20,506 \\ 55\ln(a_0) & + & 385\ln(a_1) & = & 130,71 \end{cases}$$

$$\Leftrightarrow \begin{cases} \ln(a_0) = 0,8555 \\ \ln(a_1) = 0,2173 \end{cases} \Leftrightarrow \begin{cases} a_0 = 2,353 \\ a_1 = 1,243 \end{cases}$$

Vậy hàm xu thế cần tìm là

$$\hat{y}_t = (\ 2,353) \cdot (1,243)^t.$$

6.2.4. Hàm hypebol

Hàm xu thế hypebol được vận dụng khi các mức độ của hiện tượng giảm dần theo thời gian. Dạng tổng quát của hàm xu thế hypebol là

$$\hat{y} = a_0 + \frac{a_1}{t} \tag{6.26}$$

Để xác định hai tham số a_0, a_1, áp dụng (6.19) xem t chính là $\frac{1}{t}$, ta có hệ phương trình

$$\begin{cases} na_0 & + & a_1\sum_{i=1}^{n}\frac{1}{t_i} & = & \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n}\frac{1}{t_i} & + & a_1\sum_{i=1}^{n}\frac{1}{t_i^2} & = & \sum_{i=1}^{n}\frac{y_i}{t_i} \end{cases} \tag{6.27}$$

Trong thực tế, ta còn có thể xây dựng nhiều hàm số khác nhau như hàm bậc 3, hàm hypebol… tùy theo đặc điểm và tính chất biến động của hiện tượng.

Phương pháp dự báo bằng hàm xu thế là khái quát hóa chiều hướng biến động của hiện tượng nghiên cứu bằng một hàm số toán học, nhằm mô tả một cách sát nhất, gần đúng nhất biến động thực tế của hiện tượng. Tùy theo tính chất của hiện tượng kết hợp với kinh nghiệm, sự hiểu biết thực tế về hiện tượng, ta chọn một hàm số có tính chất lý thuyết để thể hiện một cách tốt nhất xu hướng phát triển của hiện tượng.

Ví dụ 6.14. Giả sử ta có tài liệu về sản lượng của một công ty qua các năm được cho trong bảng sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016

Sản lượng	2,5	1,95	1,85	1,75	1,64	1,61	1,59	1,57	1,55	1,51

Tìm hàm xu thế hypebol phản ánh xu hướng biến động của sản lượng theo thời gian.

Giải

Gọi y là sản lượng của công ty, t là thời gian. Ta lập bảng như sau:

STT	t	$1/t$	$1/t^2$	y	y/t
1	1	1	1	2,5	2,5
2	2	1/2	1/4	1,95	0,975
3	3	1/3	1/9	1,85	0,6167
4	4	1/4	1/16	1,75	0,4375
5	5	1/5	1/25	1,64	0,3280
6	6	1/6	1/36	1,61	0,2683
7	7	1/7	1/49	1,59	0,2271
8	8	1/8	1/64	1,57	0,1963
9	9	1/9	1/81	1,55	0,1722
10	10	1/10	1/100	1,51	0,1510
Tổng	55	2,929	1,55	17,52	5,872

Thay các tổng vào (6.27), ta có hệ phương trình

$$\begin{cases} 10a_0 + 2{,}929a_1 = 17{,}52 \\ 2{,}929a_0 + 1{,}55a_1 = 5{,}872 \end{cases} \Leftrightarrow \begin{cases} a_0 = 1{,}44 \\ a_1 = 1{,}07 \end{cases}$$

Vậy hàm xu thế cần tìm là

$$\hat{y} = 1{,}44 + \frac{1{,}07}{t}.$$

6.3. Dự báo theo dãy số thời gian

Theo thời gian dự báo có thể chia thành ngắn hạn, trung hạn và dài hạn. Sau đây là một số phương pháp dự báo thống kê thường sử dụng dựa vào dãy số thời gian.

6.3.1. Dự báo dựa vào lượng tăng (giảm) tuyệt đối trung bình

Phương pháp này được sử dụng khi biến động của hiện tượng có lượng tăng (giảm) tuyệt đối liên hoàn xấp xỉ nhau.

$$\hat{y}_{n+L} = y_n + L \cdot \bar{\delta} \qquad (6.28)$$

trong đó:

+) $\hat{y}_{n+L}$: Giá trị dự đoán ở thời gian $(n+L)$,

+) y_n : Giá trị thực tế ở thời gian n,

+) $\overline{\delta}$: Lượng tăng (giảm) tuyệt đối trung bình,

+) L : Tầm xa dự đoán.

Ví dụ 6.15. Xem số liệu về kết quả hoạt động kinh doanh của doanh nghiệp A như sau

Năm	2008	2009	2010	2011	2012
Lợi nhuận sau thuế (tỷ đồng)	29,08	38,39	40,7	58,79	70,55

Qua bảng kết quả lợi nhuận cho thấy lượng tăng tuyệt đối liên hoàn của lợi nhuận xấp xỉ bằng nhau, áp dụng mô hình dự báo trên lợi nhuận của doanh nghiệp năm 2013 và 2014, ta có

$$\overline{\delta} = \frac{1}{n-1}\sum_{i=2}^{n}\delta_i = \frac{41,47}{5-1} = 10,37$$

Dự báo lợi nhuận năm 2013 :

$$\hat{y}_{2013} = y_{2012} + 1\cdot\overline{\delta} = 70,55 + 1\cdot 10,37 = 80,92 \text{ (tỷ).}$$

Dự báo lợi nhuận năm 2014:

$$\hat{y}_{2014} = y_{2012} + 2\cdot\overline{\delta} = 70,55 + 2\cdot 10,37 = 91,29 \text{ (tỷ).}$$

6.3.2. Dự báo dựa vào tốc độ phát triển bình quân

Phương pháp này được sử dụng khi hiện tượng nghiên cứu biến động với một nhịp độ tương đối ổn định, tức là tốc độ phát triển liên hoàn xấp xỉ bằng nhau.

$$\hat{y}_{n+L} = y_n\left(\overline{t}\right)^L \qquad (6.29)$$

trong đó

+) $\hat{y}_{n+L}$: Giá trị dự đoán ở thời gian $(n+L)$,

+) y_n : Giá trị thực tế ở thời gian n,

+) $\overline{t}$: Tốc độ phát triển bình quân,

+) L : Tầm xa dự đoán.

Ví dụ 6.16. Sử dụng bảng số liệu ví dụ 15, dự báo lợi nhuận năm 2013 và 2014 dựa vào tốc độ phát triển bình quân.

Ta có

$$\overline{t} = \sqrt[n-1]{\frac{y_n}{y_1}} = \sqrt[5-1]{\frac{70,55}{29,08}} = 1,24$$

Dự báo lợi nhuận năm 2013:

$$\hat{y}_{2013} = y_{2012}\left(\bar{t}\right)^1 = 70,55\cdot(1,24)^1 = 87,482 \text{ (tỷ)}$$

Dự báo lợi nhuận năm 2014:

$$\hat{y}_{2014} = y_{2012}\cdot\left(\bar{t}\right)^2 = 70,55\cdot(1,24)^2 = 108,4 \text{ (tỷ)}$$

6.3.3. Dự báo dựa vào hàm xu thế tuyến tính

Dựa trên cơ sở chiều hướng biến động của hiện tượng nghiên cứu được khái quát hóa bằng một hàm số tuyến tính có dạng:

$$\hat{y}_t = a_0 + a_1 t \tag{6.30}$$

trong đó

+) $\hat{y}_t$: Giá trị của hiện tượng tại thời gian t xác định bằng hàm số tuyến tính,

+) t : Thứ tự thời gian.

+) a_0, a_1 : Các tham số quy định vị trí của đường thẳng.

Tổng quát: $\hat{y}_t = f(t)$

Do đó:

$$\hat{y}_{n+L} = f\left(n+L\right). \tag{6.31}$$

Trong đó: $\hat{y}_{n+L}$: Giá trị dự đoán ở thời gian $(n+L)$, L: tầm xa dự đoán.

Ví dụ 6.17. Sử dụng bảng số liệu ví dụ 15, dự báo nhuận năm 2013 và 2014 bằng hàm xu thế tuyến tính.

Hàm xu thế tuyến tính

$$\hat{y}_t = 16,5 + 10,334t$$

Dự báo lợi nhuận năm 2013:

$$\hat{y}_{2013} = 16,5 + 10,334\cdot(5+1) = 78,5 \text{ (tỷ).}$$

Dự báo lợi nhuận năm 2014:

$$\hat{y}_{2014} = 16,5 + 10,334\cdot(5+2) = 88,8 \text{ (tỷ).}$$

Nhận xét: Kết quả dự báo theo 3 phương pháp là khác nhau, trong đó phương pháp dự báo dựa vào lượng tăng giảm tuyệt đối bình quân là phù hợp nhất. Do đó, điều kiện áp dụng thỏa mãn. Như vậy, tùy theo từng loại số liệu cụ thể khi phân tích có thể chọn phương pháp dự báo sao cho phù hợp với điều kiện áp dụng nhất.

6.4. Tóm tắt chương 6

1. Các chỉ tiêu phân tích dãy số thời gian

i) Mức độ trung bình theo thời gian: Giả sử ta có dãy số thời gian: $y_1, y_2, ..., y_n$

+) Dãy số thời kỳ: Đối với dãy số thời kỳ, mức độ bình quân theo thời gian được tính theo công thức:

$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

+) Dãy số thời điểm: có 2 trường hợp

- Đối với dãy số thời điểm có các khoảng cách thời gian bằng nhau:

$$\overline{y} = \frac{\frac{1}{2}y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{y_i + y_{i+1}}{2}\right).$$

- Đối với dãy số thời điểm có các khoảng cách thời gian không bằng

$$\overline{y} = \frac{t_1 y_1 + t_2 y_2 + \cdots + t_n y_n}{t_1 + t_2 + \cdots + t_n} = \frac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i}.$$

ii) Lượng tăng (giảm) tuyệt đối

+) Lượng tăng (giảm) tuyệt đối liên hoàn : $\delta_i = y_i - y_{i-1}$, $i = 2, 3, ..., n$.

+) Lượng tăng (giảm) tuyệt đối định gốc: $\Delta_i = y_i - y_1$, $i = 2, 3, ..., n$.

+) Mối liên hệ giữa lượng tăng (giảm) liên hoàn và lượng tăng (giảm) định gốc

$$\sum_{i=2}^{n} \delta_i = \Delta_n.$$

+) Lượng tăng (giảm) tuyệt đối bình quân: $\overline{\delta} = \dfrac{y_n - y_1}{n-1}$.

iii) Tốc độ phát triển

+) Tốc độ phát triển liên hoàn: $t_i = \dfrac{y_i}{y_{i-1}}$, $i = 2, 3, ..., n$.

+) Tốc độ phát triển định gốc: $T_i = \dfrac{y_i}{y_1}$, $i = 2, 3, ..., n$.

+) Mối liên hệ giữa tốc độ phát triển liên hoàn và tốc độ phát triển định gốc

$$\prod_{i=2}^{n} t_i = T_n.$$

+) Tốc độ phát triển bình quân: $\overline{t} = \sqrt[n-1]{\frac{y_n}{y_1}}$.

iv) Tốc độ tăng (giảm)

+) Tốc độ tăng giảm liên hoàn: $a_i = t_i - 1,\ i = 2,3,...,n.$

+) Tốc độ tăng giảm định gốc: $A_i = T_i - 1,\ i = 2,3,...,n.$

+) Tốc độ tăng giảm bình quân: $\overline{a} = \overline{t} - 1.$

v) Giá trị tuyệt đối của 1% của tốc độ tăng (giảm) liên hoàn: $g_i = \frac{y_{i-1}}{100}$.

2. Hàm xu thế

i) Hàm xu thế tuyến tính: $\hat{y}_t = a_0 + a_1 t$

Các tham số a_0, a_1 có thể xác định thông qua hệ phương trình

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n} t_i + a_1\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i t_i \end{cases}.$$

ii) Hàm số bậc 2: $\hat{y}_t = a_0 + a_1 t + a_2 t^2$

Các tham số a_0, a_1, a_2 có thể xác định thông qua hệ phương trình

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} t_i + a_2\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n} t_i + a_1\sum_{i=1}^{n} t_i^2 + a_2\sum_{i=1}^{n} t_i^3 = \sum_{i=1}^{n} y_i t_i \\ a_0\sum_{i=1}^{n} t_i^2 + a_1\sum_{i=1}^{n} t_i^3 + a_2\sum_{i=1}^{n} t_i^4 = \sum_{i=1}^{n} y_i t_i^2 \end{cases}.$$

iii) Hàm số mũ: $\hat{y}_t = a_0 a_1^t$

Các tham số a_0, a_1 có thể xác định thông qua hệ phương trình

$$\begin{cases} n\ln(a_0) + \ln(a_1)\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \ln(y_i) \\ \ln(a_0)\sum_{i=1}^{n} t_i + \ln(a_1)\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} t_i \ln(y_i) \end{cases}.$$

iv) Hàm hypebol: $\hat{y} = a_0 + \frac{a_1}{t}$

Các tham số a_0, a_1 có thể xác định thông qua hệ phương trình

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n}\frac{1}{t_i} = \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n}\frac{1}{t_i} + a_1\sum_{i=1}^{n}\frac{1}{t_i^2} = \sum_{i=1}^{n}\frac{y_i}{t_i} \end{cases}.$$

3. Dự báo theo dãy số thời gian

+) Dự báo dựa vào lượng tăng (giảm) tuyệt đối trung bình: $\hat{y}_{n+L} = y_n + L\cdot\bar{\delta}$.

+) Dự báo dựa vào tốc độ phát triển bình quân: $\hat{y}_{n+L} = y_n\left(\bar{t}\right)^L$.

+) Dự báo dựa vào hàm xu thế tuyến tính: $\hat{y}_{n+L} = a_0 + a_1\left(N+L\right)$.

6.5. Bài tập

Bài số 1. Có số liệu về hàng tồn kho của một công ty thương mại trong 6 tháng đầu năm 2008 như sau:

Chỉ tiêu \ Thời gian	1/1	1/2	1/3	1/4	1/5	1/6	1/7
GT hàng hóa tồn kho (tr.đ)	120	122	126	128	130	140	148

Hãy tính giá trị hàng hóa tồn kho bình quân của công ty trong các thời kỳ sau: từng tháng, từng quý, 6 tháng đầu năm.

Đáp số: Theo tháng: $\bar{y}_1 = 121$, $\bar{y}_2 = 124$, $\bar{y}_3 = 127$, $\bar{y}_4 = 129$, $\bar{y}_5 = 135$, $\bar{y}_6 = 144$

Theo quý: $\bar{Q}_1 = 124$, $\bar{Q}_2 = 136$. *Theo 6 tháng đầu năm:* 130.

Bài số 2. Số công nhân trong danh sách của 1 XN năm báo cáo như sau:

Ngày 01/01 xí nghiệp có 425 công nhân

26/01 bổ sung thêm 5 công nhân

22/03 cho thôi việc 2 công nhân

10/04 bổ sung thêm 3 công nhân

14/06 chuyển sang cơ quan khác 3 công nhân

17/08 cho thôi việc 2 công nhân

20/10 bổ sung thêm 6 công nhân

Từ đó đến cuối năm số công nhân của xí nghiệp không thay đổi. Hãy tính số công nhân trong danh sách bình quân cả năm của xí nghiệp. Biết rằng trong năm này tháng 02 có 29 ngày.

Đáp số: 429,082.

Bài số 3. Tình hình biến động giá thành sản phẩm của một xí nghiệp qua các năm như sau:

Năm 2003 so với năm 2002 giảm 3,0%

Năm 2004 so với năm 2003 giảm 2,5%

Năm 2005 so với năm 2004 tăng 2,7%

Năm 2006 so với năm 2005 giảm 3,2%

Năm 2007 so với năm 2006 giảm 2,6%

1. Hãy xây dựng một dãy số thời gian nói lên biến động về giá thành sản phẩm của xí nghiệp trong thời kỳ nói trên (lấy năm 2002 là 100%).

2. Tính tốc độ giảm giá thành bình quân hàng năm của xí nghiệp trong thời kỳ nói trên.

Đáp số 1.

2002	*2003*	*2004*	*2005*	*2006*	*2007*
100	*97*	*94,575*	*97,13*	*94,02*	*91,58*

2. 0,983.

Bài số 4. Tốc độ phát triển về giá trị tổng sản lượng của hai xí nghiệp như sau (%)

Xí nghiệp	Năm 2005 so với 2004	Năm 2006 so với 2005	Năm 2007 so với 2006
A	105	110	118
B	107	115	120

Hãy tính:

1. Tốc độ phát triển năm 2007 so với năm 2004 của mỗi xí nghiệp trong thời gian trên?

2. Tốc độ phát triển bình quân hàng năm của mỗi xí nghiệp trong thời gian trên?

3. Tốc độ phát triển năm 2007 so với năm 2004 tính chung cho cả 2 xí nghiệp ? Biết thêm rằng giá trị tổng sản lượng của xí nghiệp A năm 2004 là 45.000 triệu đồng và của xí nghiệp B là 80.000 triệu đồng.

Đáp số: 1. 1,3629 và 1,4766; 2. 1,108 và 1,139; 3. 1,4357.

Bài số 5. Có số liệu ở một xí nghiệp như sau:

Năm	Sản	Biến động so với năm trước (%)

	lượng (tỷ đồng)	Lượng tăng (giảm) tuyệt đối (tỷ đồng)	Tốc độ phát triển (%)	Tốc độ tăng (%)	Giá trị tuyệt đối 1% tăng lên (triệu đồng)
2002	18,6				
2003			112,4		
2004				6,7	
2005		2,3			
2006					
2007				12,5	271

1. Hãy tính các số liệu còn thiếu trong bảng thống kê?

2. Tính tốc độ phát triển bình quân về giá trị sản lượng của xí nghiệp trong thời kì trên?

3. Dùng các phương pháp dự báo sản lượng năm 2008?

Đáp số: 2. 1,104; 3. 32,86; 33,658; 32,033.

Bài số 6. Có số liệu về sản lượng lúa (y: đơn vị: trăm ngàn tấn) qua các năm như sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
y	35,94	38,73	38,95	40,01	42,4	43,74	44,04	44,98	45,1	43,2

1. Tính các chỉ tiêu δ_i, t_i, g_i để phân tích sự biến động sản lượng lúa theo thời gian?

2. Xác định các chỉ tiêu: Tốc độ phát triển bình quân; lượng tăng (giảm) tuyệt đối bình quân về sản lượng lúa?

3. Dự báo sản lượng lúa năm 2017 theo các phương pháp: Lượng tăng (giảm) tuyệt đối bình quân; tốc độ phát triển bình quân; hàm xu thế?

Đáp số: 1. $\delta_i = y_i - y_{i-1}$; $t_i = \dfrac{y_i}{y_{i-1}}$; $g_i = \dfrac{y_{i-1}}{100}$;

2. $\overline{t} = 1{,}021$; $\overline{\delta} = 0{,}807$; *3.* $44{,}007$; $44{,}1072$; $46{,}8264$.

Bài số 7. Có tài liệu về tổng sản lượng lúa (y: ngàn tấn) của một địa phương như sau:

Năm	2009	2010	2011	2012	2013	2014	2015	2016
y	38950	40005	42398	43737	44039	44947	45106	43610

Yêu cầu:

1. Tính các chỉ tiêu phân tích sự biến động của sản lượng lúa theo thời gian?

2. Tìm hàm xu thế, biểu thị biến động sản lượng lúa theo thời gian.

3. Dự đoán sản lượng lúa vào các năm 2017, 2018 theo các phương pháp đã học?

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}.$

2. $\hat{y}_t = 39309,3214 + 786,5952 \cdot t.$

3. Năm 2017 : $44275,714; 44320,84; 46388,678.$

Năm 2018 : $44941,42; 45043,273; 47175,2734.$

Bài số 8. Doanh số bán (y) của công ty thương mại X ghi nhận được như sau:

Năm	1998	1999	2000	2001	2002	2003	2004	2005
y	35	50	66	58	74	90	107,5	121

1. Tính các chỉ tiêu động thái phân tích sự biến động của doanh số bán theo thời gian?

2. Tìm hàm xu thế, thể hiện sự biến động doanh số bán theo thời gian?

3. Dự đoán doanh số bán các năm 2007, 2008, 2009 theo các phương pháp đã học?

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}.$

2. $\hat{y}_t = 22,8214 + 11,6369 \cdot t.$

3. Năm 2007 : $145,5714; 172,4731; 139,1904.$

Năm 2008 : $157,8571; 205,9156; 150,8273.$

Năm 2009 : $170,1428; 245,8426; 162,4642.$

Bài số 9. Có tài liệu về sản lượng mặt hàng X (triệu mét) của một công ty qua các năm như sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014
Sản lượng	5,01	5,28	5,56	5,8	6,3	6,51	6,75	6,97

1. Tính các chỉ tiêu động thái phân tích sự biến động của sản lượng mặt hàng X theo thời gian?

2. Tìm hàm xu thế biểu thị biến động sản lượng theo thời gian?

3. Dự đoán sản lượng các năm 2016, 2017, 2018?

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}.$

2. $\hat{y}_t = 4,7143 + 0,2907 \cdot t.$

3. Năm 2016 : $7,53; 7,66; 7,62.$

Năm 2017 : $7,81; 8,03; 7,91.$

Năm 2018 : $8,09; 8,42; 8,20.$

Bài số 10. Có tài liệu về doanh thu bán hàng của cửa hàng bách hoá như sau:

Năm	2004	2005	2006	2007	2008
Doanh thu bán hàng (triệu đồng)	7510	7680	8050	8380	8500

Yêu cầu:

1. Tính các chỉ tiêu động thái phân tích sự biến động của doanh thu bán hàng theo thời gian?
2. Tìm hàm xu thế, biểu thị biến động doanh thu theo thời gian?
3. Dự đoán doanh thu bán hàng các năm 2010, 2011, 2012?

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}.$

2. $\hat{y}_t = 7220 + 268 \cdot t.$

3. Năm 2010 : $8995; 9041,18; 9096.$

Năm 2011 : $9242,5; 9326,1; 9364.$

Năm 2012 : $9490; 9618,95; 9632.$

Bài số 11. Có tài liệu về sản lượng của một sản phẩm tại xí nghiệp X như sau:

Năm	1998	1999	2000	2001	2002	2003	2004	2005
Sản lượng (1000 tấn)	35,5	36,8	53,8	45,4	56,7	48,6	46,5	54,2

1. Tính các chỉ tiêu phân tích sự biến động của sản lượng theo thời gian?
2. Tìm hàm xu thế thể hiện sự biến động sản lượng theo thời gian?
3. Dự đoán sản lượng các năm 2008, 2009, 2010?

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}.$

2. $\hat{y}_t = 37,8071 + 2,0845 \cdot t.$

3. Năm 2008 : $62,21; 64,97; 60,74.$

Năm 2009 : $64,89; 69,02; 62,82.$

Năm 2010 : $67,56; 73,32; 64,91.$

Bài số 12. Cho biết tình hình về tỷ lệ tăng trưởng GDP (%) của Việt Nam như sau:

1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
8,07	8,84	9,54	9,34	8,15	5,76	4,77	6,79	6,19	6,32	6,90

Lưu ý: Chỉ dùng số liệu đến năm 2002.

1. Tính các chỉ tiêu sau: tốc độ phát triển liên hoàn mỗi năm t_i, lượng tăng (giảm) tuyệt đối liên hoàn mỗi năm δ_i, tốc độ phát triển bình quân, lượng tăng (giảm) tuyệt đối bình quân.

2. Dùng các phương pháp: tốc độ phát triển bình quân, lượng tăng (giảm) tuyệt đối bình quân và đường xu thế. Hãy dự báo lệ tăng trưởng GDP các năm 2003, 2004, 2005. Nhận xét kết quả.

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}. \overline{\delta} = -0{,}1944;\ \overline{t} = 0{,}9732.$

2. Năm 2003 : $6{,}13; 6{,}15; 5{,}24.$ *Năm 2004 :* $5{,}93; 5{,}99; 4{,}85.$

Năm 2005 : $5{,}74; 5{,}83; 4{,}46.$

Bài số 13. Cho biết tình hình về tỷ lệ tăng trưởng GDP (%) của Việt Nam như sau:

2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
7,536	7,547	6,978	7,130	5,662	5,398	6,423	6,240	5,247	5,422

Lưu ý: Chỉ dùng số liệu đến năm 2012.

1. Tính các chỉ tiêu sau: tốc độ phát triển liên hoàn mỗi năm t_i, lượng tăng (giảm) tuyệt đối liên hoàn mỗi năm δ_i, tốc độ phát triển bình quân, lượng tăng (giảm) tuyệt đối bình quân

2. Dùng các phương pháp: tốc độ phát triển bình quân, lượng tăng (giảm) tuyệt đối bình quân và đường xu thế. Hãy dự báo tỷ lệ tăng trưởng GDP của Việt Nam ở các năm 2013, 2014, 2015.

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}. \overline{\delta} = -0{,}2861; \overline{t} = 0{,}9558.$

2. Năm 2013 : $4{,}96; 5{,}02; 5{,}14.$ *Năm 2014 :* $4{,}67; 4{,}79; 4{,}87.$

Năm 2015 : $4{,}39; 4{,}58; 4{,}61.$

Bài số 14. Cho biết tình hình về tỷ lệ tăng trưởng GDP (%) của Việt Nam như sau:

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
6,899	7,536	7,547	6,978	7,130	5,662	5,398	6,423	6,240	5,247

Lưu ý: Chỉ dùng số liệu đến năm 2011.

1. Tính các chỉ tiêu sau: tốc độ phát triển liên hoàn mỗi năm t_i, lượng tăng (giảm) tuyệt đối liên hoàn mỗi năm δ_i, tốc độ phát triển bình quân, lượng tăng (giảm) tuyệt đối bình quân

2. Dùng các phương pháp: tốc độ phát triển bình quân, lượng tăng (giảm) tuyệt đối bình quân và đường xu thế. Hãy dự báo lệ tăng trưởng GDP của Việt Nam ở các năm 2012, 2013, 2014.

Đáp số: 1. $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}. \overline{\delta} = -0,0811; \overline{t} = 0,9877.$

2. Năm 2012 : $6,159; 6,163; 5,68.$ *Năm 2013 :* $6,08; 6,09; 5,49.$

Năm 2014 : $6,0; 6,013; 5,30.$

Bài số 15. Dữ liệu về GDP (tỷ USD) của Việt Nam qua các năm được cho như sau:

Năm	2008	2009	2010	2011	2012	2013	2014	2015
GDP	99,13	106	115,9	---	155,8	171,2	186,2	193,6

Yêu cầu:

1. Dự báo GDP của Việt Nam năm 2011 bằng phương pháp tốc độ phát triển bình quân $(\overline{t})$.

2. Tìm GDP mỗi năm, tốc độ phát triển mỗi năm, tốc độ tăng (giảm) mỗi năm, lượng tăng (giảm) tuyệt đối mỗi năm.

3. Dùng các phương pháp sau: Đường xu thế, tốc độ phát triển bình quân, lượng tăng (giảm) tuyệt đối bình quân. Hãy dự báo GDP của Việt Nam các năm 2016, 2017.

Đáp số: 1. $y_{2011} = 125,323;$ *2.* $\delta_i = y_i - y_{i-1}; t_i = \dfrac{y_i}{y_{i-1}}; g_i = \dfrac{y_{i-1}}{100}; a_i = t_i - 1.$

2. Năm 2016 : $207,1; 213,02; 211,57.$ *Năm 2017 :* $220,59; 234,38; 226,56.$

Bài số 16. Có tài liệu về sản lượng sản xuất một loại sản phẩm của một doanh nghiệp như sau

Năm	2008	2009	2010	2011	2012
Sản lượng (1000 tấn)	50	54	60	65	70

Yêu cầu: Dự đoán sản lượng của doanh nghiệp vào năm 2013 dựa vào:

1. Lượng tăng (giảm) tuyệt đối bình quân.

2. Tốc độ phát triển bình quân.

3. Ngoại suy hàm xu thế tuyến tính.

4. Trong các phương pháp dự đoán trên, phương pháp nào cho kết quả tốt nhất?

Bài số 17. Tốc độ phát triển về doanh thu về du lịch của một địa phương năm 2007 so với năm 2002 bằng 2,2 lần, năm 2012 so với năm 2007 doanh thu bằng 4,4 lần. Hãy tính tốc độ phát triển về doanh thu bình quân hàng năm giai đoạn từ năm 2003 – 2012.

Bài số 18. Có tài liệu về giá trị hàng hóa dự trữ của một công ty trong quý III/2012 như sau:

- Ngày 1/7, giá trị dự trữ là 850 triệu đồng.
- Ngày 30/7, giá trị dự trữ là 980 triệu đồng.
- Ngày 31/8, giá trị dự trữ là là 870 triệu đồng.
- Ngày 5/9, dự trữ thêm 200 triệu đồng.
- Ngày 18/9, xuất dự trữ 250 triệu đồng.
- Ngày 25/9, dự trữ thêm 100 triệu đồng.

Yêu cầu:

1. Tính giá trị hàng hóa dự trữ bình quân của từng tháng trong quý III/2012.
2. Tính giá trị hàng hóa dự trữ bình quân của quý III/2012.

6.6. Tài liệu tham khảo

[1] Hà văn Sơn, Giáo trình Lý thuyết Thống kê, ứng dụng trong Quản trị và kinh tế.

[2] Phạm Văn Chững, Lê Thanh Hoa, Nguyễn Đình Uông, Thống kê ứng dụng, NXB Đại học Quốc gia Thành Phố Hồ Chí Minh, 2016.

[3] S.P. Gordon, Contemporary Statistics, Mc Graw – Hill, Inc. 1994sworth.

[4] Anderson, Sweeney, and William [2010], Statistics for Business and Economics, South-Western Cengage Learning (11^{th} Edition).

[5] Michael Barrow, Statistics for Economics, Accounting and Business Studies-Prentice Hall, 2006.

[6] Newbold Paul - Statistics for Bussiness and Economics, 5th edition - Prentice Hall, 2005.

Thuật ngữ chính chương 6	
Tiếng Anh	**Tiếng Việt**
Analysis time series	Phân tích dãy số thời gian
Absolute amount of increase (decrease)	Lượng tăng (giảm) tuyệt đối
Average growth rate	Tốc độ phát triển bình quân
Cyclical pattern	Mô hình tuần hoàn
Exponential trend equation	Phương trình xu thế mũ
Estimating the linear trend	Ước lượng hàm xu thế tuyến tính
Forecast error	Sai số dự báo
Forecast accuracy	Độ chính xác của dự báo
Forecast by time series	Dự báo theo dãy số thời gian
Forecasting methods	Các phương pháp dự báo
Intercept of the linear trend line	Hệ số chặc của đường xu thế tuyến tính
Hypebol trend equation	Phương trình xu thế hypebol
Mean absolute error	Sai số trung bình dự báo
Mean squared error	Trung bình bình phương sai số
Moving averages	Trung bình di động
Linear trend equation	Phương trình xu thế tuyến tính
Growth rate	Nhịp tăng trưởng
Quadratic trend equation	Phương trình xu thế bậc 2
Time series	Dãy số thời gian
Trend pattern	Mô hình xu hướng
Time period	Khoảng thời gian
Trend function	Hàm xu thế
Time series decompostition	Phân rã chuỗi thời gian
Stationary time series	Dãy số thời gian dừng
Seasonal pattern	Mô hình theo mùa vụ
Seasonality and trend	Tính thời vụ và xu thế
Smoothing constant	Làm mịn hằng số
Slope of the linear trend line	Độ dốc của đường xu thế tuyến tính
Weighted moving averages	Trung bình di động có trọng số

MỘT SỐ ĐỀ THAM KHẢO

Đề số 1

Câu 1. Một kho hàng có 15 hộp sản phẩm được đóng gói giống nhau bao gồm 2 loại: loại I và loại II. Trong đó có 10 hộp loại I, mỗi hộp chứa 8 chính phẩm và 2 phế phẩm; 5 hộp loại II, mỗi hộp chứa 7 chính phẩm và 3 phế phẩm.

a) Lấy ngẫu nhiên hai hộp sản phẩm từ kho hàng, tính xác suất lấy được hai hộp cùng loại.

b) Lấy ngẫu nhiên một hộp rồi từ hộp đó lấy ngẫu nhiên 2 sản phẩm. Tính xác suất cả 2 sản phẩm lấy ra đều là chính phẩm.

Câu 2. Một người chăn nuôi có nuôi 400 con gà đẻ trứng cùng một loại. Xác suất để một con gà đẻ trứng trong ngày là 90%.

a) Tìm xác suất để người nuôi có được ít nhất 350 trứng trong ngày.

b) Nếu mỗi trứng gà bán được 3000 đồng và tiền nuôi dưỡng mỗi con gà trong 1 ngày là 1500 đồng thì hãy tính số tiền trung bình mà người nuôi có được trong ngày.

Câu 3. Một cuộc điều tra về chiều cao của một loại cây giống một năm tuổi. Lấy ngẫu nhiên 100 cây và tiến hành đo chiều cao, kết quả được cho trong bảng sau:

Chiều cao (cm)	140 − 150	150 − 160	160 − 170	170 − 180	180 − 190
Số cây	10	25	40	20	5

Giả sử chiều cao của một loại cây giống một năm tuổi tuân theo quy luật phân phối chuẩn.

a) Hãy ước lượng chiều cao trung bình của cây giống một năm tuổi với độ tin cậy 95%.

b) Dựa vào mẫu quan sát trên, với mức ý nghĩa 1% có thể kết luận rằng chiều cao trung bình của cây giống 1 năm tuổi lớn hơn 160 cm được không?

Câu 4. Hệ thống bán vé máy bay online của công ty hàng không AP vừa được cải tiến quy trình và được theo dõi để ghi nhận tình trạng hủy vé sau khi đặt chỗ. Khảo sát ngẫu nhiên một số ngày và nhận thấy trong 169 lần đặt vé thì có 15 lần hủy vé.

a) Với độ tin cậy 98%, hãy ước lượng tỷ lệ hủy vé sau khi đặt chỗ qua hệ thống?

b) Một tài liệu có được trước khi cải tiến hệ thống cho biết tỷ lệ hủy vé sau khi đặt chỗ là 15%. Với mức ý nghĩa 2%, hãy kiểm định xem hệ thống được cải tiến này có thực sự làm giảm tỷ lệ hủy vé hay không?

Câu 5. Doanh số bán (y) của công ty thương mại X ghi nhận được như sau:

Năm	2000	2001	2002	2003	2004	2005	2006	2007
y	35	50	66	58	74	90	108	121

Dự đoán doanh số bán năm 2010 bằng hàm xu thế tuyến tính.

Đề số 2

Câu 1. Có 30 hộp bóng đèn được đóng gói giống nhau bao gồm: 15 hộp loại I, 10 hộp loại II, 5 hộp loại III. Biết rằng mỗi hộp đều có 50 bóng đèn, trong đó:

Hộp loại I: có 1 bóng có tuổi thọ kém

Hộp loại II: có 3 bóng có tuổi thọ kém

Hộp loại III: có 6 bóng có tuổi thọ kém.

Lấy ngẫu nhiên một hộp, rồi từ hộp đó lấy ngẫu nhiên một bóng đèn

a) Tính xác suất bóng đèn lấy ra là bóng đèn có tuổi thọ kém.

b) Giả sử bóng đèn lấy ra là bóng có tuổi thọ kém, tính xác suất bóng đèn đó thuộc hộp loại II.

Câu 2. Biến ngẫu nhiên X có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \dfrac{3}{4}(x-1)(3-x), & x \in [1,3] \\ 0 & x \notin [1,3] \end{cases}$$

a) Tìm hàm phân bố xác suất F(x)?

b) Tìm xác suất để trong 3 phép thử độc lập có ít nhất một lần $X \in (1,2)$?

Câu 3. Kiểm tra độ bền (X : kg/mm^2) của một loại thép cốt có đường kính 30mm do nhà máy M sản xuất. Người ta tiến hành lấy mẫu gồm 200 thanh thép để đo độ bền, kết quả đo được cho trong bảng:

X	90 – 95	95 – 100	100 – 105	105 – 110	110 – 115	115 – 120
Số thanh	18	24	72	49	22	15

Giả sử độ bền của một loại thép cốt có đường kính 30mm do nhà máy M sản xuất tuân theo quy luật phân phối chuẩn.

a) Tìm khoảng ước lượng cho độ bền trung bình của một thanh thép loại trên do nhà máy M sản xuất với độ tin cậy 95%.

b) Với mức ý nghĩa 1% có thể kết luận rằng, độ bền trung bình của một thanh thép do nhà máy M sản xuất lớn hơn 102kg/mm^2 được không?

Câu 4. Khảo sát về tỷ lệ sản phẩm loại I do nhà máy M sản xuất, người ta kiểm tra ngẫu nhiên 400 sản phẩm do nhà máy đó sản xuất thì thấy có 154 sản phẩm loại I.

a) Hãy ước lượng tỷ lệ sản phẩm loại I của nhà máy M với độ tin cậy 95%.

b) Giám đốc nhà máy nói rằng tỷ lệ sản phẩm loại I của nhà máy tối thiểu là 40% thì có chấp nhận được không, hãy kết luận với mức ý nghĩa 5%.

Câu 5. Doanh số bán (y) của công ty thương mại X ghi nhận được như sau:

Năm	2000	2001	2002	2003	2004	2005	2006	2007
y	35	50	66	58	74	90	108	121

Dự đoán doanh số bán năm 2010 bằng tốc độ phát triển bình quân.

Đề số 3

Câu 1. Ba nhà máy A, B, C cùng sản xuất một loại sản phẩm X. Tỷ lệ chính phẩm của các nhà máy A, B và C lần lượt là 97%; 98% và 95%. Giả sử sản phẩm X bày bán ở một siêu thị chỉ do ba nhà máy A, B và C này cung cấp với tỷ lệ lần lượt là 30%; 45% và 25%. Mua một sản phẩm X ở siêu thị.

a) Tính xác suất để sản phẩm đó là chính phẩm.

b) Giả sử mua một sản phẩm X ở siêu thị và thấy sản phẩm đó là chính phẩm. Tính xác suất để sản phẩm đó do nhà máy A sản xuất.

Câu 2. Một lô hàng gồm có 6 sản phẩm loại I và 4 sản phẩm loại II. Lấy ngẫu nhiên đồng thời 3 sản phẩm từ lô hàng trên. Gọi X là số sản phẩm loại II được lấy ra.

a) Lập bảng phân phối xác suất của X.

b) Tính số sản phẩm loại II trung bình được lấy ra và cho biết xác suất có ít nhất 1 sản phẩm loại II được lấy ra là bao nhiêu?

Câu 3. Kiểm tra về trọng lượng (X) mỗi bao xi măng được đóng bao tại nhà máy M. Cân ngẫu nhiên 250 bao xi măng người ta thu được bảng số liệu sau: (đơn vị: kg)

X	49,2−49,4	49,4−49,6	49,6−49,8	49,8−50,0	50,0−50,2	50,2−50,4	50,4−50,6
Số bao	10	15	60	80	50	20	15

Giả sử trọng lượng mỗi bao xi măng được đóng bao tại nhà máy M tuân theo quy luật phân phối chuẩn.

a) Hãy ước lượng trọng lượng trung bình của một bao xi măng được sản xuất tại nhà máy M với độ tin cậy 95%?

b) Trọng lượng quy định của mỗi bao xi măng là 50 kg. Với mức ý nghĩa 1%, hãy kiểm định xem xi măng được đóng bao tại nhà máy M có đạt trọng lượng theo quy định hay không?

Câu 4. Kiểm tra ngẫu nhiên 400 sản phẩm do một nhà máy sản xuất thấy có 380 sản phẩm đạt tiêu chuẩn.

a) Người ta muốn tìm khoảng ước lượng cho tỷ lệ sản phẩm đạt tiêu chuẩn với độ chính xác không vượt quá 0,02 và độ tin cậy 95% thì cần kiểm tra ít nhất bao nhiêu sản phẩm của nhà máy sản xuất ?

b) Giám đốc nhà máy nói rằng tỷ lệ sản phẩm đạt tiêu chuẩn của nhà máy không dưới 98% thì có chấp nhận được không? Kết luận với mức ý nghĩa 5%.

Câu 5. Doanh số bán (y) của công ty thương mại X ghi nhận được như sau:

Năm	2000	2001	2002	2003	2004	2005	2006	2007
y	35	50	66	58	74	90	108	121

Dự đoán doanh số bán năm 2010 bằng lượng tăng giảm tuyệt đối bình quân.

Đề số 4

Câu 1. Tỷ lệ sinh viên tốt nghiệp trường Đại học M được xếp loại Giỏi; Khá; Trung bình lần lượt là 25%; 35%; 40%. Xác suất có việc làm sau khi tốt nghiệp trường M đối với sinh viên Giỏi là 95%; đối với sinh viên Khá là 80%; đối với sinh viên Trung bình là 60%.

a) Chọn ngẫu nhiên một sinh viên đã tốt nghiệp trường Đại học M, tính xác suất sinh viên đó sẽ có việc làm.

b) Chọn ngẫu nhiên một sinh viên tốt nghiệp trường M đã có việc làm. Tính xác suất đó là sinh viên tốt nghiệp loại Giỏi.

Câu 2. Có ba kiện hàng: kiện thứ nhất có 7 sản phẩm loại A và 3 sản phẩm loại B, kiện thứ hai có 8 sản phẩm loại A và 2 sản phẩm loại B, kiện thứ ba có 6 sản phẩm loại A và 4 sản phẩm loại B. Chọn ngẫu nhiên mỗi kiện hàng 3 sản phẩm.

a) Tính xác suất để chọn được 9 sản phẩm loại A.

b) Tính xác suất để chọn được ít nhất 2 sản phẩm loại A.

Câu 3. Điều tra mức lương (đơn vị : triệu đồng/tháng) của một số công nhân tại một xí nghiệp, ta có bảng số liệu như sau:

Mức lương	1,8 – 2	2 – 2,2	2,2 – 2,4	2,4 – 2,6	2,6 – 2,8
Số công nhân	10	15	25	20	8

Biết rằng mức lương của công nhân tại xí nghiệp này là biến ngẫu nhiên phân phối chuẩn.

a) Hãy ước lượng mức lương trung bình của các công nhân thuộc xí nghiệp này, với độ tin cậy 95%.

b) Giám đốc của xí nghiệp cho rằng lương trung bình của công nhân thuộc xí nghiệp này là 2,5 triệu đồng/tháng. Với mức ý nghĩa 5%, hãy cho biết lời của vị giám đốc này có phù hợp thực tế không?

Câu 4. Để ước lượng tỷ lệ sinh viên bị cận thị tại một trường Đại học, người ta chọn ngẫu nhiên 200 sinh viên thì thấy có 118 sinh viên bị cận thị.

a) Với độ tin cậy 90%, hãy ước lượng tỷ lệ sinh viên bị cận thị của trường Đại học này?

b) Trước đây 5 năm, tỷ lệ sinh viên bị cận thị tại trường Đại học này là 45%. Với mức ý nghĩa 5%, hãy cho biết tỷ lệ sinh viên bị cận thị hiện nay tại trường Đại học này có tăng lên so với 5 năm trước không?

Câu 5. Cho biết tình hình về tỷ lệ tăng trưởng GDP (%) của Việt Nam như sau:

2004	2005	2006	2007	2008	2009	2010	2011	2012
7,536	7,547	6,978	7,130	5,662	5,398	6,423	6,240	5,247

Dự đoán tỷ lệ tăng trưởng GDP của Việt Nam năm 2014 bằng hàm xu thế tuyến tính.

Đề số 5

Câu 1. Một cỗ máy có 3 bộ phận 1, 2, 3. Xác suất hỏng của các bộ phận trong thời gian làm việc theo thứ tự là 0,2; 0,4 và 0,3.

a) Tính xác suất có hai bộ phận hỏng.

b) Cuối ngày làm việc được biết rằng có 2 bộ phận bị hỏng. Tính xác suất để hai bộ phận hỏng đó là bộ phận 1 và 2.

Câu 2. Một loại chi tiết máy do một phân xưởng sản xuất được coi là đạt tiêu chuẩn kỹ thuật nếu đường kính của nó sai lệch so với đường kính thiết kế không quá 0,012mm. Biết rằng đường kính của các chi tiết máy do phân xưởng sản xuất là biến ngẫu nhiên tuân theo phân phối chuẩn, với độ lệch chuẩn là 0,005mm.

a) Tính tỷ lệ chi tiết máy đạt tiêu chuẩn kỹ thuật do phân xưởng này sản xuất.

b) Chọn ngẫu nhiên 20 chi tiết máy do phân xưởng này sản xuất. Tính xác suất để chọn được ít nhất 18 chi tiết máy đạt tiêu chuẩn kỹ thuật.

Câu 3. Tại một nông trường để điều tra kết quả sử dụng loại phân bón mới trên một loại trái cây, người ta lấy một mẫu ngẫu nhiên cân thử và thu được số liệu như sau:

Trọng lượng (gam)	45 – 55	55 – 60	60 – 65	65 – 70	70 – 75	75 – 80
Số trái cây	10	35	75	130	35	15

Giả sử trọng lượng một loại trái cây tuân theo quy luật phân phối chuẩn.

a) Với độ tin cậy 95%, hãy ước lượng trọng lượng trung bình của loại trái cây này của nông trường.

b) Trước kia, trọng lượng trung bình của mỗi loại trái cây này là 64gam. Với mức ý nghĩa 5%, hãy đánh giá xem loại phân bón mới có mang lại hiệu quả thực sự là làm tăng trọng lượng trung bình của trái cây lên hay không?

Câu 4. Điều tra về việc sử dụng mạng ADSL tại một địa phương, người ta khảo sát 600 hộ thì thấy có 240 hộ có sử dụng mạng Internet.

a) Hãy ước lượng số hộ có sử dụng mạng ADSL tại địa phương này với độ tin cậy 95%, biết rằng địa phương có 160000 hộ.

b) Năm trước tỷ lệ hộ dân có sử dụng mạng ADSL tại địa phương này là 35%. Với mức ý nghĩa 5%, hãy cho biết tỷ lệ hộ dân có sử dụng mạng ADSL tại địa phương này năm nay có cao hơn so với năm trước không?

Câu 5. Có tài liệu về doanh thu bán hàng của cửa hàng bách hoá như sau:

Năm	2008	2009	2010	2011	2012
Doanh thu bán hàng (triệu đồng)	7510	7680	8050	8380	8500

Dự đoán doanh thu bán hàng năm 2014 bằng lượng tăng giảm tuyệt đố bình quân.

Đề số 6

Câu 1. Tỷ lệ người dân nghiện rượu ở một nước lạc hậu là 30%. Biết rằng tỷ lệ người bị viêm gan trong số người nghiện rượu là 60%, còn tỷ lệ người bị viêm gan trong số người không nghiện rượu là 40%.

a) Lấy ngẫu nhiên một người, tính xác suất để người đó bị viêm gan.

b) Nếu người đó không bị viêm gan, tính xác suất để người đó nghiện rượu.

Câu 2. Trọng lượng các sản phẩm do một máy sản xuất ra là một ĐLNN có phân phối chuẩn với độ lệch chuẩn 1,2kg. Sản phẩm được xếp loại I nếu có trọng lượng sai lệch so với trọng lượng trung bình không quá 1,5kg.

a) Tính tỉ lệ sản phẩm loại I.

b) Muốn tỉ lệ sản phẩm loại I là 90% thì độ đồng đều (phương sai) của các sản phẩm này là bao nhiêu?

Câu 3. Đo lượng cholesterol (X: đơn vị mg%) cho nhóm người có độ tuổi từ 45 đến 60 tuổi, người ta thu được bảng số liệu sau:

X	150 – 160	160 – 170	170 – 180	180 – 190	190 – 200	200 – 210
Số người	7	18	34	20	12	9

Giả sử rằng lượng cholesterol tuân theo quy luật phân phối chuẩn.

a) Hãy ước lượng lượng cholesterol trung bình cho nhóm người có độ tuổi từ 45 đến 60 với độ tin cậy 95%.

b) Có tài liệu cho rằng lượng cholesterol trung bình là 177 (mg/%), tài liệu này có đáng tin cậy không? Hãy kết luận với mức ý nghĩa 5%.

Câu 4. Tỷ lệ phế phẩm của một nhà máy là 5%. Sau khi tiến hành một phương án cải tiến kỹ thuật nhằm giảm tỷ lệ phế phẩm, người ta kiểm tra 400 sản phẩm thì thấy có 12 phế phẩm.

a) Hãy ước lượng tỷ lệ phế phẩm của nhà máy sau khi cải tiến kỹ thuật với độ tin cậy 99%?

b) Với mức ý nghĩa 5%, hãy kết luận việc cải tiến kỹ thuật có hiệu quả hay không?

Câu 5. Có tài liệu về doanh thu bán hàng của cửa hàng bách hoá như sau:

Năm	2008	2009	2010	2011	2012
Doanh thu bán hàng (triệu đồng)	7510	7680	8050	8380	8500

Dự đoán doanh thu bán hàng năm 2014 bằng tốc độ phát triển bình quận.

Đề số 7

Câu 1. Hai nhà máy cùng sản xuất một loại bóng đèn. Năng suất nhà máy thứ nhất gấp 2 lần nhà máy thứ hai. Tỷ lệ bóng đèn không đạt chất lượng của nhà máy thứ nhất là 3% và của nhà máy thứ hai là 5%. Giả sử bóng đèn bán trên thị trường chỉ do hai nhà máy này sản xuất, nếu mua một bóng đèn trên thị trường:

a) Tính xác suất mua phải bóng đèn không đạt chất lượng.

b) Giả sử mua phải bóng đèn không đạt chất lượng. Theo bạn bóng đèn đó do nhà máy nào sản xuất với khả năng cao hơn.

Câu 2. Một nhân viên bán hàng mỗi ngày đi chào hàng ở ba nơi với xác suất bán được hàng mỗi nơi là 0,4. Tiền hoa hồng cho mỗi lần bán được hàng là 100$, chi phí cho một ngày đi chào hàng là 20$.

a) Hãy lập bảng phân phối xác suất cho số tiền hoa hồng mà nhân viên này nhận được một ngày.

b) Tính trung bình và phương sai của số tiền hoa hồng mà nhân viên này nhận được một ngày.

Câu 3. Năng suất (đơn vị tính: tấn/ha) lúa thử nghiệm của một giống lúa mới trên 49 lô đất cho kết quả như sau:

Năng suất	6 – 7	7 – 8	8 – 9	9 – 10	10 – 11	11 – 12	12 – 13
Số lô	2	3	8	20	9	5	2

Giả sử năng suất của giống lúa này tuân theo quy luật phân phối chuẩn.

a) Hãy ước lượng năng suất trung bình của giống lúa này với độ tin cậy 95%.

b) Các nhà nghiên cứu hy vọng rằng năng suất của giống lúa này ít nhất sẽ là 10,5 tấn/ha. Điều hy vọng này có thể trở thành hiện thực không với mức ý nghĩa 5%.

Câu 4. Khảo sát 500 người để lấy ý kiến về một loại sản phẩm mới người ta thấy có 348 người đánh giá tốt về loại sản phẩm này, hãy:

a) Ước lượng tỷ lệ khách hàng có đánh giá tốt đối với loại sản phẩm này, với độ tin cậy 95%.

b) Nếu muốn sai số của bài toán ước lượng là 1% thì phải khảo sát tối thiểu bao nhiêu khách hàng với độ tin cậy 95%.

Câu 5. Có tài liệu về doanh thu bán hàng của cửa hàng bách hoá như sau:

Năm	2008	2009	2010	2011	2012
Doanh thu bán hàng (triệu đồng)	7510	7680	8050	8380	8500

Dự đoán doanh thu bán hàng năm 2014 bằng hàm xu thế tuyến tính.

Đề số 8

Câu 1. Có 6 lô hàng loại I, mỗi lô có 7 sản phẩm tốt và 3 sản phẩm không tốt và 4 lô hàng loại II mỗi lô có 5 sản phẩm tốt và 5 sản phẩm không tốt. Chọn ngẫu nhiên 1 lô rồi từ đó lấy 2 sản phẩm.

a) Tính xác suất để cả hai sản phẩm đều là sản phẩm tốt.

b) Tính xác suất để hai sản phẩm lấy ra có ít nhất 1 sản phẩm tốt.

Câu 2. Trọng lượng các sản phẩm do một máy sản xuất ra là một đại lượng ngẫu nhiên có phân phối chuẩn với độ lệch chuẩn 1,2 kg. Sản phẩm được xếp loại I nếu có trọng lượng sai lệch so với trọng lượng trung bình không quá 1,5 kg.

a) Tính tỉ lệ sản phẩm loại I.

b) Muốn tỉ lệ sản phẩm loại I là 90% thì độ đồng đều (phương sai) của các sản phẩm này là bao nhiêu?

Câu 3. Theo dõi thời gian hoàn thành sản phẩm (X: đơn vị phút) của 64 công nhân, người ta thu được bảng số liệu sau :

X	10 – 12	12 – 14	14 – 16	16 – 18	18 – 20
Số công nhân	4	11	29	14	6

Giả sử thời gian hoàn thành một sản phẩm tuân theo quy luật phân phối chuẩn.

a) Hãy ước lượng thời gian hoàn thành sản phẩm trung bình của một công nhân, với độ tin cậy 95%.

b) Trước đây, định mức thời gian hoàn thành một sản phẩm là 14,5 phút. Hãy cho biết có cần thay đổi định mức này không, với mức ý nghĩa 5%?

Câu 4. Khảo sát ngẫu nhiên 200 người ở một địa phương A, người ta thấy có 42 người hút thuốc lá.

a) Hãy ước lượng tỷ lệ người hút thuốc lá tại địa phương A với độ tin cậy là 95%.

b) Người ta cho rằng có không quá 20% số người hút thuốc lá tại địa phương A, nhận định trên đúng không với mức ý nghĩa 5%.

Câu 5. Có tài liệu về sản lượng mặt hàng X (triệu mét) của một công ty qua các năm như sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014
Sản lượng	5,01	5,28	5,56	5,8	6,3	6,51	6,75	6,97

Dự đoán sản lượng năm 2016 bằng lượng tăng giảm tuyệt đối bình quân.

Đề số 9

Câu 1. Có 2 kiện hàng, kiện thứ nhất có 5 sản phẩm loại A và 10 sản phẩm loại B, kiện thứ hai có 6 sản phẩm loại A và 4 sản phẩm loại B. Xác suất chọn kiện thứ nhất và thứ hai là 1/3 và 2/3. Chọn ngẫu nhiên một kiện rồi từ đó lấy ngẫu nhiên 3 sản phẩm.

a) Tính xác suất chọn được 3 sản phẩm loại A.

b) Biết rằng 3 sản phẩm chọn được đều loại A. Hãy cho biết khả năng 3 sản phẩm này là của kiện nào?

Câu 2. Có ba kiện hàng: kiện thứ nhất có 7 sản phẩm loại A và 3 sản phẩm loại B, kiện thứ hai có 8 sản phẩm loại A và 2 sản phẩm loại B, kiện thứ ba có 6 sản phẩm loại A và 4 sản phẩm loại B. Chọn ngẫu nhiên mỗi kiện hàng 3 sản phẩm.

a) Tính xác suất để chọn được 9 sản phẩm loại A.

b) Tính xác suất để chọn được ít nhất 2 sản phẩm loại A.

Câu 3. Trọng lượng của một loại bánh A được cho là biến ngẫu nhiên tuân theo quy luật chuẩn. Chọn ngẫu nhiên 100 chiếc bánh A thấy có số liệu như sau:

Trọng lượng (g)	90	100	110	120	130	140
Số lượng bánh	4	16	25	30	15	10

a) Với độ tin cậy 95% hãy ước lượng khoảng tin cậy cho trọng lượng trung bình của loại bánh A.

b) Người chủ sản xuất cho rằng trung bình bánh của anh ta nặng không dưới 120g. Với mức ý nghĩa 5%, kết luận về ý kiến của người chủ sản xuất này?

Câu 4. Để ước lượng tỷ lệ sinh viên bị cận thị tại một trường Đại học, người ta chọn ngẫu nhiên 200 sinh viên thì thấy có 118 sinh viên bị cận thị.

a) Với độ tin cậy 90%, hãy ước lượng tỷ lệ sinh viên bị cận thị của trường Đại học này?

b) Trước đây 5 năm, tỷ lệ sinh viên bị cận thị tại trường Đại học này là 45%. Với mức ý nghĩa 5%, hãy cho biết tỷ lệ sinh viên bị cận thị hiện nay tại trường Đại học này có tăng lên so với 5 năm trước không?

Câu 5. Có tài liệu về sản lượng mặt hàng X (triệu mét) của một công ty qua các năm như sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014
Sản lượng	5,01	5,28	5,56	5,8	6,3	6,51	6,75	6,97

Dự đoán sản lượng năm 2016 bằng hàm xu thế tuyến tính.

Đề số 10

Câu 1. Có 2 kiện hàng được đóng gói giống nhau. Kiện thứ nhất 12 sản phẩm tốt và 3 sản phẩm kém chất lượng, kiện thứ hai có 10 sản phẩm tốt và 5 sản phẩm kém chất lượng. Lấy ngẫu nhiên một kiện rồi từ kiện đó lấy ngẫu nhiên 2 sản phẩm.

a) Tính xác suất 2 sản phẩm lấy ra là 2 sản phẩm tốt.

b) Biết 2 sản phẩm lấy ra là 2 sản phẩm tốt, tính xác suất 2 sản phẩm đó được lấy ra từ kiện thứ nhất.

Câu 2. Trọng lượng các sản phẩm do một máy sản xuất ra là một ĐLNN có phân phối chuẩn với độ lệch chuẩn 1,2kg. Sản phẩm được xếp loại I nếu có trọng lượng sai lệch so với trọng lượng trung bình không quá 1,5kg.

a) Tính tỉ lệ sản phẩm loại I.

b) Muốn tỉ lệ sản phẩm loại I là 90% thì độ đồng đều (phương sai) của các sản phẩm này là bao nhiêu?

Câu 3. Số liệu thống kê về doanh số bán hàng (triệu đồng/ngày) của một siêu thị trong một số ngày cho ở bảng sau:

Doanh số	24	30	36	42	48	54	60	65	70
Số ngày	5	12	25	35	24	15	12	10	6

Giả sử doanh số bán hàng tuân theo quy luật phân phối chuẩn.

a) Ước lượng doanh số bán trung bình trong 1 ngày của siêu thị, với độ tin cậy 95%?

b) Trước đây doanh số bán trung bình của siêu thị là 35 triệu đồng/ ngày. Số liệu ở bảng trên được thu thập sau khi siêu thị áp dụng một phương pháp bán hàng mới. Hãy cho biết phương thức bán hàng mới có làm tăng doanh số bán hàng của siêu thị lên hay không? Kết luận với mức ý nghĩa 5%.

Câu 4. Khảo sát 150 chung cư, khách sạn trên địa bàn thành phố người ta ghi nhận được chỉ có 60 chung cư hoặc khách sạn đảm bảo an toàn về phòng cháy chữa cháy.

a) Hãy ước lượng tỷ lệ chung cư, khách sạn đảm bảo về an toàn phòng cháy chữa cháy trên địa bàn với độ tin cậy 95%.

b) Có thông tin cho rằng có không quá 35% chung cư hoặc khách sạn đảm bảo an toàn phòng cháy chữa cháy. Hãy cho nhận định về thông tin trên với mức ý nghĩa 5%.

Câu 5. Có tài liệu về sản lượng mặt hàng X (triệu mét) của một công ty qua các năm như sau:

Năm	2007	2008	2009	2010	2011	2012	2013	2014
Sản lượng	5,01	5,28	5,56	5,8	6,3	6,51	6,75	6,97

Dự đoán sản lượng năm 2016 bằng tốc độ phát triển bình quân.

Phụ lục 1

GIẢI TÍCH TỔ HỢP

1. Quy tắc đếm

Ta chỉ khảo sát tập hữu hạn: $X = \{x_1, x_2, ..., x_n\}$, X có n phần tử,

ký hiệu $|X| = n$.

1.1. Công thức cộng

Cho X, Y là hai tập hữu hạn và $X \cap Y = \varnothing$, ta có $|X \cup Y| = |X| + |Y|$

Tổng quát: Nếu cho k tập hữu hạn $X_1, X_2, ..., X_k$ sao cho $X_i \cap Y_j = \varnothing, i \neq j$,

ta có

$$|X_1 \cup X_2 \cup ... \cup X_k| = |X_1| + |X_2| + ... + |X_k|$$

1.2. Công thức nhân

Cho X, Y là hai tập hữu hạn, định nghĩa tập tích nhý sau

$X \times Y = \{(x, y) / x \in X \wedge y \in Y\}$, ta có $|X \times Y| = |X| \cdot |Y|$

Tổng quát: Nếu cho n tập hữu hạn $X_1, X_2, ..., X_k$, ta có

$$|X_1 \times X_2 \times ... \times X_k| = |X_1| \cdot |X_2| \cdot ... \cdot |X_k|$$

1.3. Quy tắc cộng

Giả sử một công việc có thể thực hiện một trong k phương pháp, trong đó

- Phương pháp 1 có n_1 cách thực hiện,
- Phương pháp 2 có n_2 cách thực hiện,…,
- Phương pháp k có n_k cách thực hiện,

và hai phương pháp khác nhau không có cách thực hiện chung.

Khi đó, ta có $n_1 + n_2 + ... + n_k$ cách thực hiện công việc.

1.4. Quy tắc nhân

Giả sử một công việc có thể thực hiện tuần tự theo k bước, trong đó

- Bước 1 có n_1 cách thực hiện,
- Bước 2 có n_2 cách thực hiện,…,
- Bước k có n_k cách thực hiện,

Khi đó, ta có $n_1 \times n_2 \times ... \times n_k$ cách thực hiện công việc.

2. Giải tích tổ hợp

2.1. Chỉnh hợp

Định nghĩa: Chỉnh hợp chập k từ n phần tử là một bộ có kể thứ tự gồm k phần tử khác nhau lấy từ n phần tử đã cho.

Số chỉnh hợp: Số chỉnh hợp chập k từ n phần tử, ký hiệu là : A_n^k

Công thức tính : $A_n^k = n(n-1)...(n-k+1) = \frac{n!}{(n-k)!}$

Ví dụ 1. Đêm chung kết hoa khôi sinh viên thành phố có 12 thí sinh, chọn 3 thí sinh trao giải: Hoa khôi, Á khôi 1, Á khôi 2. Có bao nhiêu cách chọn ?

Giải

Nhận xét: thí sinh được trao giải, được chọn từ 12 thí sinh, và có thứ tự (A, B, C cùng được trao giải, nhưng trường hợp A là hoa khôi, khác trường hợp B là hoa khôi).

Suy ra mỗi cách chọn là một chỉnh hợp chập 3 từ 12 phần tử.

Vậy số cách chọn là: $A_{12}^3 = 12.11.10 = 1320$.

2.2. Chỉnh hợp lặp

Định nghĩa: Chỉnh hợp lặp chập k từ n phần tử là một bộ có kể thứ tự gồm k phần tử không cần khác nhau lấy từ n phần tử đã cho.

Số chỉnh hợp lặp: Số chỉnh hợp lặp chập k từ n phần tử ký, hiệu là : $\widetilde{A_n^k}$

Công thức tính: $\widetilde{A_n^k} = n^k$

Ví dụ 2. Giả sử có một vị thần có quyền phân phát ngày sinh cho con người, có bao nhiêu cách phân bố ngày sinh cho 10 em bé ra đời trong năm 1999 tại 1 khu tập thể của công nhân viên chức.

Giải

Nhận xét: Mỗi ngày sinh của một em bé là 1 trong 365 ngày của năm 1999, nên các ngày sinh có thể trùng nhau.

Suy ra mỗi cách phân bố 10 ngày sinh là một chỉnh hợp lặp chập 10 từ 365 phần tử.

Vậy số cách phân bố ngày sinh là: $\tilde{A}_{365}^{10} = 365^{10}$.

2.3. Hoán vị

Định nghĩa: Một hoán vị từ n phần tử là một bộ có kể thứ tự gồm n phần tử khác nhau đã cho.

Số hoán vị: Số hoán vị từ n phần tử, ký hiệu là P_n

Công thức tính: $P_n = n! = (n-1)(n-2)...(1)$

Ví dụ 3. Có 3 bộ sách:

Toán cao cấp : 6 tập,

Kinh tế lượng : 2 tập,

Xác suất thống kê : 3 tập,

Được đặt lên giá sách. Có bao nhiêu cách sắp:

a) Tuỳ ý;

b) Các tập sách được đặt theo từng bộ.

Giải

a) Tuỳ ý;

Ba bộ sách có tất cả 11 tập; đặt lên giá sách, mỗi cách sắp là hoán vị của 11 phần tử.

Suy ra số cách sắp tuỳ ý: $P_{11} = 11!$

b) Các tập sách được đặt theo từng bộ.

+) Xem mỗi bộ sách là một phần tử suy ra có $3!$ cách sắp xếp 3 phần tử này.

+) Các cặp sách trong mỗi bộ sách xáo trộn với nhau.

Toán cao cấp : $6!$

Kinh tế lượng : $2!$

Xác suất thống kê : $3!$

Suy ra: số cách sắp xếp 3 bộ sách theo từng bộ là: $3!6!2!3! = 51840$ cách.

2.4. Tổ hợp

Định nghĩa: Một tổ hợp chập k từ n phần tử là một tập con gồm k phần tử lấy từ n phần tử.

Số tổ hợp : Số tổ hợp chập k từ n phần tử ký hiệu là : C_n^k

Công thức tính: $C_n^k = \dfrac{n!}{k!(n-k)!}$

Ví dụ 4. Giải bóng đá ngoại hạng Anh có 20 đội bóng thi đấu vòng tròn, có bao nhiêu trận đấu được tổ chức nếu:

a) Thi đấu vòng tròn 1 lượt.

b) Thi đấu vòng tròn 2 lượt.

Giải

a) Thi đấu vòng tròn 1 lượt

Mỗi trận đấu ứng với việc chọn 2 đội chọn từ 20 đội. Suy ra mỗi trận đấu là một tổ hợp chập 2 từ 20 phần tử.

Số mỗi trận đấu được tổ chức là : $C_{20}^{2} = \dfrac{20!}{2!18!} = 190$ trận

b) Thi đấu vòng tròn 2 lượt.

Mỗi trận đấu ứng với việc chọn 2 đội chọn từ 20 đội. (đội chủ, đội khách).

Suy ra mỗi trận đấu là một chỉnh hợp chập 2 từ 20 phần tử.

Vậy số trận đấu là : $A_{20}^{2} = \dfrac{20!}{18!} = 380$ trận.

2.5. Nhị thức Newton

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k$$

$$(1+x)^n = \sum_{k=0}^{n} C_n^k x^k$$

3. Bài tập

Bài 1. Trong một lớp gồm 30 sinh viên, cần chọn ra ba sinh viên để làm lớp trưởng, lớp phó và thủ quỹ (mỗi người chỉ làm một chức). Hỏi có tất cả bao nhiêu cách bầu chọn ?

Đáp số : 24360.

Bài 2. Có bao nhiêu cách xếp 10 người ngồi thành hàng ngang sao cho A và B ngồi cạnh nhau.

Đáp số : 725760.

Bài 3. Một hộp đựng 6 bi trắng và 4 bi đen.

a) Có tất cả bao nhiêu cách lấy ra 5 bi ?

b) Có bao nhiêu cách lấy ra 5 bi trong đó có 2 bi trắng ?

Đáp số : a) 252; b) 60.

Bài 4. Trong một nhóm ứng viên gồm 7 nam và 3 nữ,

a) có bao nhiêu cách thành lập một ủy ban gồm 3 người ?

b) có bao nhiêu cách thành lập một ủy ban gồm 3 người trong đó có đúng 1 nữ ?

c) có bao nhiêu cách thành lập một ủy ban gồm 3 người trong đó có ít nhất 1 nữ ?

Đáp số : a) 120; b) 63; c) 85.

Bài 5. Cho các chữ số 0, 1, 2, 3, 4, 5. Hỏi từ các chữ số này lập được bao nhiêu số có 4 chữ số khác nhau trong đó nhất thiết có mặt chữ số 5?

Đáp số : 204.

Phụ lục 2.

Các bảng giá trị tới hạn của phân phối xác suất

1. Bảng giá trị của phân phối chuẩn tắc (Phân phối Gauss) $N(0;1)$

$$\phi_0(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt$$

Bảng 1. Giá trị hàm $\phi_0(x)$

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

2. Bảng giá trị của phân phối Student

$$P\left(|T| \ge t_\alpha\right) = \alpha \text{ với } T \sim St(n)$$

Cột 1: Giá trị độ tự do n.

Hàng 1: Giá trị nguy cơ sai lầm $\alpha / 2$

Nội dung bảng : Giá trị $t_{\alpha/2}$ tương ứng với n và α

Bảng 2. Giá trị tới hạn của phân phối Student

$\alpha / 2$	0.005	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05	0.075	0.1
1	63.656	31.821	21.205	15.894	12.706	10.579	9.058	7.916	7.026	6.314	4.165	3.078
2	9.925	6.965	5.643	4.849	4.303	3.896	3.578	3.320	3.104	2.920	2.282	1.886
3	5.841	4.541	3.896	3.482	3.182	2.951	2.763	2.605	2.471	2.353	1.924	1.638
4	4.604	3.747	3.298	2.999	2.776	2.601	2.456	2.333	2.226	2.132	1.778	1.533
5	4.032	3.365	3.003	2.757	2.571	2.422	2.297	2.191	2.098	2.015	1.699	1.476
6	3.707	3.143	2.829	2.612	2.447	2.313	2.201	2.104	2.019	1.943	1.650	1.440
7	3.499	2.998	2.715	2.517	2.365	2.241	2.136	2.046	1.966	1.895	1.617	1.415
8	3.355	2.896	2.634	2.449	2.306	2.189	2.090	2.004	1.928	1.860	1.592	1.397
9	3.250	2.821	2.574	2.398	2.262	2.150	2.055	1.973	1.899	1.833	1.574	1.383
10	3.169	2.764	2.527	2.359	2.228	2.120	2.028	1.948	1.877	1.812	1.559	1.372
11	3.106	2.718	2.491	2.328	2.201	2.096	2.007	1.928	1.859	1.796	1.548	1.363
12	3.055	2.681	2.461	2.303	2.179	2.076	1.989	1.912	1.844	1.782	1.538	1.356
13	3.012	2.650	2.436	2.282	2.160	2.060	1.974	1.899	1.832	1.771	1.530	1.350
14	2.977	2.624	2.415	2.264	2.145	2.046	1.962	1.887	1.821	1.761	1.523	1.345
15	2.947	2.602	2.397	2.249	2.131	2.034	1.951	1.878	1.812	1.753	1.517	1.341
16	2.921	2.583	2.382	2.235	2.120	2.024	1.942	1.869	1.805	1.746	1.512	1.337
17	2.898	2.567	2.368	2.224	2.110	2.015	1.934	1.862	1.798	1.740	1.508	1.333
18	2.878	2.552	2.356	2.214	2.101	2.007	1.926	1.855	1.792	1.734	1.504	1.330
19	2.861	2.539	2.346	2.205	2.093	2.000	1.920	1.850	1.786	1.729	1.500	1.328
20	2.845	2.528	2.336	2.197	2.086	1.994	1.914	1.844	1.782	1.725	1.497	1.325
21	2.831	2.518	2.328	2.189	2.080	1.988	1.909	1.840	1.777	1.721	1.494	1.323
22	2.819	2.508	2.320	2.183	2.074	1.983	1.905	1.835	1.773	1.717	1.492	1.321
23	2.807	2.500	2.313	2.177	2.069	1.978	1.900	1.832	1.770	1.714	1.489	1.319
24	2.797	2.492	2.307	2.172	2.064	1.974	1.896	1.828	1.767	1.711	1.487	1.318
25	2.787	2.485	2.301	2.167	2.060	1.970	1.893	1.825	1.764	1.708	1.485	1.316
26	2.779	2.479	2.296	2.162	2.056	1.967	1.890	1.822	1.761	1.706	1.483	1.315
27	2.771	2.473	2.291	2.158	2.052	1.963	1.887	1.819	1.758	1.703	1.482	1.314
28	2.763	2.467	2.286	2.154	2.048	1.960	1.884	1.817	1.756	1.701	1.480	1.313
29	2.756	2.462	2.282	2.150	2.045	1.957	1.881	1.814	1.754	1.699	1.479	1.311
∞	2.576	2.326	2.170	2.054	1.960	1.881	1.812	1.751	1.695	1.645	1.440	1.282

3. Bảng giá trị của phân phối Chi bình phương

$$P\left(X \geq \chi^2_{n,\alpha}\right) = \alpha \text{ khi } X \sim \chi^2(n)$$

Hàng 1 : Giá trị của α; Cột 1 : Giá trị độ tự do n.

Bảng 3. Giá trị tới hạn, $\left(\chi^2_{n,\alpha}\right)$ của phân phối chi bình phương

	0.005	**0.01**	**0.015**	**0.02**	**0.025**	**0.03**	**0.05**	**0.95**	**0.975**	**0.98**	**0.99**	**0.995**
1	7.879	6.635	5.916	5.412	5.024	4.709	3.841	0.004	0.001	0.001	0.000	0.000
2	10.597	9.210	8.399	7.824	7.378	7.013	5.991	0.103	0.051	0.040	0.020	0.010
3	12.838	11.345	10.465	9.837	9.348	8.947	7.815	0.352	0.216	0.185	0.115	0.072
4	14.860	13.277	12.339	11.668	11.143	10.712	9.488	0.711	0.484	0.429	0.297	0.207
5	16.750	15.086	14.098	13.388	12.832	12.375	11.070	1.145	0.831	0.752	0.554	0.412
6	18.548	16.812	15.777	15.033	14.449	13.968	12.592	1.635	1.237	1.134	0.872	0.676
7	20.278	18.475	17.398	16.622	16.013	15.509	14.067	2.167	1.690	1.564	1.239	0.989
8	21.955	20.090	18.974	18.168	17.535	17.011	15.507	2.733	2.180	2.032	1.647	1.344
9	23.589	21.666	20.512	19.679	19.023	18.480	16.919	3.325	2.700	2.532	2.088	1.735
10	25.188	23.209	22.021	21.161	20.483	19.922	18.307	3.940	3.247	3.059	2.558	2.156
11	26.757	24.725	23.503	22.618	21.920	21.342	19.675	4.575	3.816	3.609	3.053	2.603
12	28.300	26.217	24.963	24.054	23.337	22.742	21.026	5.226	4.404	4.178	3.571	3.074
13	29.819	27.688	26.403	25.471	24.736	24.125	22.362	5.892	5.009	4.765	4.107	3.565
14	31.319	29.141	27.827	26.873	26.119	25.493	23.685	6.571	5.629	5.368	4.660	4.075
15	32.801	30.578	29.235	28.259	27.488	26.848	24.996	7.261	6.262	5.985	5.229	4.601
16	34.267	32.000	30.629	29.633	28.845	28.191	26.296	7.962	6.908	6.614	5.812	5.142
17	35.718	33.409	32.011	30.995	30.191	29.523	27.587	8.672	7.564	7.255	6.408	5.697
18	37.156	34.805	33.382	32.346	31.526	30.845	28.869	9.390	8.231	7.906	7.015	6.265
19	38.582	36.191	34.742	33.687	32.852	32.158	30.144	10.117	8.907	8.567	7.633	6.844
20	39.997	37.566	36.093	35.020	34.170	33.462	31.410	10.851	9.591	9.237	8.260	7.434
21	41.401	38.932	37.434	36.343	35.479	34.759	32.671	11.591	10.283	9.915	8.897	8.034
22	42.796	40.289	38.768	37.659	36.781	36.049	33.924	12.338	10.982	10.600	9.542	8.643
23	44.181	41.638	40.094	38.968	38.076	37.332	35.172	13.091	11.689	11.293	10.196	9.260
24	45.558	42.980	41.413	40.270	39.364	38.609	36.415	13.848	12.401	11.992	10.856	9.886
25	46.928	44.314	42.725	41.566	40.646	39.880	37.652	14.611	13.120	12.697	11.524	10.520
26	48.290	45.642	44.031	42.856	41.923	41.146	38.885	15.379	13.844	13.409	12.198	11.160
27	49.645	46.963	45.331	44.140	43.195	42.407	40.113	16.151	14.573	14.125	12.878	11.808
28	50.994	48.278	46.626	45.419	44.461	43.662	41.337	16.928	15.308	14.847	13.565	12.461
29	52.335	49.588	47.915	46.693	45.722	44.913	42.557	17.708	16.047	15.574	14.256	13.121
30	53.672	50.892	49.199	47.962	46.979	46.160	43.773	18.493	16.791	16.306	14.953	13.787
35	60.275	57.342	55.553	54.244	53.203	52.335	49.802	22.465	20.569	20.027	18.509	17.192
40	66.766	63.691	61.812	60.436	59.342	58.428	55.758	26.509	24.433	23.838	22.164	20.707
45	73.166	69.957	67.994	66.555	65.410	64.454	61.656	30.612	28.366	27.720	25.901	24.311
50	79.490	76.154	74.111	72.613	71.420	70.423	67.505	34.764	32.357	31.664	29.707	27.991
55	85.749	82.292	80.173	78.619	77.380	76.345	73.311	38.958	36.398	35.659	33.571	31.735
60	91.952	88.379	86.188	84.580	83.298	82.225	79.082	43.188	40.482	39.699	37.485	35.534
65	98.105	94.422	92.161	90.501	89.177	88.069	84.821	47.450	44.603	43.779	41.444	39.383
70	104.215	100.425	98.098	96.387	95.023	93.881	90.531	51.739	48.758	47.893	45.442	43.275
75	110.285	106.393	104.001	102.243	100.839	99.665	96.217	56.054	52.942	52.039	49.475	47.206
80	116.321	112.329	109.874	108.069	106.629	105.422	101.879	60.391	57.153	56.213	53.540	51.172
85	122.324	118.236	115.720	113.871	112.393	111.156	107.522	64.749	61.389	60.412	57.634	55.170
90	128.299	124.116	121.542	119.648	118.136	116.869	113.145	69.126	65.647	64.635	61.754	59.196
95	134.247	129.973	127.341	125.405	123.858	122.562	118.752	73.520	69.925	68.879	65.898	63.250
100	140.170	135.807	133.120	131.142	129.561	128.237	124.342	77.929	74.222	73.142	70.065	67.328

4. Bảng giá trị của phân phối Fisher

$$P\left(X \geq f_{\alpha}(n,m)\right) = \alpha \text{ khi } X \sim F(n,m)$$

Hàng 1 : Giá trị của độ tự do (tử số) n. Cột 1: Giá trị của độ tự do (mẫu số) m.

Nội dung bảng : Giá trị $f_{\alpha}(n,m)$.

Bảng 4. Giá trị tới hạn phân phối Fisher $(\alpha = 0.05)$

	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34	2.3
23	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25
25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3	2.6	2.37	2.21	2.1	2.01	1.94	1.88	1.83

Bảng 5. Giá trị tới hạn phân phối Fisher $(\alpha = 0.05)$

	12	15	20	24	30	40	60	120	∞
1	244	246	248	249	250	251	252	253	254
2	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.5
3	8.74	8.7	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.91	5.86	5.8	5.77	5.75	5.72	5.69	5.66	5.63
5	4.68	4.62	4.56	4.53	4.5	4.46	4.43	4.4	4.37
6	4	3.94	3.87	3.84	3.81	3.77	3.74	3.7	3.67
7	3.57	3.51	3.44	3.41	3.38	3.34	3.3	3.27	3.23
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.07	3.01	2.94	2.9	2.86	2.83	2.79	2.75	2.71
10	2.91	2.85	2.77	2.74	2.7	2.66	2.62	2.58	2.54
11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.4
12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.3
13	2.6	2.53	2.46	2.42	2.38	2.34	2.3	2.25	2.21
14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.48	2.4	2.33	2.29	2.25	2.2	2.16	2.11	2.07
16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.38	2.31	2.23	2.19	2.15	2.1	2.06	2.01	1.96
18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.28	2.2	2.12	2.08	2.04	1.99	1.95	1.9	1.84
21	2.25	2.18	2.1	2.05	2.01	1.96	1.92	1.87	1.81
22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.2	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.92	1.84	1.75	1.7	1.65	1.59	1.53	1.47	1.39
120	1.83	1.75	1.66	1.61	1.55	1.5	1.43	1.35	1.25
∞	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1

Bảng 6. Giá trị tới hạn phân phối Fisher $(\alpha = 0.01)$

	1	2	3	4	5	6	7	8	9	10
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Bảng 7. Giá trị tới hạn phân phối Fisher $(\alpha = 0.01)$

	12	15	20	24	30	40	60	120	∞
1	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	99.42	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.5
3	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.1
4	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.5
5	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.6
12	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3
15	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
30	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.8
60	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.6
120	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

5. Bảng giá trị của phân phối Tukey

Bảng 7. Giá trị tới hạn phân phối Tukey $(\alpha = 0,05)$

n – k	k								
	2	3	4	5	6	7	8	9	10
1	17,97	26,98	32,82	37,08	40,41	43,12	45,40	47,36	49,07
2	6,08	8,33	9,80	10,88	11,74	12,44	13,03	13,54	13,99
3	4,50	5,91	6,82	7,50	8,04	8,48	8,85	9,18	9,46
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99
6	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92
9	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15
17	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92
30	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65
120	2,80	3,36	3,68	3,92	4,10	4,24	4,36	4,47	4,56
∞	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47

Bảng 8. Giá trị tới hạn phân phối Tukey $(\alpha = 0,05)$

n – k	k									
	11	12	13	14	15	16	17	18	19	20
1	50,59	51,96	53,20	54,33	55,36	56,32	57,22	58,04	58.83	59,56
2	14,39	14,75	15,08	15,38	15,65	15,91	16,14	16,37	16,57	16,77
3	9,72	9,95	10,2	10,3	10,5	10,7	10,8	11,0	11,1	11,2
4	8,03	8,21	8,37	8,52	8,66	8,79	8,91	9,03	9,13	9,23
5	7,17	7,32	7,47	7,60	7,72	7,83	7,93	8,03	8,12	8,21
6	6,65	6,79	6,92	7,03	7,14	7,24	7,34	7,43	7,51	7,59
7	6,30	6,43	6,55	6,66	6,76	6,85	6,94	7,02	7,10	7,17
8	6,05	6,18	6,29	6,39	6,48	6,57	6,65	6,73	6,80	6,87
9	5,87	5,98	6,09	6,19	6,28	6,36	6,44	6,51	6,58	6,64
10	5,72	5,83	5,93	6,03	6,11	6,19	6,27	6,34	6,40	6,47
11	5,61	5,71	5,81	5,90	5,98	6,06	6,13	6,20	6,27	6,33
12	5,51	5,61	5,71	5,80	5,88	5,95	6,02	6,09	6,15	6,21
13	5,43	5,53	5,63	5,71	5,79	5,86	5,93	5,99	6,05	6,11
14	5,36	5,46	5,55	5,64	5,71	5,79	5,85	5,91	5,97	6,03
15	5,31	5,40	5,49	5,57	5,65	5,72	5,78	5,85	5,90	5,96
16	5,26	5,35	5,44	5,52	5,59	5,66	5,73	5,79	5,84	5,90
17	5,21	5,31	5,39	5,47	5,54	5,61	5,67	5,73	5,79	5,84
18	5,17	5,27	5,35	5,43	5,50	5,57	5,63	5,69	5,74	5,79
19	5,14	5,23	5,31	5,39	5,46	5,53	5,59	5,65	5,70	5,75
20	5,11	5,20	5,28	5,36	5,43	5,49	5,55	5,61	5,66	5,71
24	5,01	5,10	5,18	5,25	5,32	5,38	5,44	5,49	5,55	5,59
30	4,92	5,00	5,08	5,15	5,21	5,27	5,33	5,38	5,43	5,47
40	4,82	4,90	4,98	5,04	5,11	5,16	5,22	5,27	5,31	5,36
60	4,73	4,81	4,88	4,94	5,00	5,06	5,11	5,15	5,20	5,24
120	4,74	4,71	4,78	4,84	4,90	4,95	5,00	5,04	5,09	5,13
∞	4,55	4,62	4,68	4,74	4,80	4,85	4,89	4,93	4,97	5,01

Bảng 9. Giá trị tới hạn phân phối Tukey $(\alpha = 0,01)$

n – k	k								
	2	3	4	5	6	7	8	9	10
1	90,03	135	164,3	185,6	202,2	215,8	227,2	237,0	245,6
2	14,04	19,02	22,29	24,72	26,63	28,20	29,53	30,68	31,69
3	8,26	10,62	12,17	13,33	4,24	15,00	15,64	16,20	16,69
4	6,51	8,12	9,17	9,96	10,58	11,10	11,55	11,93	12,27
5	5,70	6,97	7,80	8,42	8,91	9,32	9,67	9,97	10,24
6	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,10
7	4,95	5,92	6,54	7,01	7,37	7,68	7,64	8,17	8,37
8	4,74	5,63	6,20	6,63	6,96	7,24	7,47	7,68	7,87
9	4,60	5,43	5,96	6,35	6,66	6,91	7,13	7,32	7,49
10	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21
11	4,39	5,14	5,62	5,97	6,25	6,48	6,67	6,84	6,99
12	4,32	5,04	5,50	5,84	6,10	6,32	6,51	6,67	6,81
13	4,26	4,96	5,40	5,73	5,98	6,19	6,37	6,53	6,67
14	4,21	4,89	5,32	5,63	5,88	6,08	6,26	6,41	6,54
15	4,17	4,83	5,25	5,56	5,80	5,99	6,16	6,31	6,44
16	4,13	4,78	5,19	5,49	5,72	5,92	6,08	6,22	6,35
17	4,10	4,74	5,14	5,43	5,66	5,85	6,01	6,15	6,27
18	4,07	4,70	5,09	5,38	5,60	5,79	5,94	6,08	6,20
19	4,05	4,67	5,05	5,33	5,55	5,73	5,89	6,02	6,14
20	4,02	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09
24	3,96	4,54	4,91	5,19	5,37	5,54	5,69	5,81	5,92
30	3,89	4,45	4,80	5,05	5,24	5,40	5,54	5,65	5,76
40	3,82	4,37	4,70	4,93	5,11	5,27	5,39	5,50	5,60
60	3,76	4,28	4,60	4,82	4,99	5,13	5,25	5,36	5,45
120	3,70	4,20	4,50	4,71	4,87	5,01	5,12	5,21	5,30
∞	3,64	4,12	4,40	4,60	4,76	4,88	4,99	5,08	5,16

Bảng 10. Giá trị tới hạn phân phối Tukey $(\alpha = 0,01)$

n – k	k									
	11	12	13	14	15	16	17	18	19	20
1	253	260	266	272	277	282	286	290	294	298
2	32,6	33,4	34,1	34,8	35,4	36,0	36,5	37,0	37,5	37,9
3	17,1	13,5	17,9	18,2	18,5	18,8	19,1	19,3	19,5	19,8
4	12,6	12,8	13,1	13,3	13,5	13,7	13,9	14,1	14,2	14,4
5	10,5	10,7	10,9	11,1	11,2	11,4	11,6	11,7	11,8	11,9
6	9,30	9,49	9,65	9,81	9,95	10,1	10,2	10,3	10,4	10,5
7	8,55	8,71	8,86	9,00	9,12	9,24	9,35	9,46	9,55	9,65
8	8,03	8,18	8,31	8,44	8,55	8,66	8,76	8,85	8,94	9,03
9	7,65	7,78	7,91	8,03	8,13	8,23	8,32	8,41	8,49	8,57
10	7,36	7,48	7,60	7,71	7,81	7,91	7,99	8,07	8,15	8,22
11	7,13	7,25	7,36	7,46	7,56	7,65	7,73	7,81	7,88	7,95
12	6,94	7,06	7,17	7,26	7,36	7,44	7,52	7,59	7,66	7,73
13	6,79	6,90	7,01	7,10	7,19	7,27	7,34	7,42	7,48	7,55
14	6,66	6,77	6,87	6,96	7,05	7,12	7,20	7,27	7,33	7,39
15	6,55	6,66	6,76	6,84	6,93	7,00	7,07	7,14	7,20	7,26
16	6,46	6,56	6,66	6,74	6,82	6,90	6,97	7,03	7,09	7,15
17	6,38	6,48	6,57	6,66	6,73	6,80	6,87	6,94	7,00	7,05
18	6,31	6,41	6,50	6,58	6,65	6,72	6,79	6,85	6,91	6,96
19	6,25	6,34	6,43	6,51	6,58	6,65	6,72	6,78	6,84	6,89
20	6,19	6,29	6,37	6,45	6,52	6,59	6,65	6,71	6,76	6,82
24	6,02	6,11	6,19	6,26	6,33	6,39	6,45	6,51	6,56	6,61
30	5,85	5,93	6,01	6,08	6,14	6,20	6,26	6,31	6,36	6,41
40	5,69	5,77	5,84	5,90	5,96	6,02	6,07	6,12	6,17	6,21
60	5,53	5,60	5,67	5,73	5,79	5,84	5,89	5,93	5,98	6,02
120	5,38	5,44	5,51	5,56	5,61	5,66	5,71	5,75	5,79	5,83
∞	5,23	5,29	5,35	5,40	5,45	5,49	5,54	5,57	5,61	5,65

6. Bảng giá trị của phân phối Wilconxon

Bảng 11. Giá trị tới hạn trong kiểm định dấu và hạng Wilconxon

	Một phía $\alpha = 0,05$	Một phía $\alpha = 0,025$	Một phía $\alpha = 0,1$	Một phía $\alpha = 0,005$
	Hai phía $\alpha = 0,1$	Hai phía $\alpha = 0,05$	Hai phía $\alpha = 0,02$	Hai phía $\alpha = 0,01$
(Cận dưới; cận trên)				
5	0 ; 15	_ ; _	_ ; _	_ ; _
6	2 ; 19	0 ; 21	_ ; _	_ ; _
7	3 ; 25	2 ; 26	0 ; 28	_ ; _
8	5 ; 31	3 ; 33	1 ; 35	0 ; 36
9	8 ; 37	5 ; 40	3 ; 42	1 ; 44
10	10 ; 45	8 ; 47	5 ; 50	3 ; 52
11	13 ; 53	10 ; 56	7 ; 59	5 ; 61
12	17 ; 61	13 ; 65	10 ; 68	7 ; 71
13	21 ; 70	17 ; 74	12 ; 79	10 ; 81
14	25 ; 80	21 ; 84	16 ; 89	13 ; 92
15	39 ; 90	25 ; 95	19 ; 101	16 ; 104
16	35 ; 101	29 ; 107	23 ; 113	19 ; 117
17	41 ; 112	34 ; 119	27 ; 126	23 ; 130
18	47 ; 124	40 ; 131	32 ; 139	27 ; 144
19	53 ; 137	46 ; 144	37 ; 153	32 ; 158
20	60 ; 150	52 ; 158	43 ; 167	37 ; 173

Bảng 12. Giá trị tới hạn trong kiểm định tổng và hạng Wilconxon

n_2	Mức ý nghĩa		n_1						
	Một phía	Hai phía	4	5	6	7	8	9	10
4	0,05	0,1	11;25						
	0,025	0,05	10;26						
	0,01	0,02							
	0,005	0,01							
5	0,05	0,1	12;28	19;36					
	0,025	0,05	11;29	17;38					
	0,01	0,02	10;30	16;39					
	0,005	0,01		15;40					
6	0,05	0,1	13;31	20;40	28;50				
	0,025	0,05	12;32	18;42	26;52				
	0,01	0,02	11;33	17;43	24;54				
	0,005	0,01	10;34	16;44	23;55				
7	0,05	0,1	14;34	21;44	29;55	39;66			
	0,025	0,05	13;35	20;45	27;57	36;69			
	0,01	0,02	11;37	18;47	25;59	34;71			
	0,005	0,01	10;38	16;49	24;60	32;73			
8	0,05	0,1	15;37	23;47	31;59	41;71	51;85		
	0,025	0,05	14;38	21;49	29;61	38;74	49;87		
	0,01	0,02	12;40	19;51	27;63	35;77	45;91		
	0,005	0,01	11;41	17;53	25;65	34;78	43;93		
9	0,05	0,1	16;40	24;51	33;63	43;76	54;90	66;105	
	0,025	0,05	14;38	22;53	31;65	40;79	51;93	62;109	
	0,01	0,02	13;43	20;55	28;68	37;82	47;97	59;112	
	0,005	0,01	11;45	18;57	26;70	35;84	45;99	56;115	
10	0,05	0,1	17;43	26;54	35;67	45;81	56;96	69;111	82;128
	0,025	0,05	15;45	23;57	32;70	42;84	53;99	65;115	78;132
	0,01	0,02	13;47	21;59	29;73	39;87	49;103	61;119	74;136
	0,005	0,01	12;48	19;61	27;75	37;89	47;105	58;105	71;139