ThS. NGUYỄN ĐÌNH ÁI (Chủ biên)
ThS. Thái Bảo Khánh, ThS. Nguyễn Thị Hà, ThS. Nguyễn T. Thùy Dung,
ThS. Nguyễn Quang Tuấn, ThS. Nguyễn T. Minh Ngọc.
Hiệu đính: TS. Phạm Gia Hưng

XÁC SUẤT - THỐNG KÊ

LÝ THUYẾT VÀ BÀI TẬP

ĐẠI HỌC NHA TRANG
Bộ môn Toán, 9/2021

Chương 1.
Biến cố và xác suất của các biến cố

Lý thuyết xác suất có thể nói là được hình thành từ những trò chơi tung đồng tiền, cờ bạc hay các trò chơi may rủi. Nhiều nghịch lý được phát hiện dẫn đến những tranh cãi kịch liệt ở Thế kỷ 19, dẫn đến luồng quan điểm coi lý thuyết xác suất là "khoa học ngây thơ". Do nhu cầu phát triển như vũ bão của khoa học ở đầu Thế kỷ 20, do đòi hỏi của vật lý, thiên văn, sinh học…, và dựa trên lý thuyết tập hợp, lý thuyết độ đo đã rất phát triển, Kolmogrov, nhà bác học Nga hàng đầu đã đưa ra hệ tiên đề của Lý thuyết xác suất, làm cơ sở toán học vững chắc cho ngành toán học này. Lý thuyết xác suất là cơ sở của thống kê toán, một ngành của toán học được ứng dụng rộng rãi trong nhiều lĩnh vực hiện nay.

Chương này trình bày các khái niệm cơ bản nhất của lý thuyết xác suất, đó là khái niệm về phép thử, biến cố và các phép toán về biến cố, xác suất và các công thức tính xác suất. Sau khi học xong chương này, sinh viên có thể nắm được các khái niệm trên và vận dụng thành thạo các công cụ tính xác suất để giải quyết được các bài toán xác suất cơ bản.

1.1. Các khái niệm về phép thử, biến cố, không gian mẫu

1.1.1. Phép thử ngẫu nhiên (Random experiment)

Định nghĩa: Phép thử ngẫu nhiên (gọi tắt là phép thử) là những hành động mà chúng ta không biết trước được kết quả.

Ví dụ: Các phép thử có thể lấy từ những trò chơi mang tính cờ bạc như : xóc đĩa (tung đồng xu) ; cá ngựa/bầu cua/đổ xí ngầu (tung xúc xắc); xì dzách/phỏm/poker (rút 1 quân bài) ; mua một tờ vé số…Và lẽ dĩ nhiên, khi thực hiện những hành động trên chúng ta không biết trước sẽ nhận được kết quả nào.

1.1.2. Biến cố và phân loại các biến cố

Định nghĩa: Biến cố (*event*) hay sự kiện là những kết quả/kết cục có thể xảy ra của một phép thử tương ứng.

Ví dụ: Với việc tung 1 đồng xu, ta nhận được 2 kết quả có thể xảy ra: sấp/ngửa. Còn với việc tung 1 con xúc xắc ta nhận được 6 kết quả có thể: mặt 1,…, mặt 6.

Phân loại các biến cố: Để có thể nghiên cứu một cách sâu hơn, tổng quát hơn, có hệ thống hơn về một đối tượng nào đó người ta sẽ tiến hành phân loại, sắp xếp chúng theo các lớp/nhóm khác nhau và nghiên cứu các lý thuyết về chúng trong một cấu trúc nào đó (đối với lý thuyết xác suất người ta hay gọi là không gian xác suất).

Thực hiện một phép thử. Một biến cố của phép thử được gọi là

- Biến cố chắc chắn, ký hiệu là Ω, là biến cố nhất định xảy ra khi thực hiện phép thử.
- Biến cố không thể, ký hiệu là $\varnothing$, là biến cố nhất định không xảy ra khi thực hiện phép thử.
- Biến cố ngẫu nhiên (*random event*), ký hiệu là $A, B, C, D, \ldots$, là biến cố có thể xảy ra hoặc không xảy ra khi thực hiện 1 lần thử. Để viết một biến cố ngẫu nhiên, ta viết A="tên kết quả/kết cục".

- Biến cố sơ cấp (*simple event*), ký hiệu là ω, là biến cố không thể phân tích nhỏ hơn được nữa (theo một nghĩa nào đó).
- Không gian mẫu (*sample space*) là tập hợp tất cả các biến cố sơ cấp trong một phép thử và dùng lại ký hiệu của biến cố chắc chắn là Ω.

Ví dụ 1. Với phép thử "tung 1 con xúc xắc" (cân đối, đồng chất), ta thấy

- Biến cố chắc chắn là Ω = "xuất hiện mặt không lớn hơn 6",
- Biến cố không thể là $\varnothing$ = "xuất hiện mặt 7",
- Các biến cố ngẫu nhiên, chẳng hạn như, A= "xuất hiện mặt chẵn"; B ="xuất hiện chia hết cho 3",…,
- Các biến cố sơ cấp là ω_1 = "xuất hiện mặt 1", ω_2 = "xuất hiện mặt 2",…(hiểu theo nghĩa không chia nhỏ được các trường hợp cụ thể của kết quả). Như vậy, ta có không gian mẫu gồm 6 biến cố sơ cấp sau $\Omega = \left\{\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}\right\}$.

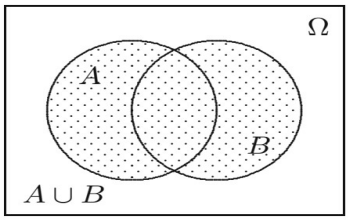
Ví dụ 2. Với phép thử "rút 1 quân trong bộ bài tây 52 quân", ta thấy

- Biến cố chắc chắn là Ω = "rút được quân đỏ hoặc quân đen",
- Biến cố không thể là $\varnothing$ = "rút được quân 11",
- Các biến cố ngẫu nhiên, chẳng hạn như, A= "rút được quân át cơ"; B= "rút được quân đầm bích", C= "rút được quân rô",…,
- Việc xác định các biến cố sơ cấp/không gian mẫu trong phép thử này, có thể là, hoặc liệt kê hết 52 quân bài, hoặc chia theo các nước bài: bích/chuồn/rô/cơ, hoặc chia theo màu của quân bài: đỏ/đen.

1.2. Các phép toán về biến cố

1.2.1. Biến cố tổng (Union of events)

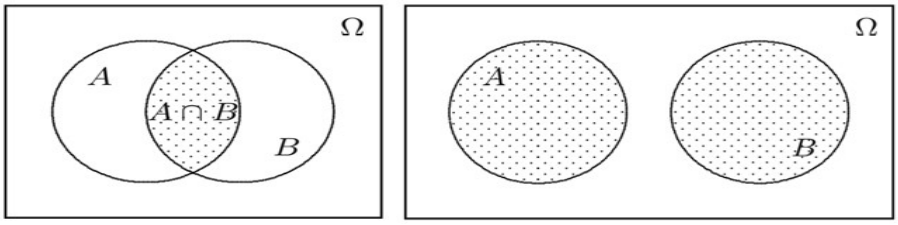
Định nghĩa: Tổng (hợp) của hai biến cố, kí hiệu $C = A \cup B$ hay $C = A + B$ là biến cố xảy ra khi và chỉ khi có ít nhất biến cố A hay biến cố B xảy ra.



Hình 1.1. Tổng của hai biến cố.

1.2.2. Biến cố tích (Intersection of events)

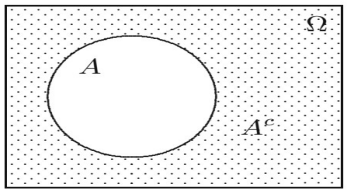
Định nghĩa: Tích (giao) của hai biến cố, kí hiệu $D = A \cap B$ hay $C = A.B$ là biến cố xảy ra khi và chỉ khi biến cố A và biến cố B đồng thời xảy ra. Nếu $A.B = \varnothing$ thì ta nói hai biến cố A và B là hai biến cố xung khắc.



Hình 1.2. Tích của hai biến cố và hai biến cố xung khắc.

1.2.3. Biến cố đối (Complement of event)

Định nghĩa: Biến cố đối của biến cố *A*, kí hiệu là $\overline{A}$ hay A^C là biến cố xảy ra khi và chỉ khi biến cố *A* không xảy ra.



Hình 1.3. Biến cố đối.

1.2.4. Hệ đầy đủ

Định nghĩa: Hệ *n* biến cố $\{A_i,\ i=\overline{1,n}\}$ được gọi là hệ đầy đủ nếu các biến cố là xung khắc từng đôi và tổng của chúng là một biến cố chắc chắn. Nghĩa là

$$A_i.A_j=\varnothing,\ i\neq j \quad \text{và} \quad A_1\cup A_2\cup\cdots\cup A_n=\Omega.$$

Hình 1.4. Hệ biến cố đầy đủ.

Tính chất.

a) $A\cup\overline{A}=\Omega$.
b) $A.\overline{A}=\varnothing$.
b) $\overline{AB}=\overline{A}\cup\overline{B}$.
d) $\overline{A\cup B}=\overline{A}.\overline{B}$.

Ví dụ 1. Có hai hộp bi ký hiệu là H_1 & H_2 chứa cả bi đỏ và bi trắng. Lấy ngẫu nhiên từ mỗi hộp ra 1 bi và gọi D_i là biến cố lấy được bi đỏ từ hộp thứ *i*, $i=1,2$. Khi đó,

- Biến cố lấy được ít nhất một bi đỏ là $A=D_1\cup D_2$.
- Biến cố lấy được hai bi đỏ là $B=D_1.D_2$.
- Biến cố lấy được (đúng) một bi đỏ là $C=\overline{D}_1.D_2\cup D_1.\overline{D}_2$. Biến cố *C* xung khắc với biến cố *B*.

Ví dụ 2. Với mọi biến cố *A*, ta có $\{A,\ \overline{A}\}$ là một hệ biến cố đầy đủ.

Ví dụ 3. Tung 1 con xúc sắc. $A_i,\ i=1,2,...,6$, khi đó hệ $\left\{A_1,A_2,A_3,A_4,A_5,A_6\right\}$ là hệ biến cố đầy đủ.

1.3. Các định nghĩa về xác suất

Để so sánh hay đánh giá khả năng xảy ra của một hay nhiều biến cố trong phép thử tương ứng, người ta gán cho mỗi biến cố một con số không âm sao cho với hai biến cố bất kỳ, biến cố nào có khả năng xảy ra nhiều hơn thì được gán cho số lớn hơn, các biến cố có cùng khả năng xảy ra thì được gán cho cùng một số.

Số gán cho biến cố A, ký hiệu là $P(A)$, là một số thỏa mãn $0 \le P(A) \le 1$, biểu thị cho khả năng xảy ra của biến cố đó khi thực hiện phép thử và được gọi là xác suất (probability) của biến cố A.

Sau đây chúng ta tìm hiểu các định nghĩa xác suất của một biến cố.

1.3.1. Định nghĩa xác suất cổ điển

Xét một phép thử có hữa hạn các biến cố sơ cấp đồng khả năng. Khi đó, xác suất xảy ra biến A được định nghĩa như sau

$$p(A) = \frac{m_A}{n},$$

trong đó n là số các biến cố sơ cấp đồng khả năng có thể xảy ra khi thực hiện phép thử; và m_A là số các biến cố sơ cấp thuận lợi cho biến cố A.

Nhận xét:

- Đây là định nghĩa đầu tiên nhất của xác suất, nó dùng để tính khả năng chiến thắng của cửa mà chúng ta đã đặt cược, do đó, bên cạnh điều kiện có hữa hạn các kết cục thì điều kiện *đồng khả năng* là yếu tố bắt buộc để tránh việc gian lận trong các trò chơi cờ bạc.
- Để tính được xác suất theo định nghĩa cổ điển, chúng ta phải thực hiện *phép đếm*. Trong một vài trường hợp đơn giản, chúng ta có thể liệt kê và đếm các kết cục bằng cách đếm cơ bản nhất là: 1,2,…Đối với các trường hợp phức tạp, chúng ta sẽ cần đến các công cụ đếm của tổ hợp, và các kiến thức về tổ hợp như: hoán vị, chỉnh hợp, tổ hợp đã được học ở chương trình phổ thông.

Ví dụ 1. Tung một con xúc xắc cân đối, đồng chất. Tính xác suất để xuất hiện mặt chia hết cho 3.

Giải. Phép thử "tung 1 con xúc xắc cân đối, đồng chất". Gọi A = "xuất hiện mặt chia hết cho 3". Ta có:

+ Số kết cục thuận lợi để A xảy ra là: $m_A = 2$ ("mặt 3" hoặc "mặt 6").

+ Số kết cục có thể xảy ra là: $n = 6$.

Vậy xác suất để xuất hiện mặt chia hết cho 3 là

$$P(A) = \frac{2}{6} = \frac{1}{3} = 0,3333.$$

Ví dụ 2. Nhà cái mở cá cược trong trận cầu chung kết C1 năm 2022 giữa PSG và Barcelona, bạn cược tỉ số của cả trận là 3-1. Hỏi xác suất để bạn thắng cược là bao nhiêu?

Giải. Phép thử "cược tỉ số trận đấu là 3-1". Gọi A là biến cố "thắng cược". Vì chỉ có một tỉ số nên số kết cục thuận lợi để A xảy ra (thắng cược) là $m_A = 1$.

Để đếm số kết cục có thể xảy ra, ta nhận xét: các tỉ số luôn có dạng $x - y$. Về mặt lý thuyết, x, y có thể nhận các giá trị là các số tự nhiên, chẳng hạn như $x = 90, y = 0$ (nghĩa là cứ 1 phút ghi 1 bàn). Tuy nhiên, trên thực tế điều đó chưa xảy ra. Thông thường, ta hay xét tập giá trị mà x, y có thể nhận là $\{0,1,2,3,4,5\}$. Như vậy, x sẽ có 6 cách chọn và y cũng có 6 cách chọn từ tập giá trị. Do đó, số kết cục có thể xảy ra, được tính theo *quy tắc nhân*, là $n = 6.6 = 36$.

Vậy xác suất để thắng cược là

$$P(A) = \frac{1}{36} = 0,027777... = 0,02\overline{7}.$$

1.3.2. Định nghĩa xác suất theo thống kê

Định nghĩa cổ điển đòi hỏi điều kiện áp dụng khắc khe: số kết cục có thể xảy ra là hữu hạn và các kết cục phải đồng khả năng. Như vậy, khi gặp các phép thử có vô hạn các kết cục có thể xảy ra (ví dụ, cho một đoạn dây, hỏi có bao nhiêu cách tạo thành 1 tam giác từ đoạn dây ấy? Câu trả lời là vô hạn cách!) hoặc điều kiện đồng khả năng là khó gặp trong thực tế nhất vì các cá thể, đối tượng hầu hết là "duy nhất và khác biệt".

Xét một phép thử và một biến cố A. Người ta tiến hành tìm số $P(A)$ – xác suất xảy ra biến A như sau

- Lặp lại phép thử n lần độc lập nhau.
- Quan sát kết quả của biến cố A: xảy ra/không xảy ra ở mỗi lần thử. Ghi lại số lần biến cố A xảy ra. Số lần biến cố A xảy ra trong n lần lặp lại phép thử được gọi là *tần số xảy ra biến cố A,* ký hiệu là f_A. Chú ý rằng, f_A sẽ nhận các giá trị trong tập $\{0,1,2,...,n\}$.
- Lập tỉ số : $p_n = f_A / n$, tỉ số này được gọi là *tần suất* xảy ra biến cố A. Dễ thấy, $0 \le p_n \le 1$ và người ta thấy rằng: khi tiến hành lặp lại càng nhiều phép thử thì tần suất *dao động xung quanh* (*/xấp xỉ*) một hằng số nào đó, chẳng hạn $p_n = \frac{f_A}{n} \approx p_0$.
- Khi đó, ta nói xác suất xảy ra biến cố A là $P(A) = p_0$.

Nhận xét. Định nghĩa xác suất xảy ra một biến cố nào đó theo thống kê mang nhiều ý nghĩa thực tế, nó phù hợp với tính không đồng nhất/không đồng khả năng ở các cá thể riêng biệt, và tính che giấu thông tin ở các hiện tượng. Tuy nhiên, ở một hiện tượng không có tính đồng bộ cao thì việc tìm ra số p_0 gặp nhiều khó khăn. Hơn nữa, giá trị p_0 chỉ mang tính xấp xỉ, do đó giá trị xác suất xảy ra biến cố A có thể khác nhau ở các nhóm nghiên cứu độc lập. Ngoài ra, việc tìm xác suất xảy ra biến cố A theo thống kê bắt buộc chúng ta phải hành động và quan sát, nghĩa là phải thực hiện phép thử trong thực tế, điều đó dẫn đến việc tốn nhiều tiền của, thời gian, công sức, kể cả có những phép thử mang tính hủy hoại dẫn đến việc chúng ta không trả lại được nguyên trạng ban đầu, ví dụ như: đạn được bắn đi, trứng được đập vỡ, hộp sữa được khui,…

Ví dụ: Khi ta nói: Xác suất khám và chữa khỏi bệnh (đối với một loại bệnh cụ thể) của một bác sĩ nào đó là 80%, có nghĩa là ta đang nói về tỉ số của số bệnh nhân khỏi bệnh chia cho tổng số bệnh nhân mắc loại bệnh này đến khám. Xác suất bán được hàng một nhân viên sale được tính là tỉ số của số lần bán được hàng chia cho số lần đến chào hàng. Xác suất sút thành công penalty của CR7 là tỉ số của số lần sút thành công chia cho tổng số lần sút,…

Tính chất:

a) $0 \le p(A) \le 1$. b) $p(\varnothing) = 0$. c) $p(\Omega) = 1$.

1.4. Công thức cộng xác suất và nhân xác suất

Việc tính xác suất của một biến cố nói chung là khó, do đó, khi thực hành ta sẽ dùng các phép toán về biến cố để biểu diễn biến cố cần tính xác suất theo các biến cố khác mà chúng ta đã biết thông tin. Từ đó việc tính xác suất sẽ dễ dàng hơn.

1.4.1. Công thức cộng xác suất

Định lý 1.1: Với A, B là hai biến cố bất kỳ, ta có

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

Hệ quả 1. Nếu $A.B = \varnothing$ (A, B là hai biến cố xung khắc) thì

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

Hệ quả 2. Ta có

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

Tổng quát, nếu $\{A_i,\ i = \overline{1,n}\}$ xung khắc từng đôi, tức là A_i, A_j xung khắc với $i \neq j$, thì

$$P(\sum_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

hay

$$P(A_1 + A_2 + ... + A_n) = P(A_1) + P(A_2) + ... + P(A_n).$$

Ví dụ 1. Ngân hàng ACB ra thông báo cần tuyển 3 nhân viên. Có 10 ứng viên tham gia dự tuyển, trong đó có 7 ứng viên nữ, 3 ứng viên nam. Tính xác suất để ngân hàng tuyển được ít nhất 2 nữ.

Giải. Phép thử: Ngân hàng ACB tuyển dụng 3 nhân viên trong 10 ứng viên tham gia. Gọi

A = "Ngân hàng tuyển được ít nhất 2 nhân viên nữ",

B = "Ngân hàng tuyển được 2 nữ và 1 nam",

C = "Ngân hàng tuyển được 3 nữ".

Ta có: $A = B + C$. Suy ra $P(A) = P(B + C)$. Hơn nữa, do B và C xung khắc với nhau nên xác suất để ngân hàng ACB tuyển được ít nhất 2 nhân viên nữ là

$$P(A) = P(B + C) = P(B) + P(C) = \frac{C_7^2.C_3^1}{C_{10}^3} + \frac{C_7^3}{C_{10}^3} = \frac{28}{40}.$$

Ví dụ 2. Tỉ lệ người dân TP. Nha Trang sử dụng mạng di động Viettel và Mobifone tương ứng là 85% và 75%. Tỉ lệ người dân dùng cả 2 mạng di động cùng lúc là 69%. Gặp ngẫu nhiên một người dân, tính xác suất để người đó sử dụng ít nhất 1 trong 2 mạng di động nói trên.

Giải. Phép thử: gặp ngẫu nhiên một người dân. Gọi

A = "Người đó sử dụng ít nhất 1 trong 2 mạng di động là Viettel và Mobifone",

B = "Người đó sử dụng mạng di động Viettel",

C = "Người đó sử dụng mạng Mobifone".

Ta có: $A = B + C$. Suy ra $P(A) = P(B + C)$. Khi đó

$$P(A) = P(B + C) = P(B) + P(C) - P(B.C) = 0,85 + 0,75 - 0,69 = 0,91.$$

Vậy xác suất gặp được người sử dụng ít nhất 1 trong 2 mạng di động Viettel và Mobifone là 0,91.

1.4.2. Xác suất có điều kiện và công thức nhân xác suất

Xác suất có điều kiện nghĩa là xác suất xảy ra một biến cố này dựa trên dữ kiện một/nhóm các biến cố nào đó đã xảy ra trước. Ví dụ, nếu ngân hàng ACB chi nhánh Khánh Hòa cần tuyển 1 nhân viên và có 10 hồ sơ ứng tuyển. Theo lẽ thường, ta nghĩ xác suất để mình được tuyển là 1/10, tuy nhiên ngân hàng lại áp điều kiện: chỉ tuyển người có hộ khẩu Khánh Hòa, và trong 10 hồ sơ ứng tuyển đó chỉ có 6 bộ hồ sơ thỏa điều kiện. Như vậy, nếu ta là 1 trong 6 bộ hồ sơ đó thì coi như xác suất được tuyển là 1/6; không có trong 6 bộ hồ sơ đó thì xác suất được tuyển là 0. Nghĩa là ta tính tỉ số của tập giao trên tập điều kiện.

Định nghĩa: Xác suất xảy ra biến A với điều kiện B đã xảy ra, ký hiệu là $P(A \mid B)$, được định nghĩa

$$P(A \mid B) = \frac{P(A.B)}{P(B)}.$$

Nhận xét. Theo trên, chúng ta thấy khi một biến cố nào đó xảy ra trước, nói chung sẽ ảnh hưởng đến xác suất xảy ra của biến cố tiếp theo. Hai biến A và B được gọi là độc lập với nhau nếu A xảy ra hay không xảy ra cũng không ảnh hưởng đến B và ngược lại. Khi đó,

$$P(A \mid B) = P(A).$$

Định lý 1.2. Cho A và B là hai biến cố trong cùng một phép thử. Khi đó

a) Nếu A và B độc lập với nhau thì $P(A.B) = P(A).P(B)$.

b) Nếu A và B là hai biến cố bất kỳ thì $P(A.B) = P(A).P(B \mid A) = P(B).P(A \mid B)$.

Ví dụ 1. Có 2 hộp chứa các viên bi. Hộp I chứa 3 bi trắng và 4 bi đen. Hộp II chứa 4 bi trắng, 3 bi đen và 2 bi vàng. Từ mỗi hộp lấy ra 1 viên bi. Tính xác suất để lấy được 2 viên bi cùng màu.

Giải. Phép thử: Từ mỗi hộp lấy ra 1 viên bi. Gọi

A = "Lấy được 2 viên bi cùng màu",

T_i="Lấy được bi trắng ở hộp thứ i", $i = 1,2$,

D_i = "Lấy được bi đen ở hộp thứ i", $i = 1,2$.

Ta có $A = T_1T_2 + D_1D_2$. Suy ra

$$p(A) = p(T_1T_2 + D_1D_2).$$

Do T_1T_2, D_1D_2 xung khắc với nhau nên

$$p(A) = p(T_1T_2) + p(D_1D_2).$$

Hơn nữa, vì T_1, T_2 độc lập và D_1, D_2 độc lập nên

$$p(A) = pT_1)p(T_2) + p(D_1)p(D_2) = \frac{3}{7}\frac{4}{9} + \frac{4}{7}\frac{3}{9} = \frac{8}{21}.$$

Vậy xác suất lấy được 2 viên bi cùng màu là $\frac{8}{21}$.

Ví dụ 2. Công ty A muốn ký một hợp đồng kinh tế với công ty B. Công ty A quyết định sẽ thương lượng các điều khoản hợp đồng với công ty B 2 lần, nếu không ký hợp đồng thành công sẽ chuyển sang đối tác khác. Xác suất để ký được hợp đồng ở lần thương lượng thứ 1 là 0,65. Nếu lần thương lượng thứ 1 không thành công thì xác suất để ký được hợp đồng ở lần thương lượng thứ 2 là 0,8. Tính xác suất để hai công ty A và B ký được hợp đồng với nhau.

Giải. Phép thử: Hai công ty A và B thương lượng để ký hợp đồng kinh tế. Gọi

A = "Hai công ty A và B ký được hợp đồng với nhau",

A_i= "Công ty A ký được hợp đồng với công ty B ở lần thương lượng thứ i", $i = 1,2$.

Ta có $A = A_1 + \overline{A}_1A_2$. Suy ra

$$p(A) = p(A_1 + \overline{A}_1A_2).$$

Khi đó, do A_1, $\overline{A}_1A_2$ xung khắc nhau nên

$$p(A) = p(A_1) + p(\overline{A}_1A_2) = p(A_1) + p(\overline{A}_1).p(A_2\overline{A}_1) = 0,65 + 0.35 \times 0,8 = 0,93.$$

Vậy xác suất để 2 công ty A và b ký hợp đồng thành công là $0,93$.

1.5. Công thức xác suất đầy đủ và công thức Bayes

1.5.1. Công thức xác suất đầy đủ

Xét về mặt hình học, một biến cố có thể xem như một tập hợp. Định nghĩa xác suất theo hình học chính là tỉ số của diện tích mỗi biến cố chia cho diện tích cả không gian mẫu. Như vậy, việc tính xác suất được mỗi biến cố coi như là tính diện tích các miền khác nhau. Một trong những ý tưởng để tính diện tích của một miền là chia nhỏ miền cần tính thành các miền nhỏ hơn, rời nhau và tính được, tính từng miền nhỏ và cuối cùng là lấy tổng. Hệ hai biến cố gồm biến cố A và biến cố đối của nó, $\{A, \overline{A}\}$, luôn chia không gian mẫu thành 2 phần rời nhau. Chúng ta cũng hoàn toàn chia nhỏ không gian mẫu thành 3, 4, 5,… phần rời nhau bởi hệ các biến cố nào đó. Một hệ biến cố như vậy được gọi là hệ đầy đủ các biến cố, gọi tắt là hệ đầy đủ.

Định nghĩa: Hệ các biến cố $\{A_1, A_2, ..., A_n\}$ được gọi là đầy đủ nếu chúng đôi một xung khắc và

$$A_1 + A_2 + ... + A_n = \Omega .$$

Định lý: Xét một phép thử, một hệ đầy đủ các biến cố $\{A_1, A_2, ..., A_n\}$ và một biến cố A. Khi đó, xác suất xảy ra biến cố A, tính theo hệ đầy đủ $\{A_1, A_2, ..., A_n\}$, là

$$P(A) = P(A_1).P(A \mid A_1) + P(A_2).P(A \mid A_2) + ... + P(A_n).P(A \mid A_n).$$

Ví dụ: Một nhà máy có ba phân xưởng 1, 2 và 3 cùng sản xuất một loại sản phẩm. Sản lượng của các phân xưởng 1, 2, 3 tương ứng là 50%, 35% và 15% tổng sản lượng của cả nhà máy cùng với tỉ lệ phế phẩm ở các phân xưởng tương ứng là 5%, 3% và 1%. Các sản phẩm sau khi làm ra sẽ được tập trung về nhà máy và đóng gói thành các kiện hàng lớn. Lấy ngẫu nhiên 1 sản phẩm từ kiện hàng để kiểm tra. Tính xác suất để

a) Sản phẩm lấy ra là phế phẩm.

b) Nếu sản phẩm lấy ra kiểm tra là phế phẩm thì theo anh/chị sản phẩm đó là của phân xưởng nào sản xuất.

Giải. Phép thử: Lấy ngẫu nhiên một sản phẩm từ kiện hàng. Gọi

A = "Sản phẩm lấy ra là phế phẩm",

A_i = "Sản phẩm lấy ra là do phân xưởng i sản suất", $i = 1, 2, 3$.

a) Hệ các biến cố $\{A_1, A_2, A_3\}$ là đầy đủ nên xác suất để sản phẩm lấy ra là phế phẩm là

$$P(A) = P(A_1).P(A \mid A_1) + P(A_2).P(A \mid A_2) + P(A_3).P(A \mid A_3)$$
$$= 0,5 \times 0,05 + 0,35 \times 0,02 + 0,15 \times 0,01 = 0,0335 .$$

b) $P(A_i \mid A)$: biết sản phẩm lấy ra là phế phẩm (A đã xảy ra), xác suất để phế phẩm đó là do phân xưởng thứ i sản xuất. Ta có

$$P(A_1 \mid A) = \frac{P(A.A_1)}{P(A)} = \frac{P(A_1.A)}{P(A)} = \frac{P(A_1).P(A \mid A_1)}{P(A)} = \frac{0,5 \times 0,05}{0,0335} = 0,746268... \approx 0,7463 .$$

Tính tương tự cho các xác suất còn lại. Kết luận.

1.5.2. Công thức Bayes

Theo Ví dụ ở mục 1.5.1.b), ta thấy rằng: sau khi lấy được một phế phẩm, với một người quản lý thì một câu hỏi được đặt ra là "phế phẩm đó của phân xưởng nào sản xuất?" Để trả lời

câu hỏi này, chúng ta tính lại các xác suất $P(A_i \mid A)$ như đã làm. Các xác suất $P(A_i \mid A)$ được gọi là các xác suất hậu nghiệm, là xác suất được điều chỉnh hoặc cập nhật lại sau khi một biến cố nào đó đã xảy ra (xem xét thông tin mới) và được tính bằng công thức Bayes như sau

Định lý 1.3. Xét một phép thử, một hệ đầy đủ các biến cố $\{A_1, A_2, ..., A_n\}$ và một biến cố A. Biết rằng biến cố A đã xảy ra. Khi đó

$$P(A_i \mid A) = \frac{P(A_i).P(A \mid A_i)}{P(A)} = \frac{P(A_i).P(A \mid A_i)}{P(A_1).P(A \mid A_1) + P(A_2).P(A \mid A_2) + ... + P(A_n).P(A \mid A_n)}.$$

Ví dụ: Hai máy tiện tự động cùng sản xuất một loại trục xe ô-tô như nhau. Các trục được đóng chung vào một kiện. Biết rằng, năng suất của máy tiện thứ 2 gấp đôi máy tiện thứ nhất. Máy tiện thứ nhất sản xuất được 64% trục loại tốt, còn máy tiện thứ hai sản xuất được 80% trục loại tốt. Lấy ngẫu nhiên một trục máy. Tính xác suất

a) Lấy được trục máy loại tốt

b) Biết rằng lấy được trục máy loại tốt. Khả năng trục máy loại tốt đó do máy tiện nào sản xuất là cao hơn.

Giải. Phép thử: Lấy ngẫu nhiên một trục máy từ kiện hàng. Gọi

A = "Trục máy lấy ra là trục loại tốt",

A_i = " Trục máy lấy ra là do máy tiện *i* sản suất", $i = 1,2$.

a) Hệ các biến cố $\{A_1, A_2\}$ là đầy đủ nên xác suất để trục máy lấy ra là trục loại tốt là

$$P(A) = P(A_1).P(A \mid A_1) + P(A_2).P(A \mid A_2)$$

$$= \frac{1}{3} \times 0,64 + \frac{2}{3} \times 0,8 = 0,7467.$$

b) $P(A_i \mid A)$: biết trục máy lấy ra là loại tốt (*A* đã xảy ra), xác suất để trục tốt đó là do máy tiện thứ *i* sản xuất là:

$$P(A_1 \mid A) = \frac{P(A.A_1)}{P(A)} = \frac{P(A_1.A)}{P(A)} = \frac{P(A_1).P(A \mid A_1)}{P(A)} = \frac{\frac{1}{3}. \times 0,64}{0,7467} = 0,2857,$$

$$P(A_2 \mid A) = \frac{P(A.A_2)}{P(A)} = \frac{P(A_2.A)}{P(A)} = \frac{P(A_2).P(A \mid A_2)}{P(A)} = \frac{\frac{2}{3} \times 0,8}{0,7467} = 0,7142.$$

Kết luận trục ô-tô loại tốt do máy tiện thứ 2 sản xuất là cao hơn.

1.6. Bài tập chương 1

1. Cho 3 xạ thủ, mỗi người bắn một phát vào một mục tiêu. Gọi A_i là biến cố người thứ $i\,(i = 1,2,3)$ bắn trúng. Hãy biểu diễn các biến cố sau qua các biến cố A_i

a) Có một người bắn trúng.

b) Có ít nhất một người bắn trúng.

c) Người thứ nhất bắn trúng.

d) Người thứ hai và ba cùng bắn trúng.

e) Người thứ nhất bắn trúng hoặc là người thứ hai và ba cùng bắn trúng.

f) Chỉ 2 người bắn trúng.

2. Trong 1 kho hàng có 10.000 sản phẩm với 500 phế phẩm. Lấy ngẫu nhiên 1 sản phẩm từ kho. Tìm xác suất để sản phẩm là phế phẩm. (Với phép thử lấy ngẫu nhiên một phần tử từ tổng thể. Xác suất để một phần tử ngẫu nhiên của tổng thể có tính chất nào đó sẽ bằng tỉ lệ có tính chất đó của tổng thể).

3. Một túi đựng 10 quả cầu, trong đó có 6 quả màu xanh và 4 quả màu vàng. Lấy ngẫu nhiên từ túi ra 3 quả cầu. Tìm xác suất để có 2 quả cầu xanh.

4. Một người cần gọi điện thoại nhưng quên mất hai chữ số cuối của số điện thoại cần gọi và chỉ nhớ là hai chữ số đó khác nhau. Ông bấm số điện thoại với 2 chữ số cuối là ngẫu nhiên theo cách nhớ. Tìm xác suất để ông gọi trúng ngay số điện thoại cần gọi.

5. Một khách sạn có 6 phòng đơn. Có 10 khách đến thuê phòng, trong đó có 6 nam và 4 nữ. Người quản lý chọn ngẫu nhiên 6 người. Tìm xác suất để:

a) Cả 6 người đều là nam.

b) Có 4 nam và 2 nữ.

c) Có ít nhất 2 nữ.

6. Một hộp có 10 sản phẩm, trong đó có 3 phế phẩm và 7 chính phẩm. Lấy ngẫu nhiên từ hộp ra 5 sản phẩm.Tính xác suất để:

a) Không có phế phẩm nào.

b) Có không quá 1 phế phẩm.

c) Có ít nhất 1 phế phẩm.

7. Một công ty có 60 nhân viên, trong đó có 20 nam và 40 nữ. Tỷ lệ nhân viên nữ có thể nói tiếng Anh lưu loát là 15% và tỷ lệ này đối với nam là 20%.

a) Gặp ngẫu nhiên một nhân viên của công ty. Tìm xác suất để gặp được nhân viên nói tiếng Anh lưu loát?

b) Gặp ngẫu nhiên hai nhân viên của công ty. Tìm xác suất để có ít nhất một người nói tiếng Anh lưu loát trong số 2 người gặp?.

8. Theo khảo sát tổ chức y tế WHO trong một vùng dân cư, tỉ lệ người mắc bệnh tim là 9%, bệnh huyết áp là 12% và mắc cả hai bệnh là 7%. Chọn ngẫu nhiên một người trong vùng. Tìm xác suất để người đó không mắc bệnh nào trong 2 bệnh.

9. Trong lớp 100 sinh viên, trong đó có 20 em giỏi môn Toán, 25 em giỏi Ngoại ngữ và có 10 em giỏi cả Toán lẫn Ngoại ngữ. Quy định giỏi ít nhất một môn thì được thưởng. Chọn ngẫu nhiên một em trong lớp. Tìm xác suất để em đó được thưởng. Suy ra tỉ lệ học sinh được thưởng của lớp.

10. Một khách sạn có 3 thang máy 1, 2, 3 hoạt động độc lập với nhau. Xác suất để thang máy 1, 2, 3 bị hỏng lần lượt là 0,4; 0,5 và 0,6. Tính xác suất để:

a) Chỉ có duy nhất một thang máy bị hỏng.

b) Ít nhất một thang máy bị hỏng.

c) Biết chỉ có một thang máy bị hỏng, tìm xác suất để đó là thang máy 1.

11. Một hãng vận tải cho 3 xe hoạt động độc lập trong năm năm. Xác suất các xe 1, 2, 3 bị hỏng tương ứng là 0,1; 0,2 và 0,15. Tính xác suất để:

a) Có một xe bị hỏng.

b) Có ít nhất một xe bị hỏng.

c) Biết chỉ có một xe hỏng, tìm xác suất để đó là xe 2.

12. Chị Lan có một chùm chìa khóa gồm 9 chiếc bề ngoài rất giống nhau nhưng trong đó chỉ có 2 chiếc mở được cửa tủ. Chị Lan thử ngẫu nhiên từng chìa và chìa nào không đúng thì bỏ ra. Tìm xác suất để chị Lan mở được cửa ở lần thử thứ 4.

13. Một bộ đề thi vấn đáp gồm 10 đề, trong đó có 4 đề về câu hỏi lý thuyết và 6 đề bài tập tính toán. Có 3 sinh viên lần lượt vào thi, mỗi sinh viên chỉ lấy một đề và không hoàn lại. Tìm xác suất để sinh viên 1 gặp đề bài tập và sinh viên 2 gặp đề lý thuyết và sinh viên 3 gặp đề bài tập.

14. Một cơ sở sản xuất mũ gồm có 3 tổ cùng sản xuất với tỉ lệ sản phẩm trong tổng số sản phẩm lần lượt là 20%, 30% và 50%. Tổ 1 có tỉ lệ phế phẩm là 5% sản phẩm tổ, tổ 2 là 2% và tổ 3 là 1% . Tất cả sản phẩm làm ra được xếp chung vào một kho. Lấy ngẫu nhiên một sản phẩm từ kho.

a) Tìm xác suất để sản phẩm đó là phế phẩm. Tỉ lệ phế phẩm của kho là bao nhiêu?

b) Biết sản phẩm là phế phẩm. Tìm xác suất để nó do tổ 2 sản xuất.

15. Cho tỉ lệ người dân nghiện thuốc lá ở một vùng là 30%. Biết tỉ lệ người viêm họng trong số người nghiện thuốc lá là 60% và tỉ lệ người viêm họng trong số người không hút thuốc là 40%. Chọn ngẫu nhiên một người trong vùng.

a) Giả sử người đó viêm họng. Tìm xác suất để người đó nghiện thuốc.

b) Giả sử người đó không viêm họng. Tìm xác suất để người đó nghiện thuốc.

Bài tập làm thêm

16. Điền các giá trị thích hợp vào ô trống

$P(A)$	$P(B)$	$P(A \cup B)$	$P(A \cap B)$	$P(A \mid B)$	$P(B \mid A)$
$\frac{3}{4}$		$\frac{9}{10}$	$\frac{1}{5}$		
			$\frac{2}{3}$	$\frac{8}{9}$	$\frac{4}{5}$
$\frac{1}{5}$	$\frac{1}{5}$				$\frac{1}{20}$
$\frac{5}{17}$	$\frac{3}{17}$		$\frac{1}{17}$		

17. Một nhân viên bán hàng mỗi năm đến bán hàng ở một công ty nọ. Xác suất để lần đầu bán được hàng là 0,8. Nếu lần trước bán được hàng thì xác suất để lần sau bán được hàng là 0,9; còn nếu lần trước không bán được hàng thì xác suất để lần sau bán được hàng là 0,4.

a) Tìm xác suất để cả ba lần đều bán được hàng.

b) Tìm xác suất để có đúng hai lần bán được hàng

18. Một test kiểm tra sự hiện diện của virus Covid19 cho kết quả dương tính nếu bệnh nhân thực sự nhiễm virus Covid19. Tuy nhiên test này cũng có sai sót, đôi khi cho kết quả dương tính đối với người không thực sự nhiễm vius, tỷ lệ sai sót là 1/20000. Giả sử cứ 10000 người thì có 1 người bị nhiễm vius H5N1. Tìm tỷ lệ người có kết quả dương tính thực sự nhiễm Covid19.

19. Người ta phỏng vấn ngẫu nhiên 500 khách hàng về một sản phẩm định đưa ra thị trường và thấy có: 100 người trả lời "sẽ mua"; 150 người trả lời " có thể sẽ mua" và 250 người trả lời "không mua". Theo kinh nghiệm cho thấy tỷ lệ khách hàng thực sự mua sản phẩm tương ứng với những cách trả lời trên là 40%; 20% và 1%.

a) Hãy đánh giá thị trường tiềm năng của sản phẩm đó (theo nghĩa tỷ lệ người thực sự mua sản phẩm đó)

b) Trong số khách hàng thực sự mua sản phẩm đó có bao nhiêu phần trăm trả lời "không mua" ? .

20. Trước tình hình dịch bệnh Covid-19 phức tạp, Nhà trường tổ chức thi học kỳ II, năm 2021 môn Xác suất thống kê, hình thức thi: online với hai bài thi là tự luận và vấn đáp trực tuyến. Giả sử, khả năng thi đạt bài thi tự luận của sinh viên An là 0,8. Nếu sinh viên An thi đạt bài thi tự luận thì khả năng An thi đạt bài thi vấn đáp trực tuyến là 0,9, còn nếu An không đạt bài thi tự luận thì khả năng để An đạt bài thi vấn đáp trực tuyến là 0,3. Tính xác suất để cho các biến cố sau: A " An thi đạt cả hai bài thi"; B: "An chỉ thi đạt một bài thi".

Chương 2.
Đại lượng ngẫu nhiên

Chương 2 trang bị cho người học các khái niệm và tính chất về đại lượng ngẫu nhiên cùng với các tham số đặc trưng và một số qui luật phân phối xác suất thông dụng của nó.

Sau khi học chương này, sinh viên có thể lập được bảng phân phối xác suất của đại lượng ngẫu nhiên rời rạc, vận dụng được hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục; tính được các tham số đặc trưng và giải thích được ý nghĩa của chúng; áp dụng được các qui luật phân phối xác suất như Nhị thức, Poisson, Chuẩn,... để tính xác suất và các tham số trong các tình huống cụ thể.

2.1. Đại lượng ngẫu nhiên (Random variable)

2.1.1. Khái niệm

Định nghĩa: Xét một phép thử ngẫu nhiên. Khi đó ánh xạ

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

được gọi là đại lượng (biến) ngẫu nhiên.

Vậy có thể hiểu đại lượng ngẫu nhiên được mô tả như một qui tắc biểu diễn các kết quả của một phép thử ngẫu nhiên nào đó dưới dạng số.

Ví dụ: Gieo đồng thời 2 đồng xu cân đối, đồng chất. Khi đó, ta có các biến cố sơ cấp là

$$\omega_1 = (HH), \omega_2 = (HT), \omega_3 = (TH), \omega_4 = (TT),$$

trong đó *H*: mặt sấp (Head), *T*: mặt ngửa (Tail).

Hình 2.1.Đồng xu

Nếu gọi X là số đồng xu xuất hiện mặt ngửa thì X nhận các giá trị sau

$$X(\omega_1) = 0, X(\omega_2) = 1, X(\omega_3) = 1, X(\omega_4) = 2.$$

Đại lượng X trong trường hợp này được gọi là đại lượng ngẫu nhiên.

2.1.2. Phân loại đại lượng ngẫu nhiên

Tùy theo giá trị mà đại lượng ngẫu nhiên nhận được, đại lượng ngẫu nhiên được chia làm 2 loại: đại lượng ngẫu nhiên rời rạc và đại lượng ngẫu nhiên liên tục.

Định nghĩa 1 (Đại lượng ngẫu nhiên rời rạc - Discrete random variable). Đại lượng ngẫu nhiên X được gọi là rời rạc (ĐLNNRR) nếu tập hợp các giá trị mà nó có thể nhận là tập hữu hạn hoặc vô hạn đếm được.

Ví dụ 1. Các đại lượng X sau đây là rời rạc:

- Số sinh viên đi học tại một thời điểm nào đó trong ngày.
- Số sản phẩm kém chất lượng trong một lô hàng.
- Số con trong một gia đình.
- Số cuộc điện thoại mà bạn nhận được trong khoảng thời gian nào đó.

- Số câu trả lời đúng khi bạn đánh hú họa bài thi trắc nghiệm Anh văn gồm 50 câu.
- Số con cá câu được trong khoảng thời gian nào đó.

Định nghĩa 2 (Đại lượng ngẫu nhiên liên tục - Continuous random variable). Đại lượng ngẫu nhiên X được gọi là liên tục (ĐLNNLT) nếu tập hợp các giá trị mà nó có thể nhận là một khoảng nào đó trên $\mathbb{R}$.

Ví dụ 2. Các đại lượng X sau đây là liên tục:

- Nhiệt độ không khí tại một thời điểm nào đó trong ngày.
- Thời gian hoạt động bình thường của một loại bóng đèn điện tử.
- Chiều cao của người trưởng thành Việt nam giai đoạn 2011-2021.

2.2. Phân phối xác suất của đại lượng ngẫu nhiên

Định nghĩa: Một hệ thức cho phép biểu diễn mối quan hệ giữa các giá trị có thể có của đại lượng ngẫu nhiên với xác suất tương ứng với các giá trị đó được gọi là qui luật phân phối xác suất của đại lượng ngẫu nhiên.

2.2.1. Bảng phân phối xác suất (Probability distribution table)

Để mô tả đại lượng ngẫu nhiên rời rạc X nhận giá trị nào đó tương ứng với xác suất là bao nhiêu, ta dùng bảng phân phối xác suất. Bảng phân phối xác suất có dạng

X	x_1	x_2	...	x_n	Σ
P	p_1	p_2	...	p_n	1

Trong đó $x_i, i=\overline{1,n}$ là các giá trị của đại lượng ngẫu nhiên X thỏa

$$x_1 < x_2 < \cdots < x_n$$

và $P(X=x_i)=p_i>0,\ i=\overline{1,n}$ thỏa $\sum_{i=1}^{n} p_i = 1$.

Ví dụ 1. Bảng phân phối xác suất của đại lượng ngẫu nhiên X là tổng số chấm xuất hiện khi tung đồng thời hai con xúc sắc là

X	2	3	4	5	6	7	8	9	10	11	12	Σ
P	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Ví dụ 2. Có một hộp kín đựng 12 sản phẩm (có 8 chính phẩm). Lấy ngẫu nhiên (không hoàn lại) từ hộp ra 2 sản phẩm. Bảng phân phối xác suất của số chính phẩm được lấy ra X là

X	0	1	2	Σ
P	$\frac{C_4^2}{C_{12}^2}=\frac{3}{33}$	$\frac{C_8^1\times C_4^1}{C_{12}^2}=\frac{16}{33}$	$\frac{C_2^2}{C_{12}^2}=\frac{14}{33}$	1

Ví dụ 3. Một công ty có 3 loại xe ôtô được sử dụng độc lập: 1 xe 50 chỗ ngồi, 1 xe 4 chỗ và 1 xe tải. Xác suất để trong một ngày làm việc các xe được công ty sử dụng lần lượt là 0,4; 0,8; 0,9. Lập bảng phân phối xác suất cho số loại xe được công ty sử dụng trong ngày.

Giải. Gọi X là số loại xe được công ty sử dụng trong ngày. Ta có tập giá trị

$$X(\Omega)=\{0,1,2,3\}.$$

Gọi C_{50}, C_4, T lần lượt là biến cố xe 50 chỗ, xe 4 chỗ, xe tải được công ty sử dụng trong ngày. Khi đó, các biến cố C_{50}, C_4, T là độc lập và

$$p(C_{50}) = 0,4;\ p(C_4) = 0,8;\ \ p(T) = 0,9\,.$$

Ta tính được các xác suất

$$p(X=3) = p(C_{50}C_4T) = 0.4 \times 0.8 \times 0.9 = 0.288 \text{ (do các biến cố trên là độc lập).}$$

$$\begin{aligned} p(X=2) &= p(C_{50}C_4\overline{T} \cup C_{50}\overline{C}_4T \cup \overline{\overline{C}_{50}}C_4T) \\ &= p(C_{50}C_4\overline{T}) + p(C_{50}\overline{C}_4T) + p(\overline{C}_{50}C_4T) = 0,536 \end{aligned}$$

Tương tự:

$$p(X=0) = 0,012;\ p(X=1) = 0,164\,.$$

Vậy bảng phân phối xác suất của số loại xe được sử dụng trong ngày của công ty là

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

2.2.2. Hàm phân phối xác suất (Cumulative distribution function)

Định nghĩa: Hàm phân phối xác suất của đại lượng ngẫu nhiên X, ký hiệu $F(x)$ và được định nghĩa như sau

$$F(x) = p(X < x),\ \ \forall x \in \mathbb{R}.$$

Ví dụ 1. Cho bảng phân phối xác suất

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

Khi đó, hàm phân phối xác suất của X là

$$F(x) = p(X < x) = \begin{cases} 0 & x \le 0 \\ 0,012 & 0 < x \le 1 \\ 0,176 & 1 < x \le 2 \\ 0,712 & 2 < x \le 3 \\ 1 & 3 < x \end{cases}.$$

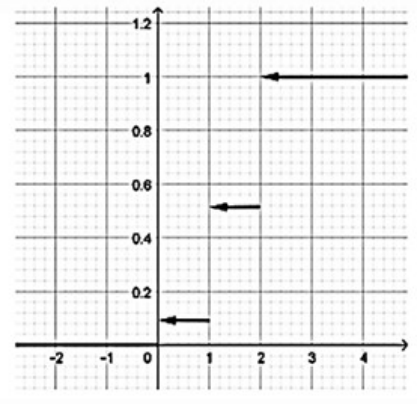
Ví dụ 2. Tìm hàm phân phối xác suất trong Ví dụ 2-Mục 2.2.1 và tính xác suất $p(1 \le X < 3)$.

Giải. Ta có

$$F(x) = p(X < x) = \begin{cases} 0 & x \le 0 \\ 3\,/\,33 & 0 < x \le 1 \\ 19\,/\,33 & 1 < x \le 2 \\ 1 & 2 < x \end{cases}.$$

Ta có

$$p(1 \le X < 3) = p(X=1) + p(X=2) = \frac{16}{33} + \frac{14}{33} = \frac{30}{33}.$$



Hình 2.2. Đồ thị của hàm phân phối xác suất

Tính chất.

a) $0 \le F(x) \le 1,\ \forall x \in \mathbb{R}$ và $F(-\infty) = 0,\ F(+\infty) = 1$.

b) $F(x)$ là hàm không giảm trên $\mathbb{R}$.

c) $F(x)$ là hàm liên tục trái mọi $x \in \mathbb{R}$. Trong trường hợp X là đại lượng ngẫu nhiên liên tục thì $F(x)$ là hàm liên tục trên $\mathbb{R}$.

d) $p(a \le X < b) = F(b) - F(a)$.

Hệ quả. Nếu X là đại lượng ngẫu nhiên liên tục thì

a) $p(X = a) = 0,\ \forall a \in \mathbb{R}$.

b) $p(a \le X \le b) = p(a \le X < b) = p(a < X \le b) = p(a < X < b)$.

Chứng minh.

a) Thật vậy

$$p(X = a) = \lim_{\Delta x \to 0} p(a \le X < a + \Delta x)$$
$$= \lim_{\Delta x \to 0} \Big[F(a + \Delta x) - F(a) \Big] = F(a) - F(a) = 0,\ \forall a \in \mathbb{R}.$$

b) Suy từ a).

2.2.3. Hàm mật độ xác suất (Probability density function)

Định nghĩa: Cho X là đại lượng ngẫu nhiên liên tục có hàm phân phối xác suất $F(x)$ khả vi, khi đó hàm mật độ xác suất của X, ký hiệu là $f(x)$, được định nghĩa

$$f(x) = F'(x),\ \ x \in \mathbb{R}.$$

Tính chất.

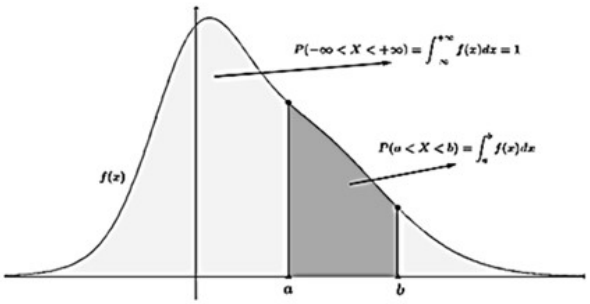
a) $f(x) \ge 0,\ \forall x \in \mathbb{R}$.

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

c) $\int_{-\infty}^{+\infty} f(x)dx = 1$.

d) $p(a \le X \le b) = p(a \le X < b) = p(a < X \le b) = p(a < X < b) = \int_{a}^{b} f(x)dx$.

Hàm mật độ xác suất của đại lượng ngẫu nhiên X, cho ta hình ảnh của sự tập trung xác suất của đại lượng ngẫu X trên từng khoảng giá trị của nó (như hình 2.3).



Hình 2.3 Hàm mật độ xác suất.

Ví dụ 1. Thời gian $T(m)$ để một khách hàng chờ thanh toán tiền ở một siêu thị vào dịp Tết là đại lượng ngẫu nhiên liên tục có hàm phân phối xác suất

$$F(t) = \begin{cases} 0 & t \le 0 \\ kt^2 & 0 < t < 20,\ k \in \mathbb{R} \\ 1 & 20 \le t \end{cases}.$$

a) Tìm k.

b) Tìm hàm mật độ $f(x)$.

c) Tính $p(T < 10)$.

Giải.

a) Do hàm phân phối xác suất liên tục trái nên

$$\lim_{t \to 20^-} F(t) = F(20) \Leftrightarrow \lim_{t \to 20^-} kt^2 = 1 \Rightarrow k = \frac{1}{400}.$$

b) Ta có hàm mật độ

$$f(t) = F'(t) = \begin{cases} 0 & t \notin (0,20) \\ \dfrac{1}{200}t & 0 < t < 20 \end{cases}.$$

c) Ta có

$$p(T < 10) = F(10) = 0,25\ .$$

Ví dụ 2. Giả sử X (ngày) là tuổi thọ của một loại sản phẩm do công ty sản xuất là biến ngẫu nhiên liên tục có hàm mật độ

$$f(d) = \begin{cases} \dfrac{100}{d^2} & d \ge 100 \\ 0 & d < 100 \end{cases}.$$

a) Tìm hàm phân phối xác suất.

b) Sản phẩm được bảo hành nếu tuổi thọ của nó dưới 120 ngày. Tính tỉ lệ sản phẩm của công ty phải bảo hành.

Giải.

a) Ta có

$$F(d) = \int_{-\infty}^{d} f(x)dx = \begin{cases} 0 & d < 100 \\ \int_{100}^{d} \dfrac{100}{x^2}dx & d \ge 100 \end{cases} = \begin{cases} 0 & d < 100 \\ 1 - \dfrac{100}{d} & d \ge 100 \end{cases}.$$

b) Tỉ lệ sản phẩm phải bảo hành là

$$p(d < 120) = F(120) = \frac{1}{6} \simeq 0{,}167 .$$

Vậy có 16.7% sản phẩm của công ty được bảo hành.

2.3. Các tham số đặc trưng của đại lượng ngẫu nhiên

2.3.1. Kỳ vọng toán (Expectation)

Định nghĩa: Kỳ vọng của đại lượng ngẫu nhiên X, ký hiệu $E(X)$, được định nghĩa

$$E(X) := \begin{cases} \sum_{i=1}^{n} x_i p_i & \text{nÕu } X \text{ lµ § LNNRR} \\ \int_{-\infty}^{+\infty} x f(x)\,dx & \text{nÕu } X \text{ lµ § LNNLT} \end{cases}.$$

Ví dụ 1. Hai người A và B tham gia một trò chơi: Có 1 hộp kín đựng 5 quả cầu (4 quả màu đỏ, 1 quả màu đen); A và B lần lượt lấy ngẫu nhiên (không hoàn lại) một quả cầu cho đến khi người nào lấy trúng quả cầu màu đen thì ván chơi kết thúc và người lấy trúng quả đen phải trả cho người còn lại số tiền bằng số bi lấy ra nhân với 10.000 đồng. Giả sử A là người lấy quả cầu trước, tính số tiền kỳ vọng A được sau một ván chơi.

Giải. Gọi X là số tiền A thu về sau ván chơi. Khi đó X là đại lượng ngẫu nhiên rời rạc. Tập giá trị

$$X(\Omega) = \{-50.000, -30.000, -10.000, +20.000, +40.000\}.$$

Ta có bảng phân phối xác suất

X	-50.000	-30.000	-10.000	20.000	40.000	Σ
P	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	1

Vì nếu gọi D_i, $(i = 1,2,3,4,5)$ là biến cố ở lần lấy thứ i lấy trúng quả cầu màu đỏ. Ta có

$$p(X = -10.000) = p(\overline{D}_1) = \frac{1}{5}.$$

$$p(X = -50.000) = p(D_1 D_2 D_3 D_4 \overline{D}_5) = \frac{4}{5}.\frac{3}{4}.\frac{2}{3}.\frac{1}{2}.1 = \frac{1}{5}.$$

Tương tự ta được các kết quả cho ở bảng trên. Khi đó, số tiền kỳ vọng A được sau ván chơi là

$$E(X) = \frac{1}{5}\left\{(-50.000) + (-30.000) + (-10.000) + 20.000 + 40.000\right\} = -\frac{30.000}{5} = -6.000 .$$

Do $E(X) < 0$ nên người A bị thiệt trong trò chơi này.

Ví dụ 2. Thời gian $T(m)$ để một khách hàng chờ thanh toán tiền ở một siêu thị vào dịp Tết là đại lượng ngẫu nhiên liên tục có hàm phân phối xác suất

$$F(t) = \begin{cases} 0 & t \le 0 \\ kt^2 & 0 < t < 20,\ k \in \mathbb{R} \\ 1 & 20 \le t \end{cases}.$$

Tính thời gian trung bình mà một khách hàng chờ thanh toán tiền ở siêu thị vào dịp Tết.

Giải. Theo Ví dụ 1-Mục 2.2.3, hàm mật độ là

$$f(t) = F'(t) = \begin{cases} 0 & t \notin (0,20) \\ \dfrac{1}{200}t & 0 < t < 20 \end{cases}.$$

nên

$$E(X) = \int_{-\infty}^{+\infty} t.f(t)dt = \int_{0}^{20} t.\frac{1}{200}.tdt = 13,33\,.$$

Vậy trung bình một khách hàng phải chờ thanh toán tiền ở siêu thị vào dịp Tết là 13,33 (phút).

Tính chất.

a) $E(C) = C$, C là hằng số.

b) $E(CX) = CE(X)$, C là hằng số.

c) $E(X+Y) = E(X) + E(Y)$.

d) $E(XY) = E(X) \times E(Y)$ nếu X, Y là các ĐLNN độc lập.

Ý nghĩa. Kỳ vọng $E(X)$ chính là giá trị trung bình (theo xác suất) của đại lượng ngẫu nhiên X khi phép thử tương ứng được thực hiện.

2.3.2. Phương sai (Variance)

Định nghĩa: Phương sai của đại lượng ngẫu nhiên X, ký hiệu là $D(X)$ hay $Var(X)$, được định nghĩa

$$D(X) := E\Big[X - E(X)\Big]^2.$$

Trong thực hành, ta thường dùng công thức sau để tính phương sai

$$D(X) = E(X^2) - [E(X)]^2$$

với

$$E(X^2) := \begin{cases} \sum_{i=1}^{n} x_i^2 p_i & \text{nếu } X \text{ là ĐLNNRR} \\ \int_{-\infty}^{+\infty} x^2 f(x)\,dx & \text{nếu } X \text{ là ĐLNNLT} \end{cases}.$$

Thật vậy, ta có

$$\begin{aligned} D(X) &= E\Big[X - E(X)\Big]^2 = E\Big[X^2 - 2X.E(X) + [E(X)]^2\Big] \\ &= E(X^2) - 2E(X)E(X) + [E(X)]^2 \text{ (do tính chất của kỳ vọng)} \\ &= E(X^2) - [E(X)]^2\,. \end{aligned}$$

Nhận xét. Phương sai của ĐLNN X có đơn vị đo là đơn vị của X bình phương. Để có một đại lượng có cùng đơn vị đo với X, người ta dùng khái niệm độ lệch chuẩn (standard deviation) $\sigma(X)$ được định nghĩa

$$\sigma(X) := \sqrt{D(X)}\,.$$

Tính chất

a) $D(X) \geq 0$.

b) $D(C) = 0$ với C là hằng số.

c) $D(CX) = C^2 D(X)$.

d) $D(X \pm Y) = D(X) + D(Y)$ nếu X và Y là độc lập.

Ý nghĩa. Phương sai là thông số đo độ phân tán các giá trị của ĐLNN X xung quanh giá trị trung bình. Trong kỹ thuật phương sai đặc trưng cho sai số của các thiết bị hoặc các phép đo, trong kinh tế nó đặc trưng cho mức độ rủi ro của các quyết định.

Ví dụ 1. Cho ĐLNN X có bảng phân phối xác suất

X	1	2	a
P	b	0,3	0,5

Tìm a,b biết $E(X)=2,3$ và tính $D(X), D(2X-3)$.

Giải. Ta có

$$\begin{cases} \sum_{i=1}^{3} p_i = 1 \\ E(X) = \sum_{i=1}^{3} x_i p_i = 2,3 \end{cases} \Leftrightarrow \begin{cases} b + 0,3 + 0,5 = 1 \\ b + 0,6 + 0,5a = 2,3 \end{cases} \Leftrightarrow \begin{cases} a = 3 \\ b = 0,2 \end{cases}.$$

và

$$D(X) = E(X^2) - [E(X)]^2 = \{1^2 \times 0,2 + 2^2 \times 0,3 + 3^2 \times 0,5\} - (2,3)^2$$
$$= 5,9 - 5,29 = 0,61,$$

$D(2X-3) = 4D(X) = 4 \times 0,61 = 2,44$ (tính chất của phương sai).

Ví dụ 2. Cho X và Y tương ứng là các ĐLNN độc lập chỉ lợi nhuận (%/năm) khi đầu tư vào hai ngành A và B nào đó. Giả sử

$$E(X) = 12(\%),\ D(X) = 25(\%)^2,\ E(Y) = 14(\%),\ D(Y) = 36(\%)^2.$$

Một người đầu tư vào cả hai ngành A và B thì người đó cần lựa chọn tỷ lệ đầu tư như thế nào để ít rủi ro nhất.

Giải. Gọi p là tỉ lệ đầu tư vào ngành A với $p \in [0,1]$. Khi đó $1-p$ là tỉ lệ đầu tư vào ngành B. Gọi π lợi nhuận của phương án đầu tư này, ta có

$$\pi = pX + (1-p)Y.$$

Suy ra

$$D(\pi) = D[pX + (1-p)Y] = p^2 D(X) + (1-p)^2 D(Y)$$
$$= 25p^2 + 36(1-p)^2 = 61p^2 - 72p + 36.$$

Để phương án đầu tư gặp ít rủi ro nhất thì

$$D(\pi) \to \min \Leftrightarrow p = \frac{36}{61} = 0,59 = 59\%.$$

Vậy người đầu tư nên chọn phương án đầu tư vào ngành A với tỉ lệ 59% và ngành B tỉ lệ 41% thì độ rủi ro là thấp nhất.

2.3.3. Giá trị tin chắc nhất (ModX)

Định nghĩa: Ta gọi $Mod(X)$ là giá trị của ĐLNN X ứng với xác suất lớn nhất với ĐLNN rời rạc hoặc ứng với giá trị làm cho hàm mật độ cực đại với ĐLNN liên tục. Ta cũng nói $Mod(X)$ là giá trị tin chắc nhất.

Nhận xét. Một ĐLNN có thể có một hoặc nhiều $Mod(X)$, chẳng hạn, nếu X là điểm thi xác suất thống kê của sinh viên NTU thì $Mod(X)$ là điểm thi mà nhiều sinh viên nhận được nhất.

2.4. Các phân phối xác suất thông dụng

2.4.1. Phân phối nhị thức (Binomial distribution)

Định nghĩa: Tiến hành n phép thử độc lập, với xác suất xuất hiện biến cố A trong mỗi phép thử là như nhau và bằng p. Ta gọi n phép thử này là n phép thử Bernoulli.

Gọi X là số lần xuất hiện biến cố A trong n phép thử Bernoulli, khi đó X là ĐLNNRR có tập giá trị là $X(\Omega) = \{0,1,\ldots,n\}$ và

$$p(X = k) := p_n(k) = C_n^k p^k (1-p)^{n-k},\ k = 0,1,..,n.$$

ĐLNN X được gọi là có phân phối nhị thức với hai tham số n và p, ký hiệu $X \sim B(n,p)$.

Nhận xét. Điều kiện để ĐLNN có phân phối nhị thức:

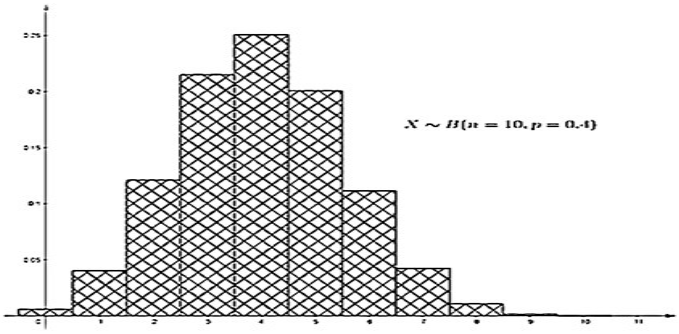
- Số phép thử cố định (n).
- Các phép thử là độc lập (kết quả của các phép thử không ảnh hưởng lẫn nhau).
- Kết quả của mỗi phép thử chỉ có hai biến cố A hoặc $\overline{A}$.
- Xác suất để biến cố A xảy ra trong mỗi phép thử là như nhau và bằng p.

Tính chất. Nếu $X \sim B(n,p)$ thì

a) $E(X) = np$.

b) $D(X) = np(1-p)$.

c) $np + p - 1 \le Mod(X) \le np + p$



Hình 2.4. Đồ thị phân phối nhị thức với $n = 10;\ p = 0,4$.

Ví dụ 1. Trong một nhà máy sản xuất chip nhớ, biết rằng những cuộc thử nghiệm kiểm tra trước đó cho thấy tỉ lệ chíp không đạt chất lượng của nhà máy là 20%. Kiểm tra ngẫu nhiên 15 chip. Tính xác suất

a) Có đúng 7 chip không đạt chất lượng.

b) Có ít nhất 2 chip không đạt chất lượng.

Giải. Phép kiểm tra trên thỏa mãn phép thử Bernoulli với $n = 15,\ p = 0,2$. Gọi X số chip không đạt chất lượng trong số 15 chip được kiểm tra ngẫu nhiên. Suy ra $X \sim B(15;0,2)$. Ta có

a) $p(X = 7) = p_{15}(7) = C_{15}^7 (0,2)^7 (0,8)^8 = 0,0138$.

b) $p(X \ge 2) = 1 - p(X < 2) = 1 - p(X = 0) - p(X = 1)$

$= 1 - p_{15}(0) - p_{15}(1)$

$= 1 - C_{15}^0 (0,2)^0 (0,8)^{15} - C_{15}^1 (0,2)^1 (0,8)^{14} = 0,8329$.

Ví dụ 2. Sinh viên Bình thi vấn đáp trả lời 5 câu hỏi một cách độc lập. Giả sử, khả năng để Bình trả lời đúng mỗi câu là như nhau và bằng 0,6. Mỗi câu trả lời đúng thì Bình sẽ được 4 điểm, mỗi câu trả lời sai thì bị trừ 2 điểm. Tính xác suất để

a) Bình trả lời đúng 3 câu hỏi.

b) Tính số điểm trung bình mà Bình đạt được.

c) Sinh viên An cũng vào thi vấn đáp với khả năng trả lời đúng mỗi câu là như nhau và cho rằng điểm thi trung bình của An sẽ đạt được ít nhất 14 điểm. Hỏi khả năng An trả lời đúng mỗi câu tối thiểu là bao nhiêu?.

Giải. Bình trả lời thi vấn đáp như trên thỏa mãn phép thử Bernoulli với $n = 5;\ p = 0,6$. Gọi X, Y tương ứng là số câu trả lời đúng và điểm số của Bình. Suy ra $X \sim B(5;\ 0,6)$.

a) Xác suất để Bình trả lời đúng 3 câu hỏi là

$$p(X = 3) = p_5(3) = C_5^3(0,6)^3(0,4)^2 = 0,3456.$$

b) Ta có điểm của Bình là

$$Y = 4X + (5 - X)(-2) = 6X - 10.$$

Suy ra điểm trung bình Bình đạt được là

$$E(Y) = E(6X - 10) = 6E(X) - 10 = 6 \times n \times p - 10 = 6 \times 5 \times 0,6 - 10 = 8 \text{ (điểm).}$$

c) Gọi p là khả năng trả lời đúng mỗi câu của An và Z, T tương ứng là số câu trả lời đúng và điểm số của An. Ta thấy $Z \sim B(5; p)$ và $T = 6Z - 10$. Ta có

$$E(T) \geq 14 \Leftrightarrow 6 \times 5 \times p - 10 \geq 14 \Leftrightarrow p \geq 0,8.$$

Vậy để đạt điểm trung bình ít nhất 14 điểm thì khả năng trả lời đúng mỗi câu của An là phải ít nhất 0,8.

2.4.2. Phân phối Poisson (Poisson distribution)

Định nghĩa: ĐLNN X được gọi là có phân phối Poisson với tham số λ, ký hiệu $X \sim P(\lambda)$, nếu

$$p(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0,1,\ldots,n,\ldots$$

Tính chất. Cho $X \sim P(\lambda)$. Khi đó

a) $EX = DX = \lambda$.

b) $\lambda - 1 \leq Mod(X) \leq \lambda$.

Nhận xét. Phân phối Poisson thường được biết như phân phối xác suất của các biến cố hiếm trong một đơn vị thời gian (hoặc không gian) như số tai nạn giao thông tại một giao lộ, số trận động đất, số người chết do bị ngựa đá,... trong một năm. Đây là ĐLNN rời rạc nhưng lại được xét trong một khoảng thời gian (hoặc không gian) liên tục nhưng hữu hạn.

Định lý 2.1. Cho $X \sim B(n, p)$, khi n khá lớn $(n \geq 1000)$ và p khá nhỏ $(np \leq 10)$ thì

$$p(X = k) \approx e^{-\lambda}\frac{\lambda^k}{k!},\ \lambda = np.$$

Điều này có nghĩa là khi n khá lớn và p khá nhỏ thì phân phối nhị thức xấp xỉ phân phối Poisson với tham số $\lambda = np$.

Ví dụ 1. Quan sát tại siêu thị A thấy trung bình 4 phút có 20 khách đến mua hàng. Tính xác suất để

a) Trong 7 phút có 30 khách đến mua hàng ở siêu thị.

b) Trong 2 phút có từ 3 đến 5 khách đến mua hàng ở siêu thị.

Giải.

a) Gọi X là số khách đến siêu thị A trong 7 phút. Khi đó

$$X \sim P(\lambda),\ \lambda = 20 \times \frac{7}{4} = 35\,.$$

Suy ra

$$p(X = 30) = e^{-35}\,\frac{35^{30}}{30!} = 0,0499\,.$$

b) Gọi Y là số khách đến siêu thị A trong 2 phút. Khi đó

$$Y \sim P(\lambda),\ \lambda = 20 \times \frac{2}{4} = 10\,.$$

Suy ra

$$p(3 \le Y \le 5) = p(3) + p(4) + p(5) = e^{-10}\left(\frac{10^3}{3!} + \frac{10^4}{4!} + \frac{10^5}{5!}\right) = 0,0643\,.$$

Ví dụ 2. Khả năng mỗi trang giấy bị lỗi do in ấn là 0,002. Tính khả năng trong một quyển sách có 1.000 trang có nhiều nhất 2 trang bị lỗi.

Giải. Gọi X là số trang sách bị lỗi trong quyển sách có 1.000 trang. Ta có

$$X \sim B(1.000;\ 0,002)\,.$$

Do $n = 1000$ (lớn) và $p = 0,002$ (bé) nên

$$X \sim B(1.000;\ 0,002) \approx P(2) \text{ và } \lambda = 1.000 \times 0,002 = 2\,.$$

Ta có bảng so sánh giữa phân phối nhị thức và phân phối Poisson:

$X \sim B(1.000;\ 0,002)$	$X \sim P(2)$
$P(X \le 2) = p(0) + p(1) + p(2)$ $= (0,998)^{1.000} + C^1_{1.000}(0,002)(0,998)^{999}$ $+ C^1_{1.000}(0,002)^2(0,998)^{998}$ $= 0,6766765.$	$p(X \le 1) = p(0) + p(1) + p(2)$ $= e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!}\right) = \frac{5}{e^2}$ $= 0,6766764.$

Ta thấy khi số phép thử (n) lớn và khả năng xảy ra (p) bé thì phân phối nhị thức sẽ xấp xỉ phân phối Poisson.

2.4.3. Phân phối chuẩn (Normal distribution)

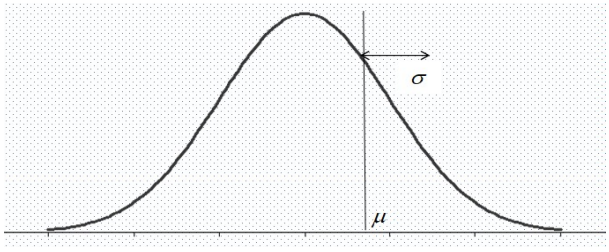
Định nghĩa: ĐLNN liên tục X có phân phối chuẩn, với hai tham số μ và σ^2, ký hiệu $X \sim N(\mu,\ \sigma^2)$ nếu hàm mật độ của nó có dạng

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

Tính chất. Giả sử $X \sim N(\mu,\ \sigma^2)$.

a) Ta có

$$E(X) = \mu, D(X) = \sigma^2, Mod(X) = \mu\,.$$



Hình 2.5. Đồ thị của ĐLNN có phân phối chuẩn

b) Nếu $\mu = 0$ và $\sigma^2 = 1$ thì $Z \sim N(0;1)$ được gọi là phân phối chuẩn tắc (*standardized normal distribution*). Khi đó hàm mật độ của ĐLNN Z có dạng

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

Hàm phân phối xác suất của Z là

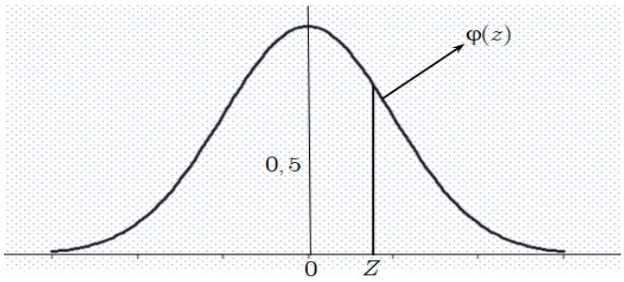
$$F(z) = p(Z < z) = \int_{-\infty}^{z} f(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{z} e^{-\frac{t^2}{2}} dt = 0,5 + \varphi(z),$$

trong đó

$$\varphi(z) := \frac{1}{\sqrt{2\pi}} \int_{0}^{z} e^{-\frac{t^2}{2}} dt$$

được gọi là hàm Laplace (tra bảng số 2). Chú ý rằng $\varphi(z)$ là hàm lẻ.



Hình 2.6. Hàm phân phối xác suất và công thức Laplace.

Nhận xét. Nếu $X \sim N(\mu, \sigma^2)$ thì $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Khi đó ta có các công thức tính

a) $p(X < a) = 0.5 + \varphi\left(\frac{a-\mu}{\sigma}\right)$.

b) $p(X \geq a) = 0.5 - \varphi\left(\frac{a-\mu}{\sigma}\right)$.

c) $p(a \leq X \leq b) = \varphi\left(\frac{b-\mu}{\sigma}\right) - \varphi\left(\frac{a-\mu}{\sigma}\right)$.

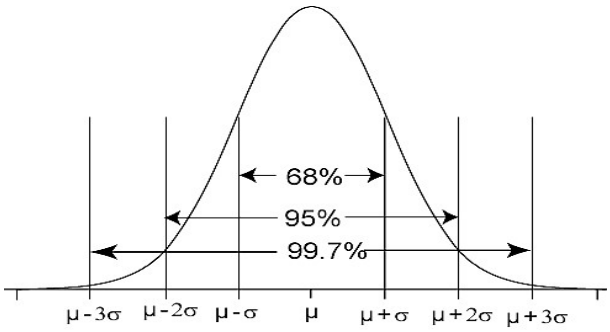
d) $p\left(\left|X-\mu\right| \leq \varepsilon\right) = 2\varphi\left(\frac{\varepsilon}{\sigma}\right)$.

e) $p\left(\left|X-\mu\right| \leq k\sigma\right) = 2\varphi(k)$.

Công thức e) được gọi là qui tắc $k-$sigma. Qui tắc này hay được dùng trong tính toán kỹ thuật vì xác suất hay độ tin cậy cao. Khi $k = 3$, ta có qui tắc $3-$sigma

$$p\left(\left|X-\mu\right| \leq 3\sigma\right) = 2\varphi(3) = 0,9973.$$

Điều này có nghĩa xác suất X nhận giá trị trong khoảng $(\mu - 3\sigma,\ \mu + 3\sigma)$ là $99,73\%$.



Hình 2.7. Qui tắc $k-$sigma

Ý nghĩa. Phân phối chuẩn là qui luật phân phối xác suất được áp dụng rộng rãi trong thực tế như các số đo về đặc tính sinh học, sai số đo lường về vật lý, sản lượng một mùa vụ, lợi tức hàng năm,...đều tuân theo qui luật chuẩn.

Ví dụ 1. Độ dài của một sợi dây kim loại do một máy tự động cắt ra tuân theo qui luật chuẩn với trung bình $10\,mm$ và độ lệch chuẩn $2\,mm$. Tính tỉ lệ sợi dây kim loại do máy tự động cắt ra có độ dài từ

a) Từ $14\,mm$ trở lên.

b) Từ $8\,mm$ đến $14\,mm$.

c) Chọn ngẫu nhiên 10 sợi dây kim loại này. Tính xác suất có đúng 9 sợi dây có chiều dài từ $8\,mm$ đến $14\,mm$.

Giải. Gọi $X\,(mm)$ là độ dài của sợi dây kim loại do máy tự động cắt ra. Ta có $X \sim N(10;4)$.

a) $p(X \geq 14) = 0,5 - \varphi\left(\dfrac{14-10}{2}\right) = 0,5 - \varphi(2) == 0,5 - 0,4772 = 0,28$.

Vậy có 2,28% sợi dây kim loại do máy tự động cắt ra có chiều dài từ $14\,mm$ trở lên.

b) $p(8 \leq X \leq 14) = \varphi\left(\dfrac{14-10}{2}\right) - \varphi\left(\dfrac{8-10}{2}\right)$

$= \varphi(2) + \varphi(1) = 0,4772 + 0,3413 = 0,8185$.

Vậy có 81,85% sợi dây kim loại do máy tự động cắt ra có chiều dài từ $8\,mm$ đến $14\,mm$.

c) Gọi T số sợi dây có chiều dài từ $8\,mm$ đến $14\,mm$ trong 10 sợi dây được lấy ra ngẫu nhiên. Ta có $T \sim B(10;\ 0,8185)$. Suy ra

$$p(T = 9) = C_{10}^{9}(0,8185)^{9}(0,1815) = 0,2993.$$

Ví dụ 2. Đường kính $X\,(mm)$ của mỗi trục máy ngẫu nhiên do một nhà máy tự động sản xuất là ĐLNN theo qui luật chuẩn với độ lệch chuẩn $\sigma = 0,04\,(mm)$. Trục máy gọi là đạt tiêu chuẩn kỷ thuật nếu đường kính của nó sai lệch so với đường kính thiết kế μ không quá $0,072\,(mm)$. Tìm tỉ lệ trục máy đạt tiêu chuẩn kỹ thuật của nhà máy?.

Giải. Ta có $X \sim N\left(\mu;\ (0,04)^2\right)$. Tỷ lệ trục máy đạt tiêu chuẩn kỹ thuật của nhà máy là

$$P\left(|\ X-\mu\ | \le 0,072\right) = 2\varphi\left(\frac{0,072}{0,04}\right) = 2\varphi(1,8) = 2\times 0,4641 = 0,9282.$$

Vậy tỉ lệ trục máy đạt tiêu chuẩn kỹ thuật của nhà máy là 92,82%.

2.4.4. Phân phối "khi bình phương" (χ^2).

Định nghĩa: Cho n ĐLNN độc lập $X_i \sim N(0;1),\ i = \overline{1,n}$. Khi đó

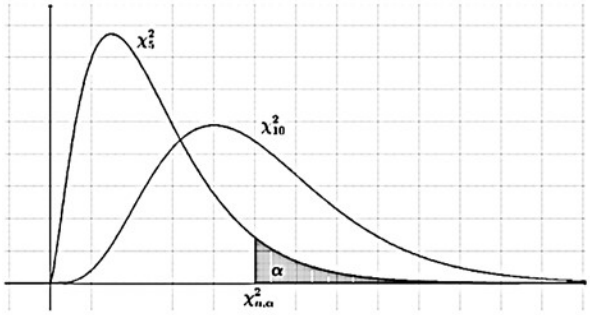
$$X := X_1^2 + X_2^2 + \cdots + X_n^2 = \sum_{i=1}^{n} X_i^2$$

có phân phối khi bình phương với n bậc tự do, ký hiệu $X \sim \chi^2(n)$.

Tính chất. Nếu $X \sim \chi^2(n)$ thì

a) $E(X) = n$. b) $D(X) = 2n$.

Nhận xét. Bảng tính xác suất $p\left(X > \chi_\alpha^2(n)\right) = \alpha$ được tính sẵn ở bảng 5. Khi n lớn thì phân phối khi bình phương xấp xỉ phân phối chuẩn.



Hình 2.8. Đồ thị phân phối khi bình phương.

2.4.5. Phân phối Student

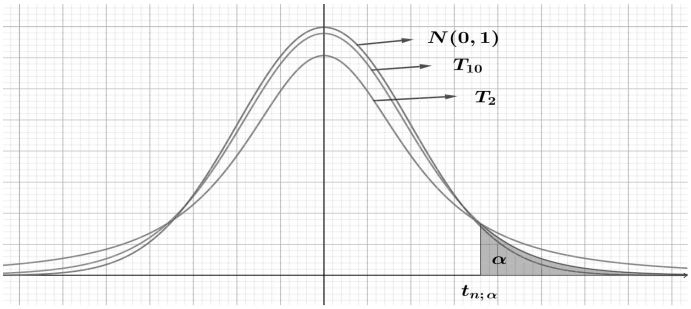
Định nghĩa: Cho $X \sim N(0;1)$ và $Y \sim \chi^2(n)$, khi đó ĐLNN

$$T := \frac{X}{\sqrt{\frac{Y}{n}}}$$

có phân phối Student với n bậc tự do và ký hiệu $T \sim T(n)$.

Tính chất. Nếu $T \sim T(n)$ thì

a) $E(X) = 0$. b) $D(X) = \dfrac{n}{n-2},\ n > 2$.



Hình 2.9. Đồ thị của phân phối Student và phân phối chuẩn tắc.

Nhận xét. Phân phối Student có cùng dạng và tính đối xứng như phân phối chuẩn tắc. Khi $n > 30$ phân phối Student có đồ thị của hàm mật độ xấp xỉ đồ thị hàm mật độ của phân phối chuẩn tắc.

2.4.6. Mối quan hệ giữa các phân phối xác suất

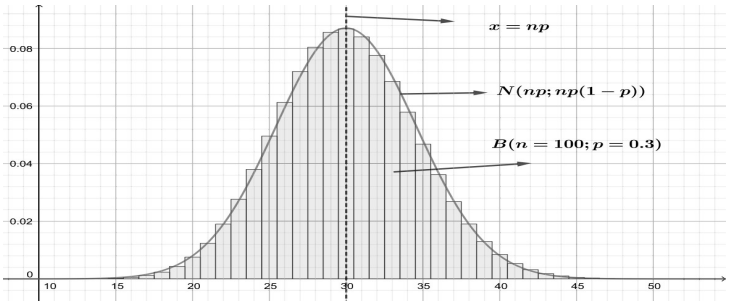
Định lý 2.2 (Moivre-Laplce). Cho $X \sim B(n, p)$. Nếu n đủ lớn và p không quá gần 0 và 1 thì

$$B(n, p) \approx N\Big(np; np(1-p)\Big)$$

và

$$p(a \le X \le b) \approx \varphi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \varphi\left(\frac{a-np}{\sqrt{np(1-p)}}\right).$$

Chú ý: Công thức xấp xỉ trên được áp dụng tốt khi $np > 5$ và $n(1-p) > 5$.



Hình 2.10. Hình minh họa xấp xỉ phân phối nhị thức và phân phối chuẩn

Ví dụ 1. Xác suất sinh được em bé trai là 0,48. Tính xác suất trong 300 trường hợp sắp sinh được chọn ngẫu nhiên có số bé trai vào khoảng từ 150 bé đến 170 bé.

Giải. Gọi X là số bé trai trong 300 trường hợp sắp sinh em bé. Khi đó $X \sim B(300;\ 0,48)$.

Ta có $n = 300$ khá lớn, $p = 0,48$ không quá gần 0 và 1 nên

$$X \sim B(300;\ 0,48) \approx N(\mu,\ \sigma^2) \text{ với } \mu = np = 144,\ \sigma = \sqrt{np(1-p)} = 8,6533.$$

Xác suất để số bé trai khoảng từ 150 đến 170 em là

$$p(150 \le X \le 170) = \varphi\left(\frac{170-144}{8,6533}\right) - \varphi\left(\frac{150-144}{8,6533}\right)$$
$$\approx \varphi(3) - \varphi(0,69) = 0,4987 - 0,2549 \approx 0,2438.$$

Ví dụ 2. Một công ty bảo hiểm có 10.000 khách hàng mua bảo hiểm xe máy ở độ tuổi 18 đến 20 tuổi với giá 300.000 đồng/năm và trung bình một khách hàng sẽ được bồi thường 3.000.000 nếu xe của họ bị tai nạn giao thông. Qua thống kê cho thấy tỉ lệ một xe máy bị tai nạn giao thông do người ở độ tuổi từ 18 tuổi đến 20 tuổi điều kiển là 0,09. Tính khả năng để công ty bảo hiểm bị lỗ sau một năm bán bảo hiểm cho khách hàng ở độ tuổi này.

Giải. Gọi X là số tai nạn giao thông trong một năm của những người ở độ tuổi 18 đến 20 tuổi và T là số tiền lời của công ty sau một năm bán bảo hiểm cho khách hàng ở độ tuổi trên. Ta có

$$X \sim B(5000;\ 0,09) \approx N(450; 409,5)$$

và số tiền lời

$$T = 300.000 \times 5000 - 3.000.000X.$$

Khả năng để công ty bị lỗ là

$$p(T<0)=p(300.000\times 5000-3.000.000X<0)=p(X>500)$$
$$\approx 0,5-\varphi\left(\frac{500-450}{\sqrt{409,5}}\right)\approx 0,5-\varphi(2,471)\approx 0,5-0,4932=0,0068.$$

Vậy khả năng bị lỗ của công ty trong trường hợp này là 0,0068.

2.5. Bài tập chương 2

1. Một dây chuyền gồm 3 bộ phận hoạt động độc lập trong thời gian 1 năm. Xác suất các bộ phận 1, 2, 3 bị hỏng trong thời gian hoạt động 1 năm là tương ứng là 0,4; 0,2 và 0,3. Gọi X là số bộ phận bị hỏng.

a) Lập bảng phân phối xác suất của ĐLNN X.

b) Tính xác suất có không quá 2 bộ phận bị hỏng.

c) Tính $E(X)$, $D(X)$, $\sigma(X)$ và giá trị tin chắc nhất của X.

2. Lãi suất thu được trong một năm (%) khi đầu tư vào công ty **A**, công ty **B** tương ứng là các ĐLNN độc lập X,Y. Cho biết qui luật phân phối xác suất của X và Y như sau

X	4	6	8	10	12
P	0,05	0,1	0,3	0,4	0,15

X	−4	2	8	10	12	16
P	0,1	0,2	0,2	0,25	0,15	0,1

a) Đầu tư vào công ty nào có lãi suất trung bình (lãi suất kỳ vọng) cao hơn.

b) Đầu tư vào công ty nào có mức độ rủi ro ít hơn.

c) Đầu tư vào cả hai công ty theo tỉ lệ nào để ít rủi ro nhất.

3. Số tiền lời trong năm tới (tính theo đơn vị: triệu đồng) thu được khi đầu tư 100 triệu đồng vào hai nghành A và B tùy thuộc vào tình hình kinh tế (THKT) trong nước và cho ở bảng sau:

Số tiền lời / THKT	Kém phát triển	Khả quan	Phát triển
Ngành *A*	10	40	80
Ngành *B*	−30	70	110
Dùng xác suất THKT	0,25	0,45	0,3

a) Số tiền lời trung bình (số tiền lơi kỳ vọng) ngành nào là cao hơn?

b) Mức độ rủi ro tiền lời ngành nào là ít hơn? (Mức độ rủi ro của tiền lời có thể hiểu là độ phân tán của tiền lời xung quanh tiền lời trung bình và chỉ ra bởi độ lệch chuẩn của tiền lời hoặc phương sai tiền lời).

4. Cho X (nghìn sản phẩm) là nhu cầu mỗi năm về một loại hàng A ở nước ta với X là một đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất

$$f(x)=\begin{cases}k(30-x) & x\in(0,30)\\ 0 & x\notin(0,30)\end{cases}.$$

a) Tìm k.

b) Tìm nhu cầu trung bình hàng năm của loại hàng A.

c) Tìm xác suất để nhu cầu về mặt hàng A không vượt quá 12 nghìn sản phẩm trong một năm?

d) Tìm phương sai và độ lệch (chuẩn) của X.

5. Tỉ lệ sản phẩm tốt của một lô hàng lớn là 90%. Kiểm tra ngẫu nhiên 10 sản phẩm. Gọi X là số sản phẩm tốt trong 10 sản phẩm lấy ra.

a) Chỉ ra qui luật phân phối xác suất của X.

b) Số sản phẩm tốt trung bình.

c) Tính xác suất để có nhiều nhất là 2 sản phẩm tốt trong 10 sản phẩm lấy ra.

6. Bắn 5 viên đạn vào mục tiêu. Xác suất trúng đích của mỗi lần bắn là như nhau và bằng 0,2. Muốn phá hủy mục tiêu phải có ít nhất 3 viên đạn trúng mục tiêu.

a) Chỉ ra phân phối xác suất của số viên đạn trúng mục tiêu.

b) Trung bình có bao nhiêu viên trúng mục tiêu.

c) Tìm xác suất mục tiêu bị phá hủy.

7. Một đại lý điện thoại di động dự định sẽ áp dụng một trong 2 phương án kinh doanh. Gọi X_1, X_2 (triệu đồng/ tháng) tương ứng là lợi nhuận thu được khi áp dụng phương án thứ nhất, phương án thứ hai. Giả sử $X_1 \sim N(140; 2.500)$; $X_2 \sim N(200; 3.600)$. Nếu biết rằng để đại lý tồn tại và phát triển thì lợi nhuận thu được từ kinh doanh điện thoại phải đạt ít nhất 80 (triệu/ tháng). Theo bạn công ty nên áp dụng phương án nào để kinh doanh điện thoại di động? Vì sao?

8. Một người cân nhắc giữa việc mua nhà bây giờ hay gửi tiết kiệm lãi suất 12% một năm để chờ năm sau sẽ mua. Biết mức tăng giá nhà sau một năm $X(\%)$ là ĐLNN phân phối chuẩn với $E(X) = 8(\%)$; $\sigma(X) = 10(\%)$. Giả sử người đó quyết định gửi tiền vào tiết kiệm. Tìm khả năng để quyết định đó là sai lầm. **ĐS.** 0,3446

9. Biết tuổi thọ X (năm) của mỗi sản phẩm là ĐLNN phân phối chuẩn với tuổi thọ trung bình là 8 năm và độ lệch chuẩn là 2 năm. Nếu bán được một sản phẩm thì cửa hàng lãi 150 (ngàn). Còn nếu sản phẩm bị hỏng trong thời gian bảo hành thì cửa hàng phải chi lại 500 (ngàn) cho bên bảo hành.Thời gian bảo hành mỗi sản phẩm được qui định là 6 năm.

a) Tìm tỉ lệ sản phẩm bị bảo hành.

b) Tiền lãi trung bình cửa hàng thu được sau khi bán mỗi sản phẩm. **ĐS.** 70,65 (ngàn)

Bài tập làm thêm

10. Theo thống kê, tỉ lệ để một người độ tuổi 40 sống thêm ít nhất 1 năm nữa là 99,5%. Một công ty nhân thọ bán mỗi thẻ bảo hiểm 1 năm cho mỗi người độ tuổi đó với giá 10 (ngàn đồng) và trường hợp người mua bảo hiểm chết sẽ có số tiền bồi thường là 1 triệu đồng. Tìm lợi nhuận kỳ vọng của công ty bảo hiểm khi bán mỗi thẻ bảo hiểm loại này.

11. Tỉ lệ bị cận thị của học sinh Việt Nam là 9%. Chọn ngẫu nhiên 100 học sinh. Gọi X là số học sinh bị cận thị.

a) Chỉ ra qui luật phân phối xác suất của X.

b) Số học sinh bị cận thị trung bình.

c) Tính xác suất có ít nhất 1 học sinh bị cận thị.

d) Tính xác suất có nhiều nhất 11 học sinh bị cận thị.

12. Thời gian hoạt động tốt X (giờ) của một TV cùng loại là một ĐLNN phân phối chuẩn, với trung bình $\mu = 4.300$ (giờ) và độ lệch chuẩn $\sigma = 250$ (giờ). Giả thiết mỗi ngày trung bình

người ta dùng TV là 10 (giờ) và thời hạn bảo hành miễn phí là 360 ngày. Tính tỉ lệ sản phẩm phải bảo hành. **ĐS.** 0,0026

13. Gọi $X\,(kWh)$ là lượng điện tiêu thụ mỗi tháng của mỗi hộ gia đình ở miền Trung. Biết X có phân phối chuẩn với trung bình $160\,kWh$, độ lệch chuẩn là $40\,kWh$. Giả sử trong $50\,kWh$ đầu tiên phải trả 1.000 (ngàn) cho mỗi kWh điện. Những kWh điện tiêu thụ tiếp theo phải trả 2.000 cho mỗi kWh điện.

a) Tính tỉ lệ hộ gia đình tiêu thụ dưới $90\,kWh$ trong 1 tháng.

b) Tính tỉ lệ hộ gia đình trả tiền điện trong 1 tháng nhiều hơn 300.000.

14. Gọi $X\,(mm)$ là chiều dài mỗi sản phẩm do một phân xưởng sản xuất; X có phân phối chuẩn với độ lệch là $0,5\,mm$. Sản phẩm gọi là đạt chất lượng cao nếu chiều dài sản phẩm sai lệch chiều dài trung bình không quá $0,1\,mm$.

a) Tìm tỉ lệ sản phẩm đạt chất lượng cao của phân xưởng

b) Chọn ngẫu nhiên 10 sản phẩm. Tính xác suất có ít nhất 2 sản phẩm đạt chất lượng cao.

c) Trung bình số sản phẩm đạt chất lượng cao là bao nhiêu?

15. Gọi $X\,(kg)$ là cân nặng mỗi con heo của các trang trại lớn ở một tỉnh; X có phân phối chuẩn với cân nặng trung bình là $80\,kg$ và độ lệch chuẩn là $10\,kg$. Heo gọi là đạt loại I nếu có cân nặng hơn $90\,kg$.

a) Tìm tỉ lệ heo đạt loại I ở các trang trại đó.

b) Chọn ngẫu nhiên 8 con heo. Tìm xác suất để có ít nhất 2 con đạt loại I.

c) Trung bình số heo đạt loại I là bao nhiêu?.

16. Lãi suất $X\,(\%)$ đầu tư vào một dự án được xem như một ĐLNN phân phối theo qui luật chuẩn. Theo đánh giá của ủy ban đầu tư thì lãi suất cao hơn 20% có xác suất là 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? **ĐS:** 99,87%.

17. Độ dài chi tiết $X\,(cm)$ do máy sản suất là biến ngẫu nhiên phân phối chuẩn với độ lệch là $10\,cm$. Biết tỉ lệ chi tiết có độ dài dưới $84\,cm$ là 84,13%.

a) Tìm độ dài trung bình mỗi chi tiết.

b) Tìm tỉ lệ chi tiết có độ dài hơn $80\,cm$.

c) Lấy ngẫu nhiên 4 chi tiết. Tìm xác suất có ít nhất 1 chi tiết có độ dài hơn $80\,cm$.

18. Chị A nuôi 160 con vịt đẻ cùng loại. Xác suất để 1 con vịt đẻ trứng trong ngày là 0,8. Qui ước mỗi con vịt chỉ đẻ nhiều nhất 1 trứng.

a) Chỉ ra phân phối xác suất của số trứng vịt đẻ trong ngày.

b) Tìm xác suất để chị A có được ít nhất 130 trứng trong ngày.

c) Nếu mỗi quả trứng bán được 2.000 đồng, tiền thức ăn cho vịt ăn trong ngày là 900 đồng. Tính số tiền lãi trung bình chị A thu được trong ngày là bao nhiêu.

19. Gọi X (m^3/tháng). Lượng nước sử dụng của mỗi hộ gia đình ở thành phố Nha Trang là biến ngẫu nhiên có phân phối chuẩn với trung bình là $\mu = 12\,(m^3)$ và độ lệch chuẩn $\sigma = 2\,(m^3)$.

a) Tính $p(X < 10)$ và $p(14 \le X \le 18)$. Minh họa các kết quả trên bằng hình vẽ.

b) Từ $0\,m^3$ 0 đến $11\,m^3$ có đơn giá là 8.000 đồng/m^3. Trên $11\,m^3$ có đơn giá là 10.000 đồng/m^3. Tính tỉ lệ hộ gia đình ở Nha trang phải trả từ 158.000 đồng/tháng trở lên.

20. Cho hai hộp. Hộp 1 có 1 tờ 500.000 đồng; 9 tờ 20.000 đồng; Hộp 2 có 5 tờ 100.000 đồng và 2 tờ 50.000 đồng và 3 tờ 10.000 đồng. Anh Lâm được quyền chọn một trong hai hộp rồi lấy ngẫu nhiên ra 1 tờ tiền. Nếu lấy được tờ tiền nào thì được nhận tờ tiền đó. Hỏi rằng anh Lâm nên chọn hộp nào thì có lợi hơn (theo nghĩa kỳ vọng số tiền anh Lâm được nhận là cao hơn).

Chương 3.
Mẫu thống kê và ước lượng tham số

Từ chương này ta bắt đầu nghiên cứu về thống kê. Khoa học thống kê ra đời nhằm mục đích nghiên cứu các phương pháp thu thập, tổ chức và phân tích các dữ liệu nhằm thu nhận thông tin chân thực về đối tượng nghiên cứu một cách khách quan, đáng tin cậy và rút ra những kết luận hợp lý. Thống kê được ứng dụng rộng rãi trong hầu hết các lĩnh vực và có vai trò cực kì quan trọng trong nhiều ngành khoa học, nhất là trong các ngành khoa học thực nghiệm như y khoa, sinh học, nông nghiệp, kinh tế,… Đặc biệt thống kê rất cần cho các cấp lãnh đạo, các nhà quản lý, các nhà hoạch định chính sách. Khoa học thống kê cung cấp cho họ các phương pháp thu thập, xử lý và diễn giải các phân tích về dân số, kinh tế, giáo dục,... để từ đó có thể vạch chính sách và ra các quyết định đúng đắn.

Chương 3 gồm hai phần: phần thứ nhất nói về mẫu và thống kê mô tả; mẫu ngẫu nhiên và các đặc trưng mẫu; phần thứ hai nghiên cứu về các bài toán ước lượng tham số như ước lượng điểm, ước lượng khoảng cho tham số, và một số bài toán cần quan tâm khi xét đến bài toán ước lượng.

3.1. Mẫu và thống kê mô tả

3.1.1. Tổng thể và mẫu ngẫu nhiên

1. Tổng thể: Tập hợp các phần tử mà ta cần khảo sát về một hay một số dấu hiệu nào đó gọi là một tổng thể.

Ví dụ. Ta cần nghiên cứu chiều cao của các cây bạch đàn trên một vùng đất. Dấu hiệu nghiên cứu là chiều cao. Tổng thể nghiên cứu là tập hợp các cây bạch đàn trên vùng đất đó.

Gọi $X(m)$ là chiều cao cây bạch đàn (chọn ngẫu nhiên) trong vùng. Khi đó X là một đại lượng ngẫu nhiên trên tổng thể và ta có thể coi việc nghiên cứu chiều cao của các cây này là nghiên cứu đại lượng ngẫu nhiên X trên tổng thể.

Bài toán nghiên cứu dấu hiệu trên tổng thể thường đưa đến bài toán nghiên cứu đại lượng ngẫu nhiên trên tổng thể.

Thông thường, tổng thể cần nghiên cứu là rất lớn, ta không thể chọn hết được tất cả các phần tử của tổng thể để nghiên cứu bởi vì: Tốn kém nhiều (về thời gian, nhân lực, kinh tế,…); Việc quan sát, kiểm tra có thể làm hư hại các phần tử của tổng thể (chẳng hạn kiểm tra hàm lượng các chất có trong các hộp sữa, viên thuốc,…); Có nhiều trường hợp không thể xác định toàn bộ các phần tử của tổng thể (chẳng hạn kiểm tra các bệnh nhân HIV, chiều cao của một loài cây,…). Do đó, ta chỉ thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

2. Mẫu: là một tập con được chọn ra từ tổng thể. Ta thường ký hiệu N để chỉ số phần tử của tổng thể và n để chỉ cỡ mẫu.

3. Phương pháp chọn mẫu - Mẫu ngẫu nhiên kích thước n : Để việc nghiên cứu trên mẫu cho ta kết quả gần với kết quả mà ta mong muốn khi nghiên cứu trên tổng thể thì mẫu được chọn phải có tính ngẫu nhiên, mang tính đại diện cho tổng thể và các số liệu phải đạt độ chính xác nào đó. Việc thu hẹp phạm vi nghiên cứu giúp việc xử lý cho ta kết quả vừa nhanh vừa đỡ tốn kém mà vẫn đạt được độ chính xác và tin cậy cần thiết.

Cụ thể, một *mẫu ngẫu nhiên* gồm n phần tử được chọn ra từ một tổng thể phải thỏa các điều kiện sau

- Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).

- Mọi cỡ mẫu n cũng có cùng khả năng được chọn từ tổng thể.

Thông thường, với kích thước mẫu đạt mức đủ lớn, việc coi mẫu thỏa yêu cầu bảo đảm những điều kiện trên nói chung là có thể chấp nhận được. Tuy nhiên, nhiều khi do thực tế không thể làm như trên hay do tổng thể không xác định, nhiều trường hợp khó lấy được mẫu bảo đảm tính ngẫu nhiên và mang tính đại diện cho tổng thể. Thường người ta yêu cầu:

- Người lấy mẫu là các chuyên gia được đào tạo bài bản và có kinh nghiệm với các dạng tổng thể cần lấy mẫu.
- Các dụng cụ, máy móc để cân đo và ghi dữ liệu phải được chọn lựa và tinh chỉnh để có độ chính xác đạt mức chấp nhận được.
- Số phần tử n trong mẫu không được quá bé (chẳng hạn, tác giả Calvin Dytham đề nghị mức tối thiểu kích thước mẫu $n \geq 20$ trong các thống kê sinh học thường gặp).
- Phương pháp lấy mẫu sẽ được các chuyên gia lựa chọn phù hợp với các dạng tổng thể phải lấy mẫu và phù hợp với thực tế. Điều này sẽ được trình bày cụ thể hơn trong các giáo trình của thống kê ứng dụng.

Hiện nay, có nhiều phương pháp khác nhau để chọn mẫu ngẫu nhiên, nhưng khó có thể nói rằng phương pháp nào là tốt nhất. Việc chọn phương pháp phù hợp phụ thuộc vào đối tượng cụ thể và thói quen hay sở trường của nhà nghiên cứu. Tuy nhiên trong giới hạn của giáo trình này, ta chỉ giới thiệu phương pháp đơn giản nhất để người đọc dễ nắm bắt:

- Đánh số các phần tử của tổng thể từ 1 đến N. Lập các phiếu cũng đánh số như vậy.
- Trộn đều các phiếu, sau đó chọn có hoàn lại n phiếu. Các phần tử của tổng thể có số thứ tự trong phiếu lấy ra sẽ được chọn làm mẫu.

Trong khuôn khổ giáo trình này, những ví dụ và các bài tập chỉ đơn giản mang tính minh họa cho lý thuyết thống kê toán, nên ta chỉ xét các tổng thể là xác định và cho phép giả định mẫu thỏa yêu cầu bảo đảm tính ngẫu nhiên và tính đại diện cho tổng thể.

Giả sử xét đại lượng ngẫu nhiên X trên một tổng thể. Ta sẽ n lần thực hiện loạt các hành động sau lấy ngẫu nhiên phần tử của tổng thể, "đo" giá trị X và hoàn lại tổng thể. Giá trị X sẽ "đo" được trên phần tử thứ i được gán là trị của $X_i, i = 1,2,...,n$. Các đại lượng $X_1, X_2,..., X_n$ là các đại lượng ngẫu nhiên độc lập, chúng là các bản sao của X và có cùng qui luật phân phối xác suất với X. Ta gọi bộ thứ tự n đại lượng ngẫu nhiên $(X_1, X_2,..., X_n)$ là *mẫu ngẫu nhiên* (*kích thước* n) của đại lượng ngẫu nhiên X.

Giả sử sau khi thực hiện xong cụ thể n lần lấy và "đo", ta được mẫu số liệu cụ thể $(x_1, x_2,..., x_n)$. Ta gọi mẫu số liệu này là một mẫu cụ thể của đại lượng ngẫu nhiên X và nó cũng là một giá trị cụ thể của mẫu ngẫu nhiên $(X_1, X_2,..., X_n)$.

Nhận xét. Với tổng thể đủ lớn , ta có thể không cần động tác hoàn lại các phần tử trong quá trình lấy mẫu.

Ví dụ. Chúng ta cần nghiên cứu khối lượng các con tôm 2 tháng tuổi ở một vùng. Gọi $X(g)$ là khối lượng con tôm 2 tháng tuổi vùng đó. Giả sử đại lượng ngẫu nhiên X có phân phối chuẩn với giá trị tring bình $\mu = 20(g)$. Ta dự định sẽ 100 lần lấy ngẫu nhiên từng con tôm 2 tháng tuổi ở trong vùng, cân (rồi hoàn lại).

Gọi $X_i(g)$ là khối lượng con tôm lấy ngẫu nhiên lần thứ i. Bộ thứ tự $(X_1, X_2,..., X_{100})$ là một mẫu ngẫu nhiên kích thước 100 của đại lượng ngẫu nhiên X. Ta có $X_1, X_2,..., X_{100}$ là các đại lượng ngẫu nhiên độc lập và chúng có phân phối chuẩn $N(70;20^2)$ giống như X.

Thực hiện cụ thể 100 lần bắt tôm, cân và hoàn lại, ta được các giá trị cụ thể $(x_1, x_2, ..., x_{100})$. Khi đó, $(x_1, x_2, ..., x_{100})$ là một mẫu cụ thể của đại lượng ngẫu nhiên X và đây cũng là giá trị cụ thể của mẫu ngẫu nhiên $(X_1, X_2, ..., X_{100})$ trên.

3.1.2. Biểu diễn số liệu cụ thể

Dữ liệu dạng ban đầu (sơ khai, thô) rất khó để quan sát, nhận dạng và đánh giá. Do vậy, ta cần phải tổ chức lại dữ liệu. Các dạng tổ chức dữ liệu thường ở dạng bảng hoặc đồ thị.

Các loại đồ thị được sử dụng sẽ phụ thuộc vào biến được tổng hợp. Các dạng đồ thị quan trọng thường dùng: đồ thị tổ chức tần số (histogram), đồ thị stem-and-leaf, đồ thị phân tán (scatter plot), biểu đồ đường (line chart), đồ thị xác suất (probability plot),…

1. Đồ thị tổ chức tần số (histogram): Phân phối tần số (frequency distribution) là một danh sách dạng bảng, chứa các khoảng được phân nhóm theo dữ liệu quan trắc và các tần số tương ứng của dữ liệu nằm bên trong từng khoảng.

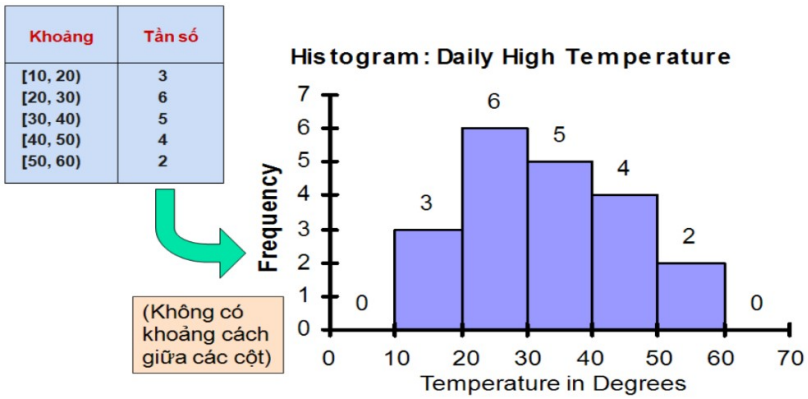
Đồ thị tổ chức tần số là một hình ảnh hiển thị của phân phối tần số. Để xây dựng một đồ thị tổ chức tần số, ta thực hiện theo các bước sau

- Đánh nhãn các khoảng trên trục hoành.
- Đánh nhãn trục tung bằng tần số hoặc tần suất.
- Trên mỗi khoảng, vẽ một hình chữ nhật với chiều cao bằng với tần số (hoặc tần suất) tương ứng với khoảng đó.

Ví dụ. Chọn ngẫu nhiên 20 ngày mùa đông có nhiệt độ cao và đo nhiệt độ (đơn vị: độ F) được số liệu như sau

24	35	17	21	24	37	26	46	58	30
32	13	12	38	41	43	44	27	53	27

Hãy lập bảng phân phối tần số cho số liệu này.



2. Đồ thị stem-and-leaf: Đồ thị stem-and-leaf là một phương pháp đơn giản để nhận biết các chi tiết của phân phối trong một tập dữ liệu. Để xây dựng một đồ thị stem-leaf, ta thực hiện theo các bước sau

- Sắp xếp dữ liệu theo thứ tự tăng dần.
- Chia các giá trị đã sắp xếp thành hai phần: phần gốc (*stem*), gồm một (hoặc vài) chữ số đầu tiên và phần lá (*leaf*), gồm các chữ số còn lại.
- Liệt kê các giá trị stem vào một cột dọc.
- Ghi lại leaf cho mỗi quan sát vào bên cạnh stem của nó.
- Viết các đơn vị cho các stem và leaf lên đồ thị.

Ví dụ. Sắp xếp dữ liệu: 21, 24, 24, 26, 27, 27, 30, 32, 38, 41.

a) Sử dụng đơn vị hàng chục cho đơn vị của stem:

	Stem	Leaf
• 21 ® î c ghi lµ →	2	1
• 38 ® î c ghi lµ →	3	8

Hoàn thành đồ thị stem-and-leaf:

Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

b) Sử dụng đơn vị hàng trăm cho stem:

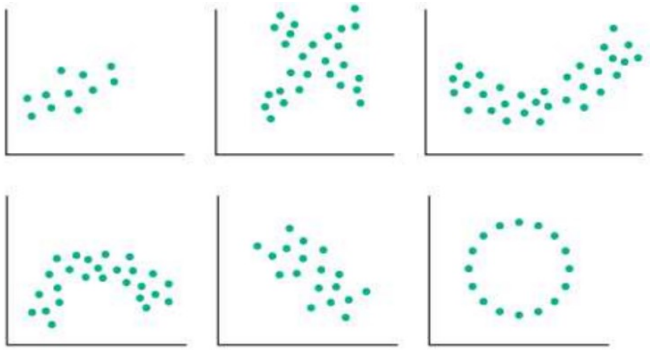
Data
613, 632, 658,
717, 722, 750, 776,
827, 841, 859, 863, 891, 894,
906, 928, 933, 955, 982,
1034, 1047, 1056,
1140, 1169,
1224.

→

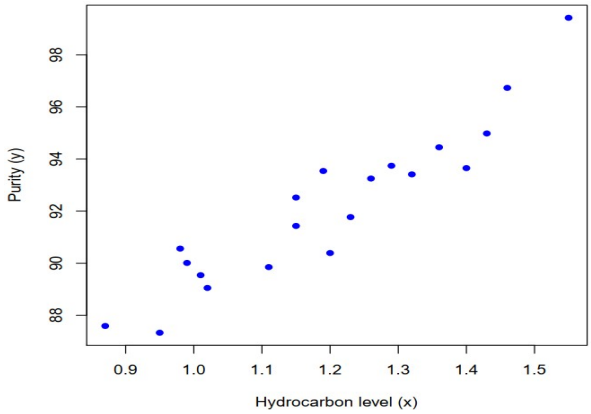
Đồ thị stem-and-leaf:

Stem	Leaaves
6	1 3 6
7	2 2 5 8
8	3 4 6 6 9 9
9	1 3 3 6 8
10	3 5 6
11	4 7
12	2

3. Đồ thị phân tán (scatter plot): Đồ thị phân tán được sử dụng để xác định mỗi liên hệ giữa hai biến X và Y.



Ví dụ. File data-oxygen.csv chứa số liệu về % độ tinh khiết của Oxygen (y) chiết xuất ra từ một chu trình hóa học và % hàm lượng Hydrocarbon (x) có trong chất chiết xuất. Dùng đồ thị phân tán xét mối quan hệ giữa x và y.



3.2. Các tham số đặc trưng của mẫu ngẫu nhiên và các tính chất

3.2.1. Các tham số đặc trưng tương ứng của tổng thể và mẫu

Xét X là đại lượng ngẫu nhiên trên tổng thể	**Xét mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ kích thước n của đại lượng ngẫu nhiên X**
Trung bình tổng thể: • $E(X) = \mu$ • $E(X^2)$: trung bình của X^2	Trung bình mẫu của X: • $\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$ • $\overline{X^2} = \frac{X_1^2 + X_2^2 + ... + X_n^2}{n}$: trung bình mẫu của X^2
Phương sai tổng thể: • $D(X) = \text{var}(X) = \sigma^2$ $= E\left(X^2\right) - \left[E(X)\right]^2$	Phương sai mẫu (hiệu chỉnh) của X: • $S^2 = S_X^{\ 2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\left(\overline{X}\right)^2}{n-1}$ $= \frac{n}{n-1}\left[\overline{X^2} - \left(\overline{X}\right)^2\right]$
Độ lệch (chuẩn) tổng thể: • $\sigma(X) = \sigma = \sqrt{D(X)}$	Phương sai mẫu (hiệu chỉnh) của X: • $S = S_X = \sqrt{S_X^{\ 2}}$
Tỉ lệ tổng thể: • $p = \frac{M_A}{N}$: tỉ lệ có tính chất A trong tổng thể N phần tử.	Tỉ lệ mẫu: • $F = \frac{m_A}{n}$: tỉ lệ có tính chất A trong mẫu (có hoàn lại) n phần tử sẽ lấy ngẫu nhiên.

Ta định nghĩa phương sai mẫu không hiệu chỉnh:

$$S_{khc}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \overline{X^2} - \left(\overline{X}\right)^2.$$

3.2.2. Các tính chất của đặc trưng mẫu

Trung bình mẫu $\overline{X}$, phương sai mẫu S^2, độ lệch mẫu S, tỉ lệ mẫu F là các đại lượng ngẫu nhiên và

a) $E(\overline{X}) = E(X) = \mu$; $D(\overline{X}) = \frac{D(X)}{n} = \frac{\sigma^2}{n}$.

b) $E(S_X^2) = D(X) = \sigma^2$.

c) $E(F) = p$; $D(F) = \frac{p(1-p)}{n}$.

Nhận xét 1. Theo nhiều nhà thống kê trong ứng dụng, khi số phần tử tổng thể N không đủ lớn so với kích thước mẫu n (chẳng hạn $n > N.\frac{5}{100}$). Nếu lấy mẫu không hoàn lại thì phải nhân thêm hệ số hiệu chỉnh vào công thức tính phương sai mẫu, độ lệch mẫu:

$$D(\bar{X}) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right); \qquad D(F) = \frac{p(1-p)}{n}.\left(\frac{N-n}{N-1}\right);$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}; \qquad \sigma(F) = \frac{\sqrt{p(1-p)}}{\sqrt{p}}\sqrt{\frac{N-n}{N-1}}.$$

Để đơn giản, trong giáo trình kể từ đây, ta luôn coi tổng thể là đủ lớn. Lúc này việc lấy mẫu có hay không hoàn lại được coi là như nhau. Do đó không cần sử dụng những công thức hiệu chỉnh ở trên.

Nhận xét 2. Khi kích thước mẫu n đủ rất lớn, thực tế có thể coi:

- Trung bình mẫu $\overline{X} \approx \mu = E(X)$: trung bình tổng thể.
- Phương sai mẫu $S^2 \approx \sigma^2 = D(X)$: phương sai tổng thể.
- Độ lệch mẫu $S \approx \sigma = \sigma(X)$: độ lệch (chuẩn) tổng thể.
- Tỉ lệ mẫu $F \approx p$: tỉ lệ tổng thể.

Chú ý 1. Với kích thước mẫu n đủ rất lớn, p không quá gần 0 và 1, $np > 5$ và $n(1-p) > 5$, có thể coi tỉ lệ mẫu F xấp xỉ tốt phân phối chuẩn $N\left(p, \frac{p(1-p)}{n}\right)$.

Chú ý 2.

* Trường hợp đại lượng ngẫu nhiên X có phân phối chuẩn $N(\mu, \sigma^2)$. Ta chứng minh được $\bar{X}$ có phân phối chuẩn $N\left(\mu, \frac{\sigma^2}{n}\right)$.

* Trường hợp X không phân phối chuẩn nhưng kích thước mẫu n đủ lớn. Từ định lý giới hạn trung tâm Liapunov (hay Lindenberg), suy ra với kích thước mẫu n đủ lớn, có thể coi trung bình mẫu $\overline{X}$ xấp xỉ phân phối chuẩn $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Trong thực tế, thường gặp các đại lượng ngẫu nhiên X có phân phối cân đối hoặc khá cân đối. Khi đó:

- Với $n \geq 30$, thường có thể coi $\bar{X}$ có phân phối xấp xỉ chuẩn $N\left(\mu, \frac{\sigma^2}{n}\right)$.
- Với $n \geq 100$, thường xấp xỉ này là khá tốt. Trong giáo trình này chúng ta chỉ xét các dạng đại lượng ngẫu nhiên thường gặp này.

3.3. Ước lượng điểm

3.3.1. Một số khái niệm về ước lượng tham số

Khi nghiên cứu đặc tính X trên một tổng thể lớn, ta thường quan tâm đến tham số đặc trưng θ của ĐLNN này. Tham số θ ở đây có thể là trung bình của tổng thể, tỉ lệ của tổng thể, phương sai của tổng thể, hoặc một đặc trưng nào đó khác. Và nếu tham số này vẫn chưa được xác định, thì một trong các bài toán cơ bản của thống kê toán là ước lượng tham số θ bằng phương pháp mẫu, tức dựa vào mẫu ngẫu nhiên thu thập được, ta đưa ra một giá trị hoặc một khoảng giá trị để đánh giá tham số θ.

1. Thống kê mẫu (Sample statistic): là một hàm của các biến ngẫu nhiên thành phần trong mẫu. Giả sử ta có mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$ thì thống kê mẫu (gọi tắt là thống kê) cũng là một ĐLNN và có dạng

$$K = f(X_1, X_2, \ldots, X_n),$$

với f là hàm bất kỳ.

2. Ước lượng (Estimator) là một giá trị được tính toán dựa trên mẫu ngẫu nhiên, nhằm xác định giá trị của một tham số đặc trưng nào đó của tổng thể.

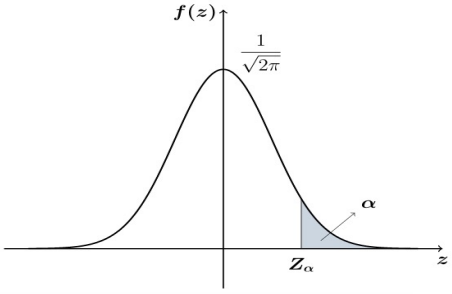
Nhận xét. Một ước lượng cũng là một thống kê mẫu. Một số các ước lượng thường gặp trong chương này gồm $\overline{X}, S, F$.

3. Phân vị mức (Critical value):

Định nghĩa. Cho X là ĐLNN có hàm phân phối xác suất $F(x)$ và số thực $\alpha \in (0,1)$. Khi đó phân vị mức $1-\alpha$ của ĐLNN X là một số thực K_α thỏa

$$P(X \geq K_\alpha) = \alpha .$$

Giả sử xét Z là ĐLNN có quy luật phân phối chuẩn tắc, hình dưới đây cho ta thấy cái nhìn trực quan về giá trị K_α, lúc này được ký hiệu là Z_α. Đó là một điểm trên trục số thực, ngăn phần diện tích α ở phía đuôi bên phải của phân phối.



Ghi chú.

- Thông thường $1-\alpha$ được chọn khá lớn cho bài toán ước lượng, tức α khá bé.
- Trong giáo trình này, ta dùng ký hiệu K_α để chỉ phân vị mức $1-\alpha$ của ĐLNN có phân phối xác suất bất kỳ. Ứng với mỗi dạng phân phối cụ thể, ta dùng ký hiệu khác nhau thay cho K_α (cụ thể xem phần Ví dụ dưới dây).
- Phần lớn các phân vị mức của các phân phối xác suất thông dụng đã được tính toán sẵn và tập hợp theo bảng. Để tìm giá trị này, ta có thể tra trên *bảng giá trị*.

Ví dụ. Với mỗi loại phân phối xác suất sau đây, bằng cách dùng bảng giá trị ta thu được

a) phân vị mức 0,975 của ĐLNN Z có phân phối chuẩn tắc $N(0,1)$ là

$Z_{0,025} = 1,96$ (tra từ bảng phân vị chuẩn tắc, bảng 3),

b) phân vị mức 0,95 của ĐLNN T có phân phối Student $T(24)$ là

$t_{0,05}^{(24)} = 1,7109$ (tra từ bảng phân vị Student, bảng 4),

c) phân vị mức 0,99 của ĐLNN χ^2 có phân phối khi-bình phương $\chi^2(2)$ là

$\chi^2_{0,01}(2) = 9,2103$ (tra bảng phân vị phân phối khi-bình phương, bảng 5).

1.3.2. Các tính chất của ước lượng điểm

Cách xấp xỉ trung bình tổng thể (hay độ lệch tổng thể) bằng các trị cụ thể của trung bình mẫu (hay độ lệch mẫu) được gọi là cách ước lượng điểm cho các tham số đặc trưng này.

Định nghĩa. Nếu lấy thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ thay cho tham số θ chưa biết thì $\hat{\theta}$ được gọi là ước lượng điểm của θ.

Nhận xét. Ước lượng điểm của θ là cũng một ĐLNN do $\hat{\theta}$ là một thống kê. Khi có mẫu cụ thể $(x_1, x_2, \ldots, x_n)$, ta thu được một trị cụ thể của ước lượng điểm.

Với cùng một mẫu ngẫu nhiên, nhiều thống kê khác nhau có thể được dùng như một ước lượng điểm cho tham số. Để đánh giá việc dùng ước lượng điểm nào thay thế tốt cho θ, ta dựa vào các tiêu chuẩn sau đây.

1. Ước lượng không chệch:

Định nghĩa. Thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ được gọi là một ước lượng không chệch của tham số θ nếu $E(\hat{\theta}) = \theta$. Nếu một ước lượng không thỏa tiêu chuẩn này, ta gọi nó là ước lượng chệch.

Ví dụ 1. Xét dấu hiệu X trên tổng thể có $E(X) = \mu$ và $(X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên từ tổng thể. Khi đó, thống kê $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ là một ước lượng không chệch của μ.

Thật vậy, vì các X_i có cùng phân phối xác suất với X nên $E(X_i) = \mu, i = \overline{1,n}$. Khi đó,

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu.$$

Ví dụ 2. Xét dấu hiệu X trên tổng thể có $E(X) = \mu$ và $D(X) = \sigma^2$. Từ X ta lập mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$. Khi đó, thống kê $S^2_{khc} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ là một ước lượng chệch của σ^2.

Thật vậy, S^2_{khc} có thể biến đổi tương đương thành

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

Do đó

$$E(S^2_{khc}) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu)^2 - E(\overline{X} - \mu)^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Vì $E(S^2_{khc}) \neq \sigma^2$ nên

$$S^2_{khc} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

là một ước lượng chệch của σ^2.

Theo cách tương tự, ta có thể chứng minh được

- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ là ước lượng không chệch của σ^2.
- F là ước lượng không chệch của p.

Ghi chú. Do không đạt tiêu chẩn này nên S^2_{khc} được gọi là phương sai mẫu chưa hiệu chỉnh, còn S^2 ngược lại được gọi là phương sai mẫu (hiệu chỉnh). Một ước lượng không chệch là một ước lượng tốt theo nghĩa nó có kỳ vọng đúng bằng tham số cần ước lượng. Nếu tồn tại

nhiều ước lượng không chệch cho một tham số, thì ước lượng nào có phương sai bé hơn sẽ được đánh giá là ước lượng tốt hơn. Từ đây ta có khái niệm ước lượng hiệu quả.

2. Ước lượng hiệu quả:

Định nghĩa. Thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ được gọi là một ước lượng hiệu quả của tham số θ nếu

a) $\hat{\theta}$ là một ước lượng không chệch của tham số θ,

b) $\hat{\theta}$ có $D(\hat{\theta})$ là nhỏ nhất so với các ước lượng không chệch khác được xây dựng trên cùng mẫu ngẫu nhiên.

Ví dụ. Xét dấu hiệu X trên tổng thể, $X \sim N(\mu, \sigma^2)$ và $(X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên có được từ tổng thể. Khi đó $\overline{X}$ là một ước lượng hiệu quả của μ.

Thật vậy, trong phần trước ta đã chứng minh được $\overline{X}$ là một ước lượng không chệch của μ. Tiếp theo ta cần chứng minh $D(\overline{X}) = \sigma^2 / n$ là bé nhất.Ta có hàm mật độ chuẩn của X là

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

Khi đó

$$\frac{\partial \ln\left(f(x,\mu)\right)}{\partial \mu} = \frac{\partial\left[-\frac{(x-\mu)^2}{2\sigma^2} - \ln\left(\sigma\sqrt{2\pi}\right)\right]}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

và

$$E\left[\frac{X-\mu}{\sigma^2}\right]^2 = \frac{1}{\sigma^4} E(X-\mu)^2 = \frac{1}{\sigma^2}.$$

Xét T là một thống kê không chệch bất kỳ của μ. Dùng bất đẳng thức Cramer-Rao (tham khảo tài liệu [6]) ta có:

$$D(T) \geq \frac{1}{nE\left(\frac{\partial \ln\left(f(x,\mu)\right)}{\partial \mu}\right)} = \frac{\sigma^2}{n} = D(\overline{X}).$$

Vậy $D(\overline{X}) = \sigma^2 / n$ là bé nhất, nên $\overline{X}$ là một ước lượng hiệu quả của μ.

Ngoài ra, ta có thể chứng minh được

- S^2 là ước lượng hiệu quả của σ^2.
- F là ước lượng hiệu quả của p.

3. Ước lượng vững:

Định nghĩa. Thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ được gọi là một ước lượng vững của tham số θ nếu với mọi $\varepsilon > 0$ ta có

$$\lim_{n \to +\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1.$$

Khi này ta nói thống kê $\hat{\theta}$ *hội tụ theo xác suất* đến tham số θ khi cỡ mẫu tiến về vô cùng (tức kích thước mẫu lớn).

Ví dụ. Xét dấu hiệu X trên tổng thể có $E(X) = \mu$ và $(X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên lấy từ tổng thể. Khi đó, thống kê $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ là một ước lượng vững của μ.

Để chứng minh khẳng định này, ta cần dùng bất đẳng thức Trêbusep với $\overline{X}$, ta có:

$$P\left(\left|\overline{X} - \mu\right| < \varepsilon\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

mà

$$\lim_{n \to +\infty}\left(1 - \frac{\sigma^2}{n\varepsilon^2}\right) = 1, \forall \varepsilon > 0.$$

Suy ra $\lim_{n \to +\infty} P\left(\left|\overline{X} - \mu\right| < \varepsilon\right) = 1$, và do đó $\overline{X}$ là một ước lượng vững của μ.

Một số ước lượng vững khác đã được chứng minh (xem tài liệu [6]):

- S^2, S^2_{khc} là ước lượng hiệu quả của σ^2.
- F là ước lượng hiệu quả của p.

3.4. Ước lượng khoảng

Giả sử ĐLNN X của tổng thể có đặc trưng số θ chưa xác định và cần ước lượng. Dựa vào mẫu ngẫu nhiên, ta chỉ ra khoảng (θ_1, θ_2) sao cho nó chứa θ với xác suất cao, tức là

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha \text{ với } \alpha > 0 \text{ đủ bé.}$$

Cách ước lượng như vậy gọi là cách ước lượng khoảng. Khoảng (θ_1, θ_2) gọi là *khoảng tin cậy* (confidence interval) và $1 - \alpha$ gọi là *độ tin cậy* (confidence level) của ước lượng trên.

Trong khuôn khổ giáo trình, ta chỉ xét bài toán ước lượng cho trung bình một tổng thể, bài toán ước lượng cho tỉ lệ một tổng thể, và bài toán ước lượng cho phương sai một tổng thể.

3.4.1. Bài toán ước lượng khoảng cho trung bình tổng thể

Xét ĐLNN X trên một tổng thể lớn và trị trung bình $E(X) = m$ chưa xác định. Ta cần ước lượng khoảng tin cậy của μ với độ tin cậy $1 - \alpha$, $\alpha > 0$ đủ bé.

Giả sử ta thu được mẫu ngẫu nhiên kích thước n của X là $W = (X_1, X_2, \ldots, X_n)$. Khi đó thống kê K được xây dựng theo ba trường hợp sau.

<u>Trường hợp 1:</u> X có phân phối chuẩn và không biết σ: Ta có thống kê

$$K = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

và K được chứng minh là có phân phối Student $T(n-1)$.

<u>Trường hợp 2:</u> Kích thước mẫu n đủ lớn: Ta có thống kê

$$K = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \text{ nếu biết } \sigma \text{ hoặc } K = \frac{\overline{X} - \mu}{S / \sqrt{n}} \text{ nếu không biết } \sigma$$

và K được chứng minh là xấp xỉ phân phối chuẩn tắc. Khi X có phân phối cân đối hoặc khá cân đối (là trường hợp thường gặp trong thực tế), ta có thể coi $n \geq 30$ là đủ lớn.

<u>Trường hợp 3:</u> *X* có phân phối chuẩn và biết σ: Ta có thống kê

$$K = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

và K được chứng minh là có phân phối chuẩn tắc.

Nhận xét. Bằng cách xây dựng khoảng ngẫu nhiên chứa μ, người ta đưa ra **quy tắc thực hành để tìm khoảng tin cậy $1-\alpha$ của trung bình tổng thể** như sau

- Lấy mẫu cụ thể $(x_1, x_2, \ldots, x_n)$ kích thước n của ĐLNN X, tính $\overline{x}$ và tính cả s nếu không biết độ lệch chuẩn σ.
- Tính độ chính xác

$\varepsilon = K_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, khi biết độ lệch chuẩn σ

$\varepsilon = K_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$, khi không biết độ lệch chuẩn σ

- Khoảng tin cậy (đối xứng) $1-\alpha$ của μ là $\left(\overline{x} - \varepsilon, \overline{x} + \varepsilon\right)$.

Trong đó, $K_{\alpha/2}$ được tra từ bảng phân vị với các trường hợp sau

<u>TH1:</u> *X* phân phối chuẩn và không biết σ:

$K_{\alpha/2} = t_{\alpha/2}^{(n-1)}$ tra từ bảng phân vị Student bậc $n-1$.

<u>TH2&3:</u> *n* đủ lớn ($n \geq 30$) hay *X* phân phối chuẩn & biết σ:

$K_{\alpha/2} = Z_{\alpha/2}$ tra từ bảng phân vị chuẩn tắc.

Ghi chú. Xét trên tập tất cả khoảng tin cậy $(1-\alpha)$ cụ thể của μ, tỉ lệ khoảng có chứa μ là $1-\alpha$. Do đó khi α đủ bé, với khả năng cao, khoảng tin cậy cụ thể có chứa μ.

Ví dụ 1. Điều tra năng suất lúa trên 100 ha lúa được chọn ngẫu nhiên của một vùng, người ta tính được trung bình mẫu cụ thể $\overline{x} = 46$ tạ và độ lệch mẫu cụ thể $s = 3,3$ tạ. Hãy ước lượng năng suất lúa trung bình cả vùng với độ tin cậy 95%.

Giải. Gọi X (tạ) là năng suất lúa trên một ha ngẫu nhiên của vùng đó. Khi đó $\mu \equiv E(X)$ là năng suất lúa trung bình toàn vùng. Ta cần ước lượng μ với độ tin cậy $1-\alpha = 0,95$ $(\alpha = 0,05)$. (Ở đây, dấu hiệu là năng suất; tổng thể là tập các ha lúa trong vùng đó.) Ta có

$$n = 100 > 30;\ \overline{x} = 46,\ \ s = 3,3.$$

Đây là TH2. Khi đó, $K_{\alpha/2} = Z_{0,025} = 1,96$. Độ chính xác

$$\varepsilon = K_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 1,96 \cdot \frac{3,3}{\sqrt{100}} = 0,6468.$$

Do đó, khoảng tin cậy 95% của μ là

$$(\overline{x} - \varepsilon; \overline{x} + \varepsilon) = (46 - 0,6468; 46 + 0,6468) = (45,3532; 46,6468).$$

Như vậy, năng suất trung bình cả vùng với độ tin cậy 95% là khoảng 45,3532 đến 46,6468 (tạ/ha).

Ví dụ 2. Lấy ngẫu nhiên 25 bao thuốc bột do một công ty dược sản xuất. Ta tính được trung bình mẫu cụ thể 39,8 g và phương sai mẫu cụ thể là 0,144. Biết trọng lượng $X(g)$ của bao thuốc bột ngẫu nhiên là ĐLNN có phân phối chuẩn.

a) Hãy ước lượng trọng lượng trung bình của các bao thuốc do công ty sản xuất với độ tin cậy 95%.

b) Hãy ước lượng trọng lượng trung bình của các bao thuốc do công ty sản xuất với độ tin cậy 95% . Biết thêm độ lệch chuẩn $\sigma = 0,4(g)$.

Giải. Gọi $\mu(g)$ là trọng lượng trung bình bao thuốc bột do công ty sản xuất. Ở đây, dấu hiệu là trọng lượng; tổng thể là tập các bao thuốc bột của công ty được.

a) Ước lượng μ với độ tin cậy $1-\alpha = 0,95$ $(\alpha = 0,05)$: Ta có

$$n = 100 > 30,\ \overline{x} = 39,8,\ \ s = \sqrt{0,144}.$$

Đây là TH1 do X phân phối chuẩn và σ^2 chưa biết. Khi đó, $K_{\alpha/2} = t^{(24)}_{0,025} = 2,0639$. Độ chính xác:

$$\varepsilon = K_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 2,0639 \cdot \frac{\sqrt{0,144}}{\sqrt{25}} = 0,1566.$$

Khoảng tin cậy 95% của μ là

$$(\overline{x} - \varepsilon; \overline{x} + \varepsilon) = (39,8 - 0,1566; 39,8 + 0,1566) = (39,6434; 39,9566).$$

Như vậy, trọng lượng trung bình các bao thuốc với độ tin cậy 95% là khoảng $39,6434$ đến $39,9566(g)$.

b) Ước lượng μ với độ tin cậy $1-\alpha = 0,95$ $(\alpha = 0,05)$: Ta có

$$n = 25 < 30,\ \overline{x} = 39,8,\ \ \sigma = \sqrt{0,144}.$$

Đây là TH3 do X có phân phối chuẩn và biết σ. Khi đó, $K_{\alpha/2} = Z_{0,025} = 1,96$. Độ chính xác:

$$\varepsilon = K_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1,96 \cdot \frac{0,4}{\sqrt{25}} = 0,1568$$

Khoảng tin cậy 95% của μ là

$$(\overline{x} - \varepsilon, \overline{x} + \varepsilon) = (39,8 - 0,1568; 39,8 + 0,1568) = (39,6432; 39,9568).$$

$$= (39,6432; 39,9568)$$

Như vậy, trọng lượng trung bình các bao thuốc với độ tin cậy 95% là khoảng $39,6432$ đến $39,9568(g)$.

3.4.2. Bài toán ước lượng khoảng cho tỉ lệ tổng thể

Xét một tổng thể lớn với tỉ lệ có tính chất A là số p chưa biết. Ta cần ước lượng tỉ lệ p với độ tin cậy $1-\alpha$, với $\alpha > 0$ đủ bé.

Giả sử ta thu được mẫu ngẫu nhiên kích thước n của tổng thể với n đủ lớn. Cho F là tỉ lệ mẫu. Thống kê K được xây dựng:

$$K = \frac{F - p}{\sqrt{p(1-p)}/\sqrt{n}}$$

và được chứng minh là xấp xỉ phân phối chuẩn tắc. Bằng cách xây dựng khoảng ngẫu nhiên chứa p, người ta đưa đến quy tắc thực hành tìm khoảng tin cậy $1-\alpha$ của tỉ lệ tổng thể p như sau

- Lấy mẫu cụ thể kích thước n đủ lớn và yêu cầu $nf > 10$ và $n(1-f) > 10$.
- Tính giá trị cụ thể tỉ lệ mẫu f.
- Tính độ chính xác:

$$\varepsilon = Z_{\alpha/2}.\frac{\sqrt{f(1-f)}}{\sqrt{n}}$$

- Khoảng tin cậy $1-\alpha$ của tỉ lệ p là $(f-\varepsilon, f+\varepsilon)$.

Ví dụ. Nghiên cứu nhu cầu tiêu dùng một mặt hàng của các người dân ở một thành phố, người ta điều tra trên 1.000 người được chọn ngẫu nhiên của thành phố và thấy có 600 người có nhu cầu. Hãy ước lượng tỉ lệ người có nhu cầu về mặt hàng đó trong toàn thành phố với độ tin cậy 95%.

Giải. Gọi p là tỉ lệ người ở thành phố đó có nhu cầu về mặt hàng này. Ta cần ước lượng p với độ tin cậy $1-\alpha = 0,95$ $(\alpha = 0,05)$. Ở đây, dấu hiệu là có nhu cầu; tổng thể là tập các người dân ở thành phố đó. Ta có

$$n = 1000 \text{ và } f = \frac{600}{1000} = 0,6$$

thỏa điều kiện $nf > 10$, $n(1-f) > 10$. Độ chính xác

$$\varepsilon = Z_{0,025} \cdot \frac{\sqrt{f(1-f)}}{\sqrt{n}} = 1,96 \cdot \frac{\sqrt{0,6 \times 0,4}}{\sqrt{1000}} = 0,0304.$$

Vậy khoảng tin cậy 95% của p là

$$(f-\varepsilon; f+\varepsilon) = (0,6-0,0304; 0,6+0,0304) = (0,5696; 0,6304).$$

Như vậy, với độ tin cậy là 95%, tỉ lệ người trong thành phố có nhu cầu là 56,96% đến 63,04%.

3.4.3. Bài toán ước lượng khoảng cho phương sai tổng thể

Giả sử ĐLNN X có phân phối chuẩn và chưa biết phương sai σ^2. Ta cần ước lượng khoảng tin cậy cho phương sai σ^2 với độ tin cậy $1-\alpha$ với α đủ bé.

Xét trường hợp không biết trung bình tổng thể E(X) như sau Từ X ta lập mẫu ngẫu nhiên kích thước n là $(X_1, X_2, \ldots, X_n)$. Thống kê được xây dựng cho bài toán này là

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}.$$

Thống kê này được chứng minh có phân phối χ^2 với $n-1$ bậc tự do. Từ đây, qui tắc thực hành tìm khoảng tin cậy của phương sai $D(X) = \sigma^2$ được đưa ra như sau

- Lấy mẫu cụ thể $(x_1, x_2, \ldots, x_n)$ kích thước n của ĐLNN X,
- Khoảng tin cậy của phương sai là

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}};\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right)$$

với $\chi^2_{\alpha/2}$ và $\chi^2_{1-\alpha/2}$ được tra từ bảng phân vị χ^2 với bậc tự do $n-1$.

Ví dụ. Lượng hao phí nguyên liệu cho một đơn vị sản phẩm là ĐLNN X có phân phối chuẩn. Quan sát ngẫu nhiên 25 sản phẩm, ta có số liệu cho ở bảng.

Träng l- î ng nguy^a n li Öu hao phi (*g*)	19,5	20,0	20,5
Sè s¶n phÈm	5	18	2

Với độ tin cậy 90%, hãy ước lượng phương sai σ^2.

Giải. Ước lượng phương sai $D(X)$ khi không biết $E(X)$. Nhắc lại rằng, khoảng tin cậy là

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}};\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right).$$

Ta có:

$$1-\alpha=0,9\Rightarrow\alpha=0,1;\alpha/2=0,05;1-\alpha/2=0,95.$$

Lập bảng tính

x_i	n_i	n_ix_i	$n_ix_i^2$
19,5	5	97,5	1.901,25
20,0	18	360	7.200
20,5	2	41	840,5
Σ	$n=25$	498,5	9.941,75

Tra bảng χ^2 với bậc $n-1=24$, ta có $\chi^2_{\alpha/2}=36,42$ và $\chi^2_{1-\alpha/2}=13,85$. Ta cần tính

$$\overline{x}=\frac{498,5}{25}=19,94;\ s^2=\frac{1}{24}(9941,75-25\times19,94^2)=0,0692;$$

$$(n-1)s^2=24\times0,0692=1,66.$$

Suy ra khoảng tin cậy

$$\left(\frac{1,66}{36,42};\frac{1,66}{13,85}\right)=(0,046;0,12).$$

Vậy phương sai của X khoảng 0,046 đến 0,12 với độ tin cậy 90%.

3.5. Các bài toán liên quan đến bài toán ước lượng

3.5.1. Bài toán xác định cỡ mẫu

Để đơn giản, ở đây ta chỉ xét hai lớp bài toán ước lượng trung bình tổng thể và ước lượng tỉ lệ tổng thể. Hơn nữa trong đó ta chỉ xét trường hợp kích thước mẫu đủ lớn và chấp nhận quan điểm đơn giản luôn có hệ số $K_{\alpha/2}=Z_{\alpha/2}$ tra ở bảng phân vị chuẩn tắc. Tùy theo bài toán cụ thể, ta chọn 1 trong 3 dạng phương trình sau

a) $\varepsilon=Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$; b) $\varepsilon=Z_{\alpha/2}.\frac{s}{\sqrt{n}}$; c) $\varepsilon=Z_{\alpha/2}.\frac{\sqrt{f(1-f)}}{\sqrt{n}}$.

Với phương trình b), người ta thường coi s mới bằng s cũ nếu không có s mới. Còn với phương trình c) thường coi f mới bằng f cũ nếu không có f mới. Đặc biệt, khi không có cơ sở nào để xấp xỉ f, ta coi $f = 0,5$ và lúc này

$$n = 0,25.\left(\frac{Z_{\alpha/2}}{\varepsilon}\right)^2.$$

Từ đó giải ra $n = a$ là số thực dương nào đó. Chọn kích thước mẫu là số tự nhiên n_1 vừa đủ lớn hơn hoặc bằng số thực a ở trên.

Ví dụ 1. Giả sử cần ước lượng năng suất lúa trung bình cho một vùng lớn. Sau khi điều tra $400\,ha$ lúa chọn ngẫu nhiên trong vùng, người ta có độ lệch mẫu $s = 6,6327$ và độ chính xác $\varepsilon = 0,65$ khi ước lượng khoảng với độ tin cậy 95%. Tìm số ha lúa phải điều tra thêm để độ chính xác như cũ cũ và có độ tin cậy là 99%.

Giải. Với $1-\alpha = 0,99$ $(\alpha = 0,01)$, $\varepsilon = 0,65$, coi $s = 6,6327$. Tìm n bằng cách xét

$$\varepsilon = Z_{0,005}.\frac{s}{\sqrt{n}} \Leftrightarrow 0,65 = 2,5758.\frac{6,6327}{\sqrt{n}}$$

$$\Rightarrow n = \left(\frac{2,5758 \times 6,6327}{0,65}\right)^2 \approx 690,84.$$

Chọn kích thước mẫu $n_1 = 691$. Khi đó, số ha lúa phải điều tra thêm là

$$n_1 - 400 = 291(ha).$$

Ví dụ 2. Nghiên cứu nhu cầu tiêu dùng một mặt hàng ở một thành phố, người ta điều tra trên 1000 người của thành phố và biết tỉ lệ người trong thành phố có nhu cầu về mặt hàng này là 57% đến 63% với độ tin cậy là 95%. Muốn độ chính xác như cũ và tăng độ tin cậy lên 99% thì số người phải điều tra thêm là bao nhiêu?

Giải. Biết khoảng tin cậy 95% của p là $(f-\varepsilon, f+\varepsilon) = (0,57; 0,63)$. Trước hết, ứng với mẫu cũ, ta tìm f và ε

$$f = \frac{0,63+0,57}{2} = 0,6 \quad \text{và} \quad \varepsilon = \frac{0,63-0,57}{2} = 0,03.$$

Ta có $1-\alpha = 0,99$ $(\alpha = 0,01)$, $\varepsilon = 0,03$, coi $f = 0,6$. Tìm n bằng cách xét

$$\varepsilon = Z_{0,005}.\frac{\sqrt{f(1-f)}}{\sqrt{n}} \Leftrightarrow 0,03 = 2,5758.\frac{\sqrt{0,6 \times 0,4}}{\sqrt{n}}$$

$$\Rightarrow n = \left(\frac{2,5758}{0,03}\right)^2 \times 0,6 \times 0,4 = 1769,27.$$

Chọn kích thước mẫu $n_1 = 1770$. Vậy số người phải điều tra thêm là

$$1770 - 1000 = 770 \text{ (người)}.$$

3.5.2. Bài toán xác định độ tin cậy

Tùy theo bài toán cụ thể, ta chọn 1 trong 3 dạng phương trình

a) $\varepsilon = Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$; b) $\varepsilon = Z_{\alpha/2}.\frac{s}{\sqrt{n}}$; c) $\varepsilon = Z_{\alpha/2}.\frac{\sqrt{f(1-f)}}{\sqrt{n}}$.

Giả sử ta giải ra được $Z_{\alpha/2} = a$ là số thực nào đó. Khi đó ta suy ra

$$\alpha/2 = P(Z > a) = \varphi(+\infty) - \varphi(a) = 0,5 - \varphi(a).$$

Từ đó sẽ tìm được α và độ tin cậy $1 - \alpha$.

Ví dụ 1. Để ước lượng trọng lượng trung bình sản phẩm của nhà máy, người ta điều tra 121 sản phẩm, được độ lệch mẫu cụ thể $s = 3,9115(kg)$ kg và trung bình mẫu cụ thể $98\,kg$. Khi ước lượng với độ chính xác là $1\,kg$ thì độ tin cậy là bao nhiêu?

Giải. Ta có $n = 121, s = 3,9115, \varepsilon = 1$. Ta cần xác định $1 - \alpha$ như sau

$$\varepsilon = Z_{\alpha/2}.\frac{s}{\sqrt{n}} \Leftrightarrow 1 = Z_{\alpha/2}.\frac{3,9115}{\sqrt{121}} \Leftrightarrow Z_{\alpha/2} = \frac{1 \times 11}{3,9115} = 2,8122.$$

Suy ra

$$\alpha/2 = P(Z > 2,8122) = \varphi(+\infty) - \varphi(2,8122) = 0,5 - \varphi(2,81) = 0,0025.$$

$$\Rightarrow \alpha = 0,005 \quad \text{và} \quad 1 - \alpha = 0,995.$$

Vậy độ tin cậy là $99,5\%$.

Ví dụ 2. Nghiên cứu nhu cầu tiêu dùng một mặt hàng ở một thành phố, người ta điều tra trên 1000 người được chọn ngẫu nhiên và được tỉ lệ mẫu $f = 60\%$. Ước lượng tỉ lệ người (trong thành phố) có nhu cầu về mặt hàng đó với độ chính xác 5% thì độ tin cậy bao nhiêu?

Giải. Ta có $f = 0,6; \varepsilon = 0,05; n = 1000$. Ta cần tìm $1 - \alpha$ như sau

$$\varepsilon = Z_{\alpha/2} \cdot \frac{\sqrt{f(1-f)}}{\sqrt{n}} \Leftrightarrow 0,05 = Z_{\alpha/2}.\frac{\sqrt{0,6 \times 0,4}}{\sqrt{1000}} \Leftrightarrow Z_{\alpha/2} = \frac{0,05 \times \sqrt{1000}}{\sqrt{0,6 \times 0,4}} = 3,2275.$$

Suy ra

$$\alpha/2 = P(Z > 3,2275) = 0,5 - \varphi(3,23) = 0,5 - 0,4994 = 0,0006.$$

$$\Rightarrow \alpha = 0,0012 \quad \text{và} \quad 1 - \alpha = 0,9988.$$

Vậy độ tin cậy là $99,88\%$.

4.6. Bài tập chương 3

1. Quan sát ngẫu nhiên về thời gian cần thiết để sản xuất một sản phẩm trong một phân xưởng, ta thu được các số liệu cho ở bảng

Khoảng thời gian (giây)	Số quan sát	Khoảng thời gian (giây)	Số quan sát
20 – 22	2	26 – 28	32
22 – 24	10	28 – 30	14
24 – 26	34	30 – 32	8

Ước lượng thời gian trung bình để sản xuất 1 sản phẩm của phân xưởng với độ tin cậy 95%.

2. Quan sát ngẫu nhiên về thời gian cần thiết để sản xuất một chi tiết máy do một máy sản xuất , ta thu được các số liệu cho ở bảng

Khoảng thời gian (giây)	Số quan sát	Khoảng thời gian (giây)	Số quan sát
122 – 124	8	128 – 130	30
124 – 126	10	130 – 132	14
126 – 128	32	132 – 134	6

Ước lượng thời gian trung bình để sản xuất một chi tiết máy với độ tin cậy 95%.

3. Kiểm tra năng suất của 1 số hecta lúa chọn ngẫu nhiên ở một vùng, người ta thu được kết quả cho ở bảng

Năng suất (tấn/ha)	Diện tích (ha)	Năng suất (tấn/ha)	Diện tích (ha)
3,0 – 3,5	2	5,0 – 5,5	16
3,5 – 4,0	5	5,5 – 6,0	25
4,0 – 4,5	8	6,0 – 6,5	17
4,5 – 5,0	10	6,5 – 7,0	13

Những thửa ruộng có năng suất hơn 4,5 (tấn/ha) gọi là ruộng tốt. Ước lượng năng suất trung bình của ruộng tốt vùng đó với độ tin cậy 95%.

4. Để nghiên cứu nhu cầu của một loại hàng hóa ở một khu vực, người ta tiến hành khảo sát 800 hộ gia đình chọn ngẫu nhiên trong khu vực. Kết quả cho ở bảng dưới

Nhu cầu (kg/tháng)	Số gia đình (n_i)	Nhu cầu (kg/tháng)	Số gia đình (n_i)
35 – 40	10	50 – 55	230
40 – 45	50	55 – 60	226
45 – 50	140	60 – 65	144

Hộ có nhu cầu hơn 45 (kg/tháng) là hộ không nghèo. Hãy

a) Ước lượng tỉ lệ hộ không nghèo trong khu vực với độ tin cậy 95%.

b) Ước lượng nhu cầu trung bình mỗi hộ gia đình không nghèo trong khu vực với độ tin cậy 98%.

5. Chọn ngẫu nhiên các ngày bán hàng của công ty A, thống kê số hàng hóa bán được mỗi ngày (SHB/N) và số ngày bán được lượng hàng tương ứng, ta có bảng số liệu sau

SHB/N (kg)	Số ngày (n_i)	SHB/N (kg)	Số ngày (n_i)
150 – 200	5	400 – 450	110
200 – 250	19	450 – 500	95
250 – 300	26	500 – 550	60
300 – 350	40	550 – 600	30
350 – 400	65		

Những ngày bán được nhiều hơn $300(kg)$ là những ngày đạt doanh thu. Hãy

a) Ước lượng tỉ lệ ngày công ty A đạt doanh thu với độ tin cậy 90%.

b) Ước lượng lượng hàng bán trung bình mỗi ngày trong những ngày đạt doanh thu của công ty A với độ tin cậy 98%.

6. Để định mức thời gian gia công một chi tiết máy trong công ty, người ta theo dõi ngẫu nhiên thời gian gia công 25 chi tiết và thu được bảng số liệu sau đây

Thêi gian (phót)	14	16	18	20	22
Sè chi tiÕt	2	6	11	4	2

Với độ tin cậy 0,95, hãy ước lượng thời gian gia công trung bình loại chi tiết trên trong công ty. Cho biết thời gian X (phút) để gia công chi tiết đó trong công ty tuân theo quy luật chuẩn.

7. Giả sử điểm trung bình môn Toán của 100 thí sinh thi vào ĐHNT được chọn ngẫu nhiên là 5,0 với độ lệch mẫu là 2,5.

a) Ước lượng điểm trung bình môn Toán của toàn thể thí sinh vào ĐHNT với độ tin cậy 95%.

b) Với độ chính xác là 0,25 điểm, hãy xác định độ tin cậy.

8. Biết tuổi thọ X (giờ) của một loại bóng đèn do xí nghiệp A sản xuất tuân theo quy luật chuẩn với độ lệch chuẩn $\sigma = 100$ (giờ).

a) Chọn ngẫu nhiên 100 bóng để thử nghiệm, thấy trung bình tuổi thọ mỗi bóng là 1000 giờ. Hãy ước lượng tuổi thọ trung bình của bóng đèn xí nghiệp A sản xuất với độ tin cậy 95%.

b) Với độ chính xác 15 giờ, hãy xác định độ tin cậy.

c) Với độ chính xác 25 giờ và độ tin cậy 99% thì cần thử nghiệm bao nhiêu bóng.

9. Chiều dài của một loại sản phẩm A do một máy tự động sản xuất là một biến số ngẫu nhiên tuân theo quy luật chuẩn với độ lệch chuẩn là $3(cm)$. Phải chọn ít nhất bao nhiêu chi tiết để đo, nếu muốn độ dài khoảng tin cậy không vượt quá 0,6 và độ tin cậy của ước lượng là 0,99.

10. Để ước lượng trọng lượng trung bình sản phẩm của nhà máy, người ta điều tra 100 sản phẩm, được phương sai mẫu cụ thể $s^2 = 16$ và trung bình mẫu cụ thể là 98 kg. Ước lượng với độ chính xác là 1 kg thì độ tin cậy là bao nhiêu?

11. Nghiên cứu nhu cầu tiêu dùng một mặt hàng ở một thành phố, người ta điều tra trên 1.000 người được chọn ngẫu nhiên và thấy có 400 người có nhu cầu.

a) Hãy ước lượng tỉ lệ người có nhu cầu về mặt hàng đó trong toàn thành phố với độ tin cậy 95%.

b) Muốn độ tin cậy lên đến 98% thì phải điều tra thêm bao nhiêu người.

12. Để điều tra số cá trong một hồ, người ta đánh bắt 1.000 con cá, đánh dấu rồi thả xuống hồ. Lần sau bắt lại 400 con thì được 40 con đánh dấu. Với độ tin cậy 95% hãy

a) Ước lượng tỷ lệ cá được đánh dấu trong hồ.

b) Ước lượng số cá trong hồ.

13. Cơ quan cảnh sát giao thông kiểm tra hệ thống phanh của 500 chiếc xe tải chạy trên đường quốc lộ được chọn ngẫu nhiên. Họ phát hiện ra 40 chiếc có phanh chưa đảm bảo an toàn. Tìm khoảng tin cậy 98% cho tỉ lệ xe tải có phanh chưa an toàn.

14. Trong đợt vận động bầu cử tổng thống người ta phỏng vấn ngẫu nhiên 1.600 cử tri thì được biết 960 người trong số đó bỏ phiếu cho ứng cử viên A. Với độ tin cậy 0,99, ước lượng tỉ lệ cử tri bỏ phiếu cho ứng cử viên A.

Bài tập làm thêm

15. Tỷ lệ nảy mầm của một loại hạt giống trong mẫu 400 hạt là 90%.

a) Ước lượng tỷ lệ nảy mầm của loại hạt giống đó với độ tin cậy 0,95.

b) Muốn độ dài khoảng tin cậy không vượt quá 0,02, độ tin cậy như cũ thì phải gieo bao nhiêu hạt?

16. Để ước lượng tỷ lệ sản phẩm xấu của một kho đồ hộp, người ta kiểm tra ngẫu nhiên 400 hộp thấy có 80 hộp xấu.

a) Ước lượng tỷ lệ sản phẩm xấu của kho đồ hộp với độ tin cậy 95%.

b) Với mẫu trên, biết độ chính xác 3%, hãy xác định độ tin cậy.

17. Nghiên cứu nhu cầu tiêu dùng một mặt hàng ở một thành phố, người ta điều tra trên 2.400 người của thành phố và biết tỉ lệ người trong thành phố có nhu cầu về mặt hàng này là 48% đến 52% với độ tin cậy là 95%.

a) Với độ dài khoảng tin cậy như cũ, muốn tăng độ tin cậy lên 96% thì số người phải điều tra thêm là bao nhiêu?

b) Với độ chính xác như cũ và số người điều tra là 3000 người, độ tin cậy là bao nhiêu?

18. Để xác định tỷ lệ người mắc chứng bướu cổ do thiếu hụt i-ốt ở một khu vực dân cư, cần khám bao nhiêu người nếu muốn cho khoảng tin cậy có

a) Độ chính xác không vượt quá 0,04 với độ tin cậy 95%?

b) Độ chính xác không vượt quá 0,02 và độ tin cậy 98%?

Đáp số và hướng dẫn

1. Khoảng tin cậy 95% của μ là (25,9525; 26,8475).

2. Khoảng tin cậy 95% của μ là (127,5064; 128,4936).

3. Khoảng tin cậy 95% của μ là (5,6574; 5,9290).

4. a) Khoảng tin cậy 95% của p là (0,9067; 0,9433); b) Khoảng tin cậy 98% của μ là (54,5954; 55,4586).

5. a) Khoảng tin cậy 90% của p là (0,8645; 0,9133); b) Khoảng tin cậy 98% của μ là (436,9419; 453,0581).

6. Khoảng tin cậy 95% của μ là (16,9834; 18,6966).

7. a) Khoảng tin cậy 95% của μ là (4,51; 5,49); b) $1-\alpha = 0,6826$.

8. a) Khoảng tin cậy 95% của μ là (980,4; 1019,6). b) $1-\alpha = 0,8664 = 86,64\%$; c) 107 (bóng đèn).

9. 664 (chi tiết).

10. $1-\alpha = 0,9876 = 98,76\%$

11. a) Khoảng tin cậy 95% của p là (0,3696; 0,4304); b) Điều tra thêm 406 (người).

12. a) Khoảng tin cậy 95% của p là (0,0706; 0,1294); b) Gọi N là số cá trong hồ. Ta có $7728 \le N \le 14164$.

13. Khoảng tin cậy 98% của p là (0,0518; 0,1082).

14. Khoảng tin cậy 99% của p là (0,5685; 0,6315).

15. a) Khoảng tin cậy 95% của p là (0,8706; 0,9294); b) 3458 (hạt).

16. a) Khoảng tin cậy 95% của p là (0,1608; 0,2392); b) $1-\alpha = 0,8684 = 86,64\%$.

17. $f = 0,5; \varepsilon = 0,02$.

a) Số người phải điều tra thêm là 2637−2400 = 237 (người); b) $1-\alpha = 0,9714 = 97,14\%$.

18. Do không có cơ sở để ước lượng f, ta cho $f = 0,5$.

a) Chọn kích thước mẫu $n_1 = 601$ (người); b) Chọn kích thước mẫu $f = 3.383$ (người)

Chương 4.
Kiểm định giả thuyết thống kê

4.1. Các khái niệm

Bài toán kiểm định giả thuyết thống kê (Statistical hypothesis testing) là một lớp bài toán quan trọng trong thống kê toán. Nói chung trong lớp bài toán này người ta cố gắng đưa đến các kết luận với khả năng oan sai là nhỏ.

Xét ví dụ sau Có một máy đóng các bao gạo xuất khẩu hoạt động một thời gian dài và người ta nghi ngờ rằng hiện nay trung bình lượng gạo mỗi bao do máy đóng không đúng trọng lượng qui định là 50 (kg). Bạn là người đi kiểm tra. Bạn chọn ngẫu nhiên 100 bao gạo để cân và xác định được trung bình mẫu, độ lệch chuẩn mẫu cụ thể nào đó. Bạn sẽ kết luận như thế nào về tình trạng hoạt động của máy sau khi có các số liệu đó?

Gọi $X\,(kg)$ là trọng lượng mỗi bao gạo do máy đó đóng hiện nay. Có hai khả năng xảy ra:

H_0: Máy hoạt động coi như bình thường $(E(X) = 50\,kg)$.

H_1: Máy đã trục trặc $(E(X) \neq 50\,kg)$.

Ta gọi đây là các giả thuyết thống kê. Ta cần kiểm định các giả thuyết này. Để làm điều này, người ta sẽ thực hiện các bước sau.

- Xây dựng biến cố *A* sao cho *A* khó xảy ra nếu H_0 là đúng. Tức là $P(A) = \alpha$ đủ bé nếu H_0 đúng.

- Kiểm tra thực tế biến cố *A* có xảy ra không?

Trường hợp 1: Nếu *A* xảy ra thì H_0 khó đúng được nên ta kết luận: bác bỏ H_0. Ta phân tích chi tiết hơn: nếu thực sự H_0 là sai thì kết luận bác bỏ H_0 là chính xác; còn nếu thực sự H_0 là đúng thì kết luận bác bỏ H_0 là oan sai với khả năng oan sai $\leq \alpha$.

Trường hợp 2: Nếu *A* không xảy ra thì tạm chấp nhận H_0 vì không đủ cơ sở bác bỏ.

4.1.1. Giả thuyết thống kê

Giả thuyết thống kê ở đây là các giả thuyết về các tham số đặc trưng hoặc qui luật phân phối xác suất của các đại lượng ngẫu nhiên hoặc tính độc lập của các dấu hiệu.

Việc tìm ra kết luận là chấp nhận hoặc bác bỏ một giả thuyết gọi là kiểm định giả thuyết thống kê. Giả thuyết cần kiểm định gọi là giả thuyết không, ký hiệu là H_0. Mệnh đề phủ định với H_0 được gọi là giả thuyết đối và ký hiệu là H_1.

4.1.2. Tiêu chuẩn kiểm định, mức ý nghĩa, miền bác bỏ, sai lầm loại 1 và sai lầm loại 2

Giả sử cần kiểm định giả thuyết:

$$H_0 : \theta = \theta_0; H_1 : \theta \neq \theta_0$$

với θ là một tham số (chưa biết được giá trị cụ thể) của đại lượng ngẫu nhiên X trên một tổng thể.

Lấy mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ có kích thước n của đại lượng ngẫu nhiên X. Xây dựng tiêu chuẩn kiểm định là một thống kê

$$K = f(X_1, X_2, ..., X_n, \theta_0)$$

sao cho nếu H_0 đúng thì ta có được hai điều

- Xác định được qui luật phân phối của *K* hoặc là xấp xỉ tốt phân phối của *K*.

- Ứng với mỗi mẫu cụ thể $(x_1, x_2, ..., x_n)$, ta tính được giá trị cụ thể của *K* là

$$k = f(x_1, x_2 \ldots, x_n, \theta_0).$$

Chọn miền bác bỏ là miền W_a sao cho biến cố bác bỏ $A \equiv (K \in W_\alpha)$ là rất khó xảy ra nếu H_0 là đúng. Cụ thể là $P(K \in W_\alpha) = \alpha$, với $\alpha > 0$ đủ bé nếu H_0 là đúng. Trị α này được gọi là mức ý nghĩa (loại 1) của kiểm định. Tiến hành kiểm tra

a) Trường hợp $k \in W_\alpha$: Biến cố khó xảy ra ($k \in W_\alpha$) đã xảy ra. Ta bác bỏ giả thuyết H_0 và thừa nhận H_1.

Giả định mẫu cụ thể là đảm bảo tính ngẫu nhiên. Nếu H_0 là đúng thì việc bác bỏ H_0 này là oan sai với mức khả năng oan sai là α. Như vậy, sai lầm loại 1 là sai lầm mắc phải khi ta bác bỏ H_0 nhưng trong thực tế H_0 là đúng.

Trong điều kiện lý tưởng, mức xác suất mắc sai lầm loại 1 là $P(K \in W_\alpha) = \alpha$ và chính là mức ý nghĩa α.

b) Trường hợp $k \notin W_\alpha$: Biến cố khó xảy ra ($k \in W_\alpha$) đã không xảy ra. Ta tạm chấp nhận giả thuyết H_0 vì chưa đủ cơ sở để bác bỏ H_0.

Như vậy, sai lầm loại 2 là sai lầm mắc phải khi ta chấp nhận H_0 nhưng trong thực tế H_0 là sai.

Trong điều kiện lý tưởng, khi xét bài toán kiểm định với mức ý nghĩa α, mức xác suất mắc sai lầm loại 2 là $\beta = P\left((K \notin W_\alpha) \mid H_1\right)$. Ta nói β là mức ý nghĩa (loại 2) của bài toán kiểm định.

4.2. Một số bài toán kiểm định thường gặp

4.2.1. Bài toán kiểm định giả thuyết trung bình tổng thể

Xét đại lượng ngẫu nhiên X trên một tổng thể lớn với trung bình tổng thể $\mu = E(X)$ chưa biết. Ta cần kiểm định giả thuyết

$$H_0 : \mu = m_0 \text{ và giả thuyết đối } H_1 \text{ với mức ý nghĩa } \alpha > 0 \text{ đủ bé.}$$

Lấy mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ kích thước n của X. Ta xây dựng tiêu chuẩn kiểm định K.

Trường hợp 1: X **có phân phối chuẩn và chưa biết** σ**:** Chọn tiêu chuẩn

$$K = \frac{\overline{X} - m_0}{S/\sqrt{n}}.$$

Giả sử H_0 đúng. Người ta chứng minh được K có phân phối Student bậc $n-1$.

Trường hợp 2: n **đủ lớn:** Trong giáo trình này ta chỉ xét các đại lượng ngẫu nhiên X cân đối hoặc khá cân đối và do đó có thể coi kích thước mẫu n là đủ lớn với $n \geq 30$. Ta chọn tiêu chuẩn

$$K = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \text{ (khi biết } \sigma) \text{ hay } K = \frac{\overline{X} - m_0}{S/\sqrt{n}} \text{ (khi không biết } \sigma).$$

Giả sử H_0 đúng. Người ta chứng minh được K xấp xỉ phân phối chuẩn tắc $N(0,1)$.

Trường hợp 3. X **có phân phối chuẩn và biết** σ**:** Chọn tiêu chuẩn

$$K = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}}.$$

Giả sử H_0 đúng. Ta chứng minh được K có phân phối chuẩn tắc $N(0,1)$.

Nhận xét (Quy tắc thực hành): Lấy mẫu cụ thể $(x_1, x_2, ..., x_n)$ kích thước n của X. Tính giá trị cụ thể $\overline{x}$ (và tính s nếu chưa biết σ). Tính giá trị cụ thể k của K

$$k = \frac{\overline{x} - m_0}{\sigma/\sqrt{n}} \text{ (nếu biết } \sigma) \text{ hoặc } k = \frac{\overline{x} - m_0}{s/\sqrt{n}} \text{ (nếu không biết } \sigma).$$

Giả thuyết	$H_0 : \mu = m_0$ $H_1 : \mu \neq m_0$	$H_0 : \mu = m_0$ $H_1 : \mu > m_0$	$H_0 : \mu = m_0$ $H_1 : \mu < m_0$
Miền bác bỏ W_α	$\lvert K \rvert > K_{\alpha/2}$	$K > K_\alpha$	$K < -K_\alpha$

Kiểm tra:

- Nếu $k \in W_\alpha$ thì bác bỏ H_0.
- Nếu $k \notin W_\alpha$ thì tạm chấp nhận H_0 vì chưa đủ cơ sở để bác bỏ H_0.

TH1: X có phân phối chuẩn và không biết σ:

$$K_{\alpha/2} = t_{\alpha/2}^{(n-1)};\ K_\alpha = t_\alpha^{(n-1)} \text{ tra từ bảng phân vị Student bậc } n-1.$$

TH2&3: Mẫu n đủ lớn $(n \geq 30)$ hay X phân phối chuẩn và biết σ.

$$K_{\alpha/2} = Z_{\alpha/2};\ K_\alpha = Z_\alpha \text{ tra từ bảng phân vị chuẩn tắc.}$$

Ví dụ 1. Trọng lượng $X(kg)$ mỗi bao hàng do một máy đóng bao sản xuất là đại lượng ngẫu nhiên phân phối chuẩn. Trọng lượng trung bình mỗi bao hàng đã qui định là $50\,kg$. Người ta cân thử 25 bao hàng và được $\overline{x} = 49,71(kg)$ và $s = 0,5(kg)$.

a) Nhà máy nói máy đóng các bao hàng đúng trọng lượng trung bình qui định. Với mức ý nghĩa 1% , bạn hãy kết luận ý kiến đó.

b) Với mức ý nghĩa 1%, hãy kết luận xem trọng lượng trung bình các bao hàng có nhỏ hơn qui định không?

c) Tìm khoảng tin cậy 99% của trọng lượng trung bình các bao hàng và nhận xét vị trí tương đối của nó so với giá trị $50kg$.

Giải. Gọi $\mu(kg)$ là trọng lượng trung bình của các bao hàng do máy đóng bao sản xuất.

a) Ý kiến $\mu \neq 50(kg)$: Kiểm định giả thuyết

$$H_0 : \mu = 50; H_1 : \mu \neq 50.$$

Do cỡ mẫu $n = 25(n < 30)$, trung bình mẫu $\overline{x} = 49,71(kg)$ và độ lệch chuẩn mẫu $s = 0,5(kg)$ nên đây là trường hợp 1 (chưa biết σ). Ta có tiêu chuẩn kiểm định

$$k = \frac{\overline{x} - 50}{s / \sqrt{25}} = \frac{(49,71 - 50)}{0,5 / 5} = -2,9.$$

và miền bác bỏ

$$W_a : \left|K\right| > K_{a/2} = t^{(24)}_{0,005} = 2,7969.$$

Kiểm tra thấy $k \in W_\alpha$ nên ta bác bỏ H_0. Vậy với mức ý nghĩa 1%, máy đã đóng các bao hàng có trọng lương trung bình thực tế không đúng qui định $(\mu \neq 50)$.

b) Ý kiến $\mu < 50$: Kiểm định giả thuyết

$$H_0 : \mu = 50; H_1 : \mu < 50.$$

Do cỡ mẫu $n = 25(n < 30)$, trung bình mẫu $\overline{x} = 49,71(kg)$ và độ lệch chuẩn mẫu $s = 0,5(kg)$ nên đây là trường hợp 1 (chưa biết σ). Ta có tiêu chuẩn kiểm định

$$\frac{\overline{x} - 50}{s / \sqrt{25}} = \frac{(49,71 - 50)}{0,5 / 5} = -2,9.$$

và miền bác bỏ

$$W_a : K < -K_a = -t^{(24)}_{0,01} = -2,4922.$$

Kiểm tra thấy $k \in W_\alpha$ nên ta bác bỏ H_0. Vậy với mức ý nghĩa 1%, các bao hàng do máy đóng có trọng lượng trung bình thực tế nhỏ hơn qui định $(\mu < 50)$.

c) Khoảng tin cậy 99% của μ: Tính

$$1 - \alpha = 0,99 \Rightarrow \alpha = 0,01.$$

Ta có

$$\varepsilon = t^{(24)}_{0,005}.\frac{s}{\sqrt{n}} = 2,7969.\frac{0,5}{\sqrt{25}} = 0,2797.$$

Suy ra khoảng tin cậy 99% của trọng lượng trung bình các bao hàng là

$$(\overline{x} - \varepsilon, \overline{x} + \varepsilon) = (49,71 - 0,2797; 49,71 + 0,2797) = (49,4303; 49,9897).$$

Khoảng này không chứa giá trị 50 và ở phía trái của 50 (phù hợp các kết luận ở a) và b)).

Ví dụ 2. Kiểm tra ngẫu nhiên 100 sản phẩm do một máy sản xuất thấy chiều dài trung bình của chúng là 150,5 (mm) và độ lệch mẫu là 5 (mm). Chiều dài qui định sản phẩm là 150 (mm).

a) Với mức ý nghĩa 10%, hãy kiểm tra ý kiến rằng chiều dài trung bình sản phẩm do máy sản xuất là đúng qui định.

b) Tìm khoảng tin cậy 90% của chiều dài trung bình sản phẩm và nhận xét vị trí tương đối của nó so với giá trị 150 (mm).

Giải.

a) Gọi $X(mm)$ là chiều dài mỗi sản phẩm ngẫu nhiên do máy đó sản xuất. và $\mu(mm)$ là chiều dài trung bình sản phẩm do máy sản xuất.

Ý kiến $\mu \neq 150$. Kiểm định giả thuyết

$$H_0 : \mu = 150; H_1 : \mu \neq 150.$$

Cỡ mẫu $n = 100$ $(n > 30)$, trung bình mẫu $\overline{x} = 150,5(mm)$ và độ lệch chuẩn mẫu $s = 5(mm)$ (Trường hợp 2). Tiêu chuẩn kiểm định

$$k = \frac{\overline{x} - m_0}{s / \sqrt{n}} = \frac{150,5 - 150}{5 / \sqrt{100}} = 1.$$

và miền bác bỏ

$$W_\alpha : \left|K\right| > Z_{0,05} = 1,6449.$$

Kiểm tra thấy $k \notin W_\alpha$ nên tạm chấp nhận giả thuyết H_0. Vậy với $a = 10\%$, tạm coi $\mu \neq 150(mm)$ vì chưa đủ cơ sở bác bỏ (tạm chấp nhận chiều dài trung bình sản phẩm do máy sản xuất là đúng qui định.)

b) Khoảng tin cậy 90% của μ: Tính

$$1 - \alpha = 0,9 \Rightarrow \alpha = 0,1.$$

Ta có

$$\varepsilon = Z_{0,05}.\frac{s}{\sqrt{n}} = 1,6449.\frac{5}{\sqrt{100}} = 0,8225.$$

Suy ra khoảng tin cậy 90% của chiều dài trung bình sản phẩm

$$(\overline{x} - \varepsilon, \overline{x} + \varepsilon) = (150,5 - 0,8225; 150,5 + 0,8225) = (149,6775; 151,3225)$$

Khoảng này chứa giá trị 150 (phù hợp kết luận câu a)).

4.2.2. Bài toán kiểm định giả thuyết tỉ lệ tổng thể

Xét một tổng thể lớn với tỉ lệ các phần tử có tính chất A của tổng thể là p chưa biết. Cần kiểm định giả thuyết

$H_0 : \mu = p_0$ và giả thuyết đối H_1 với mức ý nghĩa $\alpha > 0$ đủ bé.

Lấy mẫu ngẫu nhiên kích thước n. Cho F là tỉ lệ mẫu của mẫu n phần tử ngẫu nhiên. Giả sử H_0 là đúng. Cho kích thước mẫu n đủ lớn và thỏa yêu cầu $np_0 > 0$ và $n(1 - p_0) > 5$. Ta có F xấp xỉ phân phối chuẩn $N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$. Ta xây dựng tiêu chuẩn kiểm định

$$K = \frac{F - p_0}{\sqrt{p_0(1 - p_0)} / \sqrt{n}}.$$

Khi đó K xấp xỉ phân phối chuẩn tắc $N(0,1)$.

Nhận xét (Quy tắc thực hành): Lấy mẫu cụ thể n phần tử và tính tỉ lệ mẫu cụ thể f. Yêu cầu n đủ lớn, $nf > 5$ và $n(1-f) > 5$. Tính giá trị cụ thể k của K

$$k = \frac{f - p_0}{\sqrt{p_0(1-p_0)} / \sqrt{n}}.$$

Giả thuyết	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p = p_0$ $H_1 : p > p_0$	$H_0 : p = p_0$ $H_1 : p < p_0$
Miền bác bỏ W_α	$\lvert K \rvert > Z_{\alpha/2}$	$K > Z_\alpha$	$K < -Z_\alpha$

Kiểm tra:

- Nếu $k \in W_\alpha$ thì bác bỏ H_0.
- Nếu $k \notin W_\alpha$ thì tạm chấp nhận H_0 vì chưa đủ cơ sở để bác bỏ H_0.

Ví dụ 3. Tỉ lệ phế phẩm của một dây chuyền sản xuất là 5%. Sau khi tiến hành một cải tiến kỹ thuật người ta kiểm tra ngẫu nhiên 3.000 sản phẩm thì thấy có 120 phế phẩm.

a) Với mức ý nghĩa 1%, hãy kết luận việc cải tiến kỹ thuật có làm giảm tỉ lệ phế phẩm không?

b) Tìm khoảng tin cậy 98% của tỉ lệ phế phẩm mới.

Giải.

a) Gọi p là tỉ lệ phế phẩm mới của dây chuyền sau cải tiến. Ý kiến $p < 0,05$? Kiểm định giả thuyết

$$H_0 : p = 0,05;\ H_1 : p < 0,05.$$

Ta có cỡ mẫu $n = 3.000$, tỉ lệ mẫu $f = \frac{120}{3.000} = 0,04$ $(nf > 5,\ n(1-f) > 5)$. Tiêu chuẩn kiểm định

$$k = \frac{f - p_0}{\sqrt{p_0(1-p_0)} / \sqrt{n}} \frac{0,04 - 0,05}{\sqrt{0,05.(1-0,05)}}.\sqrt{3.000} = -2,513.$$

và miền bác bỏ

$$W_\alpha : K < -K_\alpha = -Z_{0,01} = -2,3263.$$

Kiểm tra thấy $k \notin W_\alpha$ nên ta bác bỏ giả thuyết H_0. Vậy với mức ý nghĩa 1%, cải tiến kỹ thuật có làm giảm tỉ lệ phế phẩm $(p < 5\%)$.

b) Ước lượng p với $1 - \alpha = 0,98 \Rightarrow \alpha = 0,02$: Cỡ mẫu $n = 3.000$. Tỉ lệ mẫu $f = \frac{120}{3.000} = 0,04$ $(nf > 10,\ n(1-f) > 10)$. Ta có

$$\varepsilon = Z_{0,01}.\frac{\sqrt{f(1-f)}}{\sqrt{n}} = 2,3263.\frac{\sqrt{0,04 \times 0,96}}{\sqrt{3000}} = 0,00083.$$

Suy ra khoảng tin cậy 98% của p là $(3,92\%; 4,08\%)$. Khoảng tin cậy này ở phía bên trái của 5% (phù hợp kết luận).

4.2.3. Bài toán so sánh hai trung bình tổng thể

Xét hai đại lượng ngẫu nhiên X và Y trên hai tổng thể với $E(X)$ và $E(Y)$ chưa biết. Cần kiểm định giả thuyết

$$H_0 : E(X) = E(Y) \text{ với mức ý nghĩa } \alpha .$$

Từ X lập mẫu ngẫu nhiên W_X kích thước n_1 và từ Y lập mẫu ngẫu nhiên W_Y kích thước n_2.

● **Trường hợp X, Y có phân phối chuẩn và biết $D(X), D(Y)$:** Xét thống kê

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{D(X)}{n_1} + \frac{D(Y)}{n_2}}}.$$

Giả sử H_0 là đúng. Người ta chứng minh được $Z \sim N(0,1)$.

● **Trường hợp không biết $D(X), D(Y)$ và n_1, n_2 đủ lớn $(n_1, n_2 > 30)$:** Xét thống kê

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}.$$

Giả sử H_0 là đúng. Người ta chứng minh được $Z \sim N(0,1)$.

Nhận xét (Qui tắc thực hành): Với các mẫu cụ thể kích thước n_1 của X và kích thước n_2 của Y, tính giá trị cụ thể Z.

Giả thuyết	$H_0 : E(X) = E(Y)$ $H_1 : E(X) \neq E(Y)$	$H_0 : E(X) = E(Y)$ $H_1 : E(X) > E(Y)$	$H_0 : E(X) = E(Y)$ $H_1 : E(X) < E(Y)$
Miền bác bỏ W_α	$\lvert Z \rvert > z_{\alpha/2}$	$Z > Z_\alpha$	$Z < -z_\alpha$

Ví dụ 1. Trọng lượng một loại sản phẩm do hai nhà máy sản xuất ra là đại lượng ngẫu nhiên X, Y có phân phối chuẩn và có cùng độ lệch tiêu chuẩn là $\sigma = 1(kg)$. Cân thử 25 sản phẩm của nhà máy thứ nhất ta có $\overline{x} = 50(kg)$ và cân thử 20 sản phẩm của nhà máy thứ hai ta có $\overline{y} = 50,58(kg)$. Với mức ý nghĩa $\alpha = 0,05$, hãy kết luận xem trọng lượng trung bình của sản phẩm do hai nhà máy sản xuất ra có như nhau không?

Giải. Kiểm định giả thuyết

$$H_0 : E(X) = E(Y);\ H_1 : E(X) \neq E(Y).$$

Ở đây X và Y có phân phối chuẩn, biết $D(X) = D(Y) = 1$ và cỡ mẫu $n_1 = 25; n_2 = 20$. Tính

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{D(X)}{n_1} + \frac{D(Y)}{n_2}}} = \frac{50 - 50,58}{\sqrt{\frac{1}{25} + \frac{1}{20}}} = -1,9333.$$

Tra bảng phân vị chuẩn tắc, ta có $z_{\alpha/2} = z_{0,025} = 1,96$ và miền bác bỏ $W_a : \lvert Z \rvert > z_{\alpha/2}$. Kiểm tra thấy $z \notin W_\alpha$ nên tạm chấp nhận H_0. Vậy với mức ý nghĩa $\alpha = 0,05$, tạm xem trọng lượng trung bình sản phẩm hai nhà máy như nhau.

Ví dụ 2. Sử dụng giả thiết ở Ví dụ 1. Với mức ý nghĩa 0,05, hãy kết luận trọng lượng trung bình của nhà máy thứ nhất có bé hơn của nhà máy thứ hai không?

Giải. Kiểm định giả thuyết

$$H_0 : E(X) = E(Y);\ H_1 : E(X) < E(Y).$$

Tương tự như trên đây, ta tính được $z = -1,9333$. Tra bảng phân vị chuẩn tắc, ta có

$$z_\alpha = z_{0,05} = 1,6449 \text{ và } W_\alpha : Z < -z_\alpha \text{ (miền bác bỏ).}$$

Kiểm tra thấy $z \in W_\alpha$ nên ta bác bỏ H_0. Vậy với mức ý nghĩa $\alpha = 0,05$, trọng lượng trung bình sản phẩm nhà máy thứ nhất nhỏ hơn của nhà máy thứ hai.

● **Trường hợp X, Y có phân phối chuẩn và $D(X) = D(Y)$ nhưng chưa biết giá trị cụ thể của chúng:** Người ta chứng minh được: Nếu H_0 đúng, tức là $E(X) = E(Y)$ thì

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \text{ có phân phối Student bậc } n_1 + n_2 - 2$$

với

$$S_*^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}.$$

Nhận xét (Qui tắc thực hành khi n_1 hay n_2 không đủ lớn: Với các mẫu cụ thể kích thước n_1 của X và kích thước n_2 của Y, tính s_*^2 và rồi tính giá trị cụ thể t.

Giả thuyết	$H_0 : E(X) = E(Y)$ $H_1 : E(X) \neq E(Y)$	$H_0 : E(X) = E(Y)$ $H_1 : E(X) > E(Y)$	$H_0 : E(X) = E(Y)$ $H_1 : E(X) < E(Y)$
Miền bác bỏ W_α	$\lvert T \rvert > t_{\alpha/2}$. Tra bảng phân vị Student bậc $(n_1 + n_2 - 2)$	$T > t_\alpha$	$T < -t_\alpha$

Ví dụ 3. Trọng lượng một loại sản phẩm do hai nhà máy sản xuất ra là đại lượng ngẫu nhiên X, Y có phân phối chuẩn và biết là có cùng độ lệch chuẩn. Cân thử 25 sản phẩm của nhà máy thứ nhất ta có $\overline{x} = 150(kg)$ và $s_X^2 = 1(kg^2)$. Cân thử 30 sản phẩm của nhà máy thứ hai ta có $\overline{y} = 150,6(kg)$ và $s_Y^2 = 1,01(kg^2)$. Với mức ý nghĩa $\alpha = 0,02$, hãy kết luận xem trọng lượng trung bình của sản phẩm nhà máy thứ nhất sản xuất ra có nhỏ hơn của nhà máy thứ hai không?

Giải. Đây là trường hợp $D(X) = D(Y)$ nhưng chưa biết giá trị cụ thể và kích thước mẫu $n_1 = 25, n_2 = 30$ chưa đủ lớn. Kiểm định giả thuyết

$$H_0 : E(X) = E(Y);\ H_1 : E(X) < E(Y).$$

Tính

$$s^* = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2} = \frac{24 + 29 \times 1,01}{53} = 1,0055,$$

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{(s_*^2 / n_1) + (s_*^2 / n_2)}} = \frac{50 - 50,6}{\sqrt{s_*^2 \left(\frac{1}{25} + \frac{1}{30}\right)}} = 2,1055.$$

Tra bảng phân vị Student bậc $n_1 + n_2 - 2 = 53$, ta có

$$t_\alpha = t_{0,02} = 2,1055 \text{ và } W_\alpha : T < -t_\alpha \text{ (miền bác bỏ).}$$

Kiểm tra ta thấy $t \in W_\alpha$ nên bác bỏ H_0. Vậy với mức ý nghĩa 2%, có thể xem trọng lượng trung bình của sản phẩm nhà máy thứ nhất nhỏ hơn của nhà máy thứ hai.

4.2.4. Bài toán so sánh hai tỉ lệ tổng thể

Giả sử p_1 và p_2 lần lượt là tỉ lệ các phần tử có tính chất A trong tổng thể thứ nhất và thứ hai. Cần kiểm định giả thuyết

$$H_0 : p_1 = p_2 \text{ với mức ý nghĩa } \alpha.$$

Với hai mẫu ngẫu nhiên kích thước n_1 và n_2 đủ lớn lần lượt trong tổng thể thứ nhất và thứ hai; F_1 và F_2 lần lượt là các tỉ lệ mẫu và $P^* = \frac{n_1F_1 + n_2F_2}{n_1 + n_2}$ là tỉ lệ mẫu của mẫu gộp. Giả sử H_0 là đúng. Người ta chứng minh được

$$Z = \frac{F_1 - F_2}{\sqrt{P^*(1-P^*).\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

xấp xỉ phân phối chuẩn tắc.

Nhận xét (Qui tắc thực hành): Với hai mẫu cụ thể kích thước n_1 và n_2 đủ lớn của tổng thể 1 và tổng thể 2, tính p^* (p^* là tỉ lệ có tính chất A của mẫu cụ thể gồm hai mẫu gộp lại), rồi tính giá trị cụ thể của z.

Giả thuyết	$H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$	$H_0 : p_1 = p_2$ $H_1 : p_1 > p_2$	$H_0 : p_1 = p_2$ $H_1 : p_1 < p_2$
Miền bác bỏ W_α	$\lvert Z \rvert > z_{\alpha/2}$	$Z > z_\alpha$	$Z < -z_\alpha$

Ví dụ. Kiểm tra các sản phẩm chọn ngẫu nhiên do hai nhà máy sản xuất. Ta có bảng

Nhµ m¸y	Sè SP ®-îc kiÓm tra	Sè phÕ phÈm
A	1.000	20
B	900	30

Với mức ý nghĩa $\alpha = 0,05$, có thể coi tỉ lệ phế phẩm hai nhà máy là như nhau không?

Giải. Gọi p_1 và p_2 là các tỉ lệ phế phẩm của nhà máy thứ nhất và thứ hai. Kiểm định giả thuyết

$$H_0 : p_1 = p_2;\ H_1 : p_1 \neq p_2.$$

Tỉ lệ phế phẩm của nhà máy thứ nhất và thứ hai, tương ứng là

$$f_1 = 0,02;\ f_2 = 0,0333.$$

Ta có

$$p^* = \frac{20 + 30}{1000 + 900} = \frac{1}{38} \text{ và } 1 - p^* = \frac{37}{38}.$$

Tính

$$z = \frac{f_1 - f_2}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0,02-0,033}{\sqrt{\frac{1}{38}\cdot\frac{37}{38}\left(\frac{1}{1000}+\frac{1}{900}\right)}} = -1,81.$$

Tra bảng phân vị chuẩn tắc, ta có

$$z_{\alpha/2} = z_{0,025} = 1,96 \text{ và } W_\alpha : |Z| > z_{\alpha/2} \text{ miền bác bỏ).}$$

Kiểm tra thấy $z \notin W_\alpha$ nên ta tạm chấp nhận H_0. Vậy với mức ý nghĩa 5%, tạm có thể coi tỉ lệ phế phẩm của hai nhà máy như nhau.

4.2.5. Bài toán kiểm định giả thuyết phương sai tổng thể

Giả sử đại lượng ngẫu nhiên X phân phối theo qui luật chuẩn và chưa biết phương sai σ^2. Cần kiểm định giả thuyết

$$H_0 : \sigma^2 = \sigma_0^2 \text{ với mức ý nghĩa } \alpha.$$

Để đơn giản, chỉ xét trường hợp không biết trung bình tổng thể μ. Từ X lập mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$. Lập thống kê kiểm định

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}.$$

Người ta chứng minh được χ^2 có phân phối χ^2 với bậc tự do $n-1$.

Nhận xét (Qui tắc thực hành): Với mẫu cụ thể $(x_1, x_2, ..., x_n)$ kích thước n, tính giá trị cụ thể

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$

Giả thuyết	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$
Miền bác bỏ W_α	$(-\infty, \chi^2_{1-\alpha/2}) \cup (\chi^2_{\alpha/2}, +\infty)$	$\chi^2 > \chi_\alpha$. Tra bảng χ^2 với bậc tự do $n-1$	$\chi < \chi_{1-\alpha}$

Ví dụ 8. Khi máy hoạt động bình thường, trọng lượng sản phẩm $X(kg)$ phân phối theo qui luật chuẩn với $D(X) = 0,12(kg^2)$. Nghi ngờ máy hoạt động không ổn định như bình thường, người ta cân thử 25 sản phẩm và tính được $\sigma^2 = 0,13(kg^2)$. Với mức ý nghĩa $\alpha = 0,05$, hãy kết luận điều nghi ngờ trên có đúng không?

Giải. Ta kiểm định giả thuyết

$$H_0 : \sigma^2 = 12;\ H_1 : \sigma^2 \neq 12.$$

Tính

$$\chi^2 = \frac{(n-1).s^2}{\sigma_o^2} = \frac{24 \times 0,13}{0,12} = 26 \text{ và } \alpha/2 = 0,025 \Rightarrow 1-\alpha/2 = 0,975.$$

Tra bảng χ^2 bậc $n-1=24$, ta có

$\chi^2_{1-\alpha/2}=12,401$, $\chi^2_{\alpha/2}=39,3641$ và $W_\alpha=(-\infty;12,401)\cup(39,3641;+\infty)$ (miền bác bỏ).

Kiểm tra thấy $\chi^2=26\notin W_\alpha$ nên tạm chấp nhận H_0. Vậy với mức ý nghĩa $\alpha=5\%$, tạm chấp nhận coi máy móc hoạt động bình thường.

4.2.6. Bài toán kiểm định giả thuyết về tính độc lập

Giả sử ta quan sát đồng thời hai dấu hiệu A và B trên một tổng thể lớn. Dấu hiệu A có k dấu hiệu thành phần $A_1,A_2,\ldots,A_k$ và dấu hiệu B có h dấu hiệu thành phần $B_1,B_2,\ldots,B_h$. Ta cần kiểm định giả thuyết

H_0: Hai dấu hiệu A và B là độc lập với mức ý nghĩa $\alpha>0$ đủ bé.

Nhận xét (Quy tắc thực hành): Chọn mẫu cụ thể kích thước n đủ lớn và lập bảng các tần số thực tế. Tính các tần số lý thuyết

$$n.\hat{P}_{ij}=\frac{n_i m_j}{n}$$

và ghi kề dưới tần số thực tế n_{ij} tại ô (i,j). Đồng thời kiểm tra điều kiện

$$n.\hat{P}_{ij}>5 \text{ tại mỗi ô } (i,j).$$

Ta được bảng tần số hoàn chỉnh cụ thể dạng sau

B / A	B_1	B_2	...	B_h	Σ
A_1	n_{11} $(n\hat{P}_{11})$	n_{12} $(n\hat{P}_{12})$	...	n_{1h} $(n\hat{P}_{1h})$	n_1
A_2	n_{21} $(n\hat{P}_{21})$	n_{22} $(n\hat{P}_{22})$	...	n_{2h} $(n\hat{P}_{2h})$	n_2
...	...	...	...	...	...
A_k	n_{k1} $(n\hat{P}_{k1})$	n_{k2} $(n\hat{P}_{k2})$	...	n_{kh} $(n\hat{P}_{kh})$	n_k
Σ	m_1	m_2	...	m_h	n

Tính giá trị tiêu chuẩn kiểm định

$$\chi^2=\sum_{i,j}\frac{n_{ij}^2}{n.\hat{P}_{ij}}-n.$$

Miền bác bỏ $W_\alpha:\chi^2>\chi^2_\alpha$. Tra χ^2_α ở bảng phân vị khi-bình-phương bậc $(h-1)(k-1)$. Kiểm tra:

- Nếu $\chi^2\in W_\alpha$ thì bác bỏ H_0.
- Nếu $\chi^2\notin W_\alpha$ thì tạm chấp nhận H_0.

Chú ý rằng, hai dấu hiệu A và B độc lập khi và chỉ khi tỉ lệ các thành phần dấu hiệu A trong các nhóm mang dấu hiệu $B_1,B_2,\ldots,B_h$ là như nhau.

Ví dụ 9. Cần nghiên cứu tác dụng của các loại phân bón 1, 2, 3 đối với việc ra hoa và không ra hoa của một loài hoa. Thí nghiệm bón ba loại phân cho một số cây loài hoa đó, ta có tình hình ra hoa của các cây cho ở bảng

Phân bón Tình trạng	Phân bón 1	Phân bón 2	Phân bón 3
Cã ra hoa	40	68	53
Kh«ng ra hoa	15	12	12

Với mức ý nghĩa 5%, hãy kiểm định việc bón các loại phân khác nhau có ảnh hưởng tới việc ra hoa của loài hoa trên không? (Tỉ lệ ra hoa của loài hoa này khi bón các loại phân 1, 2, 3 có như nhau không?)

Giải. Kiểm định giả thuyết

$$H_0 : A \text{ và B độc lập.}$$

Lập bảng tần số

B A	B1 Phân bón 1	B2 Phân bón 2	B3 Phân bón 3	$\sum$ Tæng céng
A1 Cã ra hoa	40 (44,275)	68 (64,4)	53 (52,325)	161
A2 Kh«ng ra hoa	15 (10,725)	12 (15,6)	12 (12,675)	39
$\sum$	55	80	65	$n = 200$

Tiêu chuẩn kiểm định

$$\chi^2 = \sum_{i,j} \frac{n_{ij}^2}{\widehat{n.P_{ij}}} - n$$

$$= \frac{40^2}{44,275} + \frac{68^2}{64,4} + \frac{53^2}{52,325} + \frac{15^2}{10,725} + \frac{12^2}{15,6} + \frac{12^2}{12,675} - 200 = 3,1935 .$$

Miền bác bỏ

$$\chi^2 > W_\alpha : \chi^2 > \chi^2_{0,05}(2) .$$

Kiểm tra thấy $\chi^2 \notin W_\alpha$ nên tạm chấp nhận H_0. Vậy với mức ý nghĩa $\alpha = 5\%$, ta tạm coi dấu hiệu ra hoa và các loại phân là độc lập. (Tạm coi tỉ lệ ra hoa của loài hoa này khi bón các loại phân 1, 2, 3 là như nhau).

4.3. Bài tập chương 4

1. Lấy ngẫu nhiên 25 sản phẩm do một công ty sản xuất ra. Ta tính được trọng lượng trung bình của chúng là $995,8(g)$ và phương sai mẫu là $0,144(g^2)$. Giả thiết trọng lượng các sản phẩm là đại lượng ngẫu nhiên $X(g)$ có phân phối chuẩn.

a) Ước lượng trọng lượng trung bình các sản phẩm do công ty sản xuất với độ tin cậy 95%.

b) Có ý kiến trọng lượng trung bình sản phẩm công ty là khác $996(g)$. Hãy kết luận ý kiến trên với mức ý nghĩa $\alpha = 5\%$.

2. Để định mức thời gian gia công một chi tiết máy trong công ty, người ta theo dõi ngẫu nhiên thời gian gia công 225 chi tiết và thu được bảng số liệu sau đây

Thêi gian (phót)	14	16	18	20	22
Sè chi tiÕt	12	56	91	54	12

a) Với độ tin cậy 99%, hãy ước lượng thời gian gia công trung bình loại chi tiết trên.

b) Có ý kiến thời gian gia công trung bình loại chi tiết trên là dưới 18,5 (phút). Hãy kết luận ý kiến với mức ý nghĩa 1%.

3. Năng suất ngô của một vùng A được báo cáo lên qua 100 điểm thu hoạch được chọn ngẫu nhiên là

N¨ng suÊt (t¹/ha)	7	9	11	13	15
Sè ®iÓm thu ho¹ch	2	27	40	30	1

Cho biết năng suất ngô tuân theo quy luật chuẩn.

a) Với độ tin cậy là 95%, hãy ước lượng năng suất ngô trung bình của vùng này.

b) Có ý kiến năng suất ngô trong vùng là hơn 10,4 (tạ/ha). Với mức ý nghĩa 2%, hãy kết luận.

4. Chọn ngẫu nhiên 100 công nhân của một xí nghiệp thì thấy lương tháng trung bình là 5,8 (triệu đồng). Giả sử lương X (triệu đồng) của mỗi công nhân tuân theo quy luật chuẩn, với độ lệch chuẩn $\sigma = 1,2$ (triệu đồng).

a) Độ tin cậy 98%, hãy ước lượng mức lương trung bình của công nhân trong toàn xí nghiệp.

b) Có ý kiến mức lương trung bình của công nhân là trên 5,5 (triệu đồng). Với mức ý nghĩa 1%, hãy kết luận ý kiến đó.

5. Trọng lượng trung bình qui định cho các bao hàng do một máy đóng bao sản xuất là 50 (kg). Sau một thời gian hoạt động, người ta nghi ngờ máy có trục trặc. Cho trọng lượng X (kg) mỗi bao hàng là đại lượng ngẫu nhiên phân phối chuẩn.

a) Cân thử 100 bao hàng và tính được $\overline{x} = 49,88(kg)$ và $s = 0,4(kg)$. Với mức ý nghĩa 1%, hãy kết luận trọng lượng trung bình các bao hàng do máy đóng có sai khác qui định không?

b) Cân thử 400 bao hàng và tính được $\overline{x} = 49,93(kg)$ và $s = 0,5(kg)$. Với mức ý nghĩa 1%, hãy kết luận trọng lượng trung bình các bao hàng do máy đóng có thấp hơn qui định không?

6. Tỉ lệ phế phẩm của một dây chuyền sản xuất là 5%. Sau khi tiến hành một cải tiến kỹ thuật người ta kiểm tra ngẫu nhiên 2000 sản phẩm thì thấy có 95 phế phẩm. Với mức ý nghĩa 1%, hãy kết luận việc cải tiến kỹ thuật có làm giảm tỉ lệ phế phẩm không?

7. Nghiên cứu trọng lượng các trẻ sơ sinh của hai nhóm mẹ nghiện thuốc lá (Nhóm 1) và nhóm mẹ không hút thuốc lá (Nhóm 2), ta có kết quả (kg):

Nhãm 1	Nhãm 2	Nhãm 1	Nhãm 2
3,99	3,18	3,61	2,76
3,79	2,84	3,83	3,60
3,60	2,90	3,31	3,75
3,73	3,27	4,13	3,59
3,21	3,85	3,26	3,63
3,60	3,52	3,54	2,38
4,08	3,23		

Với mức ý nghĩa 5% có thể nói rằng trẻ sơ sinh của nhóm mẹ nghiện thuốc lá nhẹ cân hơn trẻ sơ sinh của nhóm mẹ không hút thuốc lá được không? Cho biết trọng lượng trẻ sơ sinh của hai nhóm trên tuân theo quy luật chuẩn và có cùng phương sai.

8. Trọng lượng một loại sản phẩm do hai nhà máy sản xuất ra là đại lượng ngẫu nhiên X,Y (kg) có phân phối chuẩn và có cùng độ lệch tiêu chuẩn là $\sigma = 0,5(kg)$. Cân thử 29 sản phẩm của nhà máy thứ nhất ta có $\overline{x} = 50(kg)$ và cân thử 30 sản phẩm của nhà máy thứ hai ta có $\overline{y} = 50,58(kg)$. Với mức ý nghĩa 5%, hãy kết luận xem trọng lượng trung bình của sản phẩm do hai nhà máy sản xuất ra có như nhau không?

9. Đường kính một loại chi tiết do hai nhà máy sản xuất ra là hai đại lượng ngẫu nhiên X,Y (mm). Kiểm tra ngẫu nhiên 800 chi tiết do nhà máy thứ nhất sản xuất, ta được $\overline{x} = 100,1(mm)$ và $s_X^2 = 0,001(mm^2)$. Kiểm tra ngẫu nhiên 750 chi tiết của nhà máy thứ hai sản xuất, ta được $\overline{y} = 100,05(mm)$ và $s_Y^2 = 0,0012(mm^2)$. Người ta nói đường kính trung bình của chi tiết nhà máy thứ nhất lớn hơn của chi tiết nhà máy thứ hai. Với mức ý nghĩa 2%, bạn hãy cho ý kiến của mình.

10. Trọng lượng một loại sản phẩm do hai nhà máy sản xuất ra là đại lượng ngẫu nhiên X,Y có phân phối chuẩn và có cùng độ lệch chuẩn. Cân thử 35 sản phẩm của nhà máy thứ nhất ta có $\overline{x} = 200(kg)$ và $s_X^2 = 1(kg^2)$. Cân thử 36 sản phẩm của nhà máy thứ hai ta có $\overline{y} = 200,6(kg)$ và $s_Y^2 = 1,01(kg^2)$. Với mức ý nghĩa 2%, hãy kết luận trọng lượng trung bình sản phẩm nhà máy thứ nhất sản xuất có nhỏ hơn của nhà máy thứ hai không?

11. Để tìm hiểu hiệu quả của việc giảng dạy một vấn đề nào đó theo phương pháp (PP) cũ và mới, người ta làm một bảng kiểm tra trên 15 sinh viên và có kết quả tính bằng điểm số với điểm tối đa là 100 cho bởi bảng sau đây

Sinh viªn	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
§iÓm (PP cò)	54	79	91	75	68	43	33	85	22	56	73	63	29	75	87
§iÓm (PP mí i)	66	85	83	88	93	40	38	91	44	82	59	81	64	83	81

Với mức ý nghĩa 5% có thể xem việc học theo phương pháp mới có hiệu quả hơn được không, theo nghĩa điểm trung bình của sinh viên học theo phương pháp mới cao hơn điểm trung bình của sinh viên khi học theo phương pháp cũ? Cho biết điểm số của sinh viên tuân theo quy luật chuẩn.

12. Kiểm tra các sản phẩm chọn ngẫu nhiên do hai nhà máy sản xuất. Ta có bảng số liệu

Nhµ m¸y	Sè SP ®-î c kiÓm tra	Sè phÕ phÈm
A	1.000	20
B	1.000	25

Với mức ý nghĩa 5%, có thể kết luận tỉ lệ phế phẩm nhà máy B là lớn hơn không?

13. Tỉ lệ phế phẩm trước kia của một dây chuyền sản xuất là 5%. Sau khi tiến hành một thay đổi kỹ thuật, kiểm tra ngẫu nhiên 4.000 sản phẩm thì thấy có 164 phế phẩm.

a) Với $\alpha = 1\%$, hãy kết luận việc thay đổi kỹ thuật có giảm tỉ lệ phế phẩm không?

b) Ước lượng tỉ lệ phế phẩm của dây chuyền sau thay đổi kỹ thuật với độ tin cậy 99%.

14. Một đảng chính trị dự đoán rằng trong cuộc bầu cử tổng thống sắp tới, ứng viên đảng mình sẽ giành được 45% số phiếu bầu. Chọn ngẫu nhiên 800 cử tri để thăm dò ý kiến cho thấy 320 người nói rằng họ sẽ bỏ phiếu cho ứng viên của đảng đó.

a) Với mức ý nghĩa 1%, nhận định như thế nào về dự đoán của đảng đó?

b) Với mức ý nghĩa 1%, tỉ lệ cử tri bỏ phiếu cho ứng viên đó có ít hơn 45% không?

c) Tìm khoảng tin cậy 99% của tỉ lệ cử tri bỏ phiếu cho ứng viên đó.

15. Tỉ lệ học sinh tốt nghiệp phổ thông năm ngoái của tỉnh A là 88%. Trong kỳ thi năm nay trong 100 em được chọn ngẫu nhiên có 82 em thi đỗ. Với mức ý nghĩa 5%, có thể kết luận rằng tỉ lệ hoc sinh thi đỗ năm nay thấp hơn năm ngoái hay không?

16. Điều tra năng suất lúa vụ hè thu năm 2017 trên một số hecta lúa chọn ngẫu nhiên ở một vùng, người ta thu được bảng số liệu sau

N¨ng suÊt (tÊn/ha)	3,5 – 4	4 – 4,5	4,5 – 5	5 – 5,5	5,5 – 6	6 – 6,5	6,5 – 7	7 – 7,5
Di Ön tÝch (ha)	2	10	10	16	28	32	26	12

a) Những thửa ruộng có năng suất trên 5 (tấn/ha) là những thửa ruộng tốt. Ước lượng tỉ lệ thửa ruộng tốt của các hecta lúa trong vùng với độ tin cậy 95%.

b) Ước lượng năng suất vụ hè thu 2017 các thửa ruộng tốt trong vùng với độ tin cậy 95%.

c) Điều tra vụ hè thu năm 2016 thì thấy năng suất lúa tốt ở vùng này là 6 (tấn/ha). Vụ hè thu năm 2017 người ta áp dụng một biện pháp kỹ thuật mới. Biện pháp kỹ thuật mới này có làm tăng năng suất đối với các diện tích lúa tốt, với mức ý nghĩa 1%?

17. Kiểm tra chất lượng của hai lô sản phẩm, người ta thấy trong lô thứ nhất có 50 phế phẩm trên tổng số 500 sản phẩm kiểm tra và lô thứ hai có 60 phế phẩm trên tổng số 400 sản phẩm kiểm tra. Với mức ý nghĩa 5%, có thể xem lô hàng thứ nhất chất lượng tốt hơn lô thứ hai không?

18. Nghiên cứu tình trạng hôn nhân trước ngày cưới của 800 cặp vợ chồng được chọn ngẫu nhiên ở nước ta được số liệu

Vî Chång	Ch- a kÕt h«n lÇn nµo	Ly h«n	Gãa
Ch- a kÕt h«n lÇn nµo	150	124	66
Ly h«n	138	108	54
Gãa	52	48	60

Với mức ý nghĩa 1%, có thể coi tình trạng hôn nhân trước ngày cưới của vợ chồng là độc lập không?

19 Nghiên cứu về màu tóc và giới tính của 500 người được chọn ngẫu nhiên ở châu Âu, ta có số liệu sau

Giíi tÝnh Mµu tãc	Nam	N÷
§en	70	30
Hung	75	40
N©u	80	55
Vµng	95	55

Với mức ý nghĩa 5%, có thể coi giữa màu tóc và giới tính có độc lập nhau không?

20. Phỏng vấn ngẫu nhiên 200 người thuộc các vùng địa lý ở nước ta về tiêu dùng một loại sản phẩm nào đó ta thu được kết quả sau

Vï ng Þa lý Tiªu dï ng	Thµnh thÞ	N«ng th«n	MiÒn nói
Cã tiªu dï ng	26	50	24
Kh«ng tiªu dï ng	47	45	8

Với mức ý nghĩa 1%, có thể coi yếu tố địa lý và việc tiêu dùng loại sản phẩm nói trên độc lập nhau? (Với mức ý nghĩa 1%, có thể nói tỉ lệ tiêu dùng sản phẩm đó ở thành thị, nông thôn và miền núi là như nhau không?)

21. Một công ty xuất khẩu gạo nói gạo của họ ở các kho 1, 2, 3 là cùng chất lượng hạt tức là chất lượng hạt và các kho là độc lập với nhau. Lấy mẫu cụ thể các hạt gạo ở các kho, ta có số liệu

Kho ChÊt l-î ng h¹t	Kho 1	Kho 2	Kho 3
Cßn nguyªn h¹t	600	460	500
Cßn h¬n 2/3 h¹t	170	100	70
Cßn d-í i 2/3 h¹t	30	40	30

Với mức ý nghĩa 1%, bạn hãy cho ý kiến?

22 Một công ty có bốn kho chứa các sản phẩm cùng loại. Giám đốc công ty đó nói chất lượng sản phẩm và kho hàng là độc lập (tỉ lệ thành phần sản phẩm loại 1, 2, 3 trong các kho hàng là như nhau). Kiểm tra ngẫu nhiên 2.000 sản phẩm ở các kho 1, 2, 3, 4 và có số liệu sau

Kho ChÊt l-î ng	Kho 1	Kho 2	Kho 3	Kho 4
Lo¹i 1	110	120	100	115
Lo¹i 2	210	210	220	225
Lo¹i 3	160	190	180	160

Với mức ý nghĩa 5%, hãy kết luận về ý kiến trên.

Chương 5.
Tương quan và hồi quy

5.1. Đại lượng ngẫu nhiên hai chiều

5.1.1. Khái niệm

Cho X,Y là hai đại lượng ngẫu nhiên. Cặp đại lượng ngẫu nhiên (X,Y) gọi là *đại lượng ngẫu nhiên 2 chiều*. Nếu X,Y là các đại lượng ngẫu nhiên rời rạc thì gọi đại lượng ngẫu nhiên 2 chiều (X,Y) là rời rạc. Nếu X,Y là các đại lượng ngẫu nhiên liên tục thì gọi đại lượng ngẫu nhiên 2 chiều (X,Y) là liên tục.

Ví dụ. Khi khảo sát các siêu thị ở một thành phố, ta quan tâm cùng lúc doanh số bán ra X (triệu đồng) và lượng vốn Y (triệu đồng). Phép thử: chọn ngẫu nhiên một siêu thị trong thành phố và kiểm tra giá trị X và Y. Ta có X,Y là hai đại lượng ngẫu nhiên trong cùng phép thử và (X,Y) là một đại lượng ngẫu nhiên 2 chiều.

Tương tự người ta còn định nghĩa đại lượng ngẫu nhiên 3 chiều, 4 chiều, …

5.1.2. Bảng phân phối xác suất của đại lượng ngẫu nhiên hai chiều rời rạc

Bảng phân phối xác suất của đại lượng ngẫu nhiên 2 chiều rời rạc (*X*, *Y*) có dạng

X \ Y	y_1	y_2	…	y_n	P_X
x_1	p_{11}	p_{12}	…	p_{1n}	p_1
x_2	p_{21}	p_{22}	…	p_{2n}	p_2
⋮	⋮	⋮		⋮	⋮
x_m	p_{m1}	p_{m2}	…	p_{mn}	p_k
P_Y	q_1	q_2	…	q_n	1

Với

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} = 1,\ \sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j = 1\ (\forall p_{ij},\ p_i,\ q_j \geq 0).$$

Khi đó X có thể nhận các trị khác nhau: $x_1,\ldots,x_m$; Y có thể nhận các trị khác nhau: $y_1,\ldots,y_n$ và

$$P(X = x_i, Y = y_j) = p_{ij}$$
$$P(X = x_i) = p_i = p_{i1} + p_{i2} + \ldots + p_{in}$$
$$P(Y = y_j) = q_j = p_{1j} + p_{2j} + \ldots + p_{mj}$$

Ghi chú. Trong thực hành, bảng phân phối xác suất trên có thể bỏ bớt cột phải P_X và hàng dưới P_Y.

Hai đại lượng ngẫu nhiên X và Y được gọi là độc lập nếu biến cố đại lượng ngẫu nhiên này nhận một giá trị bất kỳ nào đó không ảnh hưởng gì tới phân phối xác suất của đại lượng ngẫu nhiên kia. Cụ thể hơn ta luôn có

$$P(X = x \mid Y = y) = P(X = x), \forall x, y$$

hay

$$P(Y = y \mid X = x) = P(Y = y), \forall x,\ y\,.$$

Hai đại lượng ngẫu nhiên X và Y là độc lập khi và chỉ khi

$$P(X = x, Y = y) = P(X = x).P(Y = y)\ \forall x, y$$

Ví dụ. Cho đại lượng ngẫu nhiên 2 chiều (X,Y) có bảng phân phối

X \ Y	1	2	3
1	0	0,3	0,2
2	0,1	0,2	0,2

Hai đại lượng ngẫu nhiên X,Y có độc lập không?

Giải. Ta có

$$P(X = 1) = 0 + 0,3 + 0,2 = 0,5\ \text{P}(X = 1) = 0 + 0,3 + 0,2 = 0,5.$$

$$P(Y = 1) = 0,1\,.$$

$$P(X = 1, Y = 1) = 0 \neq P(X = 1)P(Y = 1) = 0,5 \times 0,1 = 0,05$$

Do đó hai đại lượng ngẫu nhiên X và Y không độc lập.

5.1.3. Hiệp phương sai

Hiệp phương sai (covariance) của cặp đại lượng ngẫu nhiên X và Y, ký hiệu là $Cov(X,Y)$ được định nghĩa là trị số

$$Cov(X,Y) = E\Big[\big(X - E(X)\big)\big(Y - E(Y)\big)\Big].$$

Đặt $E(X) = \mu$ và $E(Y) = \eta$. Ta có

$$Cov(X, Y) = \text{E}[(X - \mu)(Y - \eta)] = \text{E}[XY - \eta X - \mu Y + \mu\eta]$$
$$= \text{E}(XY) - \eta\,\text{E}(X) - \mu\,\text{E}(Y) + \mu\eta\ = \text{E}(XY) - \mu\eta.$$

Ta suy ra công thức

$$Cov(X,Y) = E(XY) - E(X)E(Y)\,.$$

Nhận xét.

(n_1) Cho X,Y là hai đại lượng ngẫu nhiên độc lập. Ta có

$$E(X) = E(X)E(Y)\,.$$

Suy ra $Cov(X,Y) = 0$. Do đó nếu $Cov(X,Y) \neq 0$ thì X,Y *phụ thuộc nhau* (hay nói X,Y *có tương quan*). Người ta còn gọi $Cov(X,Y)$ là moment tương quan của X,Y. Ký hiệu

$$Cov(X,Y) := \mu_{XY}.$$

(n_2) Từ các định nghĩa suy ra

$$Cov(X,Y) = Cov(Y,X) \text{ và } Cov(X,X) = D(X) = Var(X).$$

Mệnh đề. Cho X và Y là hai đại lượng ngẫu nhiên. Khi đó

$$D(aX \pm bY) = a^2 D(X) + b^2 D(Y) \pm 2ab Cov(X,Y).$$

Chứng minh. Ta có

$$\begin{aligned} \mathrm{D}(aX \pm bY) &= \mathrm{E}[(aX \pm bY)^2] - [\mathrm{E}(aX \pm bY)]^2 \\ &= \mathrm{E}[a^2X^2 + b^2Y^2 \pm 2abXY] - [a\,\mathrm{E}(X) \pm b\,\mathrm{E}(Y)]^2 \\ &= a^2\,\mathrm{E}(X^2) + b^2\,\mathrm{E}(Y^2) \pm 2ab\,\mathrm{E}(XY) - [a^2\,[\mathrm{E}(X)]^2 + b^2\,[\mathrm{E}(Y)]^2 \pm 2ab\,\mathrm{E}(X)\,\mathrm{E}(Y)] \\ &= a^2\,[\mathrm{E}(X^2) - [\mathrm{E}(X)]^2] + b^2\,[\mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 \pm 2ab\,[\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)] \\ &= a^2\,\mathrm{D}(X) + b^2\,\mathrm{D}(Y) \pm 2ab\,\mathrm{Cov}(X, Y). \square \end{aligned}$$

Ví dụ. Một công ty có lãi suất hằng năm của trái phiếu là $T(\%)$ và của cổ phiếu là $S(\%)$. Cho bảng phân phối xác suất của đại lượng ngẫu nhiên (T,S)

T \ S	−10	0	10	20	P_T
6	0	0	0,1	0,1	0,2
8	0	0,1	0,3	0,2	0,6
10	0,1	0,1	0	0	0,2
P_S	0,1	0,2	0,4	0,3	**1**

Nếu muốn đầu tư vào cả trái phiếu và cổ phiếu thì nên đầu tư theo tỷ lệ bao nhiêu để:

a) Lãi suất kỳ vọng (lãi suất trung bình) thu được là lớn nhất.

b) Mức độ rủi ro về lãi suất là nhỏ nhất.

Giải.

a) Từ bảng trên, ta tính được:

$$\mathrm{E}(T) = 6 \times 0,2 + 8 \times 0,6 + 10 \times 0,2 = 8\,(\%).$$

$$\mathrm{D}(T) = \mathrm{E}(T^2) - [\mathrm{E}(T)]^2 = (6^2 \times 0,2 + 8^2 \times 0,6 + 10^2 \times 0,2) - 8^2 = 1,6.$$

Tương tự

$$\mathrm{E}(S) = 9\ (\%).$$

$$\mathrm{D}(S) = 89.$$

$$\mathrm{Cov}(T, S) = \mathrm{E}(TS) - \mathrm{E}(T)\,\mathrm{E}(S) = \ldots = -8.$$

Giả sử đầu tư vào trái phiếu và cổ phiếu theo tỉ lệ p và $(1 - p)$ $(0 \le p \le 1)$. Gọi $X\,(\%)$ là lãi suất thu được khi đầu tư vào cả trái phiếu và cổ phiếu theo các tỉ lệ trên. Ta có

$$X = pT + (1 - p)\,S\,(\%).$$

Lãi suất kỳ vọng

$$\mathrm{E}(X) = p\,\mathrm{E}(T) + (1 - p)\,\mathrm{E}(S) = 8p + 9\,(1 - p) = 9 - p.$$

Ta thấy E(X) đạt giá trị lớn nhất khi $p = 0$. Suy ra đầu tư toàn bộ vào cổ phiếu thì lãi suất kỳ vọng thu được là lớn nhất.

b) Độ rủi ro về lãi suất biểu thị bằng phương sai của X. Ta có

$$D(X) = p^2 D(T) + (1-p)^2 D(Y) + 2p(1-p)\,\mathrm{Cov}(T, S)$$
$$= 106{,}6\,p^2 - 194\,p + 89$$

D(X) đạt giá trị bé nhất khi

$$p = \frac{194}{2.106{,}6} = 0{,}9099.$$

Vậy đầu tư vào trái phiếu với tỷ lệ 90,99% và vào cổ phiếu 8,01% thì độ rủi ro về lãi suất là nhỏ nhất.

5.1.4. Hàm mật độ của đại lượng ngẫu nhiên 2 chiều liên tục

Hàm $f(x,y)$ được gọi là hàm mật độ xác suất (probability density function) của cặp đại lượng ngẫu nhiên liên tục (X,Y) nếu

$$P(X < x, Y < y) = \int_{-\infty}^{x}\left(\int_{-\infty}^{y} f(u,v)dv\right)du,\ \forall x, y.$$

Tính chất.

a) $P\Big[(x_1 < X < x_2),(y_1 < Y < y_2)\Big] = \int_{x_1}^{x_2}\left(\int_{y_1}^{y_2} f(x,y)dy\right)dx$.

b) $f(x,y) = \dfrac{\partial^2 F(x,y)}{\partial x \partial y}$ tại các điểm liên tục của $f(x,y)$.

c) $f(x,y) \geq 0$ tại các điểm liên tục của $f(x,y)$ và $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$.

Ví dụ. Cho hàm mật độ của đại lượng ngẫu nhiên 2 chiều (X, Y) như sau

$$f(x,y) = \begin{cases} (2x+y) & \text{khi} \quad (x,y) \in [2,6]\times[0,5] \\ 0 & \text{khi} \quad (x,y) \notin [2,6]\times[0,5] \end{cases}.$$

a) Xác định C.

b) Tìm xác suất $P\Big[(3 < X < 4),(Y > 3)\Big]$.

Giải. Ta có

a) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\left(x,y\right)dxdy = \int_{2}^{6}\left(\int_{0}^{5} C\left(2x+y\right)dy\right)dx = 210C = 1 \Rightarrow C = \dfrac{1}{210}$.

b) $P\Big[(3 < X < 4),(3 < Y < +\infty)\Big] = \int_{3}^{+\infty}\left(\int_{3}^{4} f(x,y)dx\right)dy$

$$= \frac{1}{210}\int_{3}^{5}\left(\int_{3}^{4}(2x+y)dx\right)dy = \frac{1}{210}\int_{3}^{5}\left(x^2+xy\right)\Big|_{x=3}^{x=4}dy$$

$$= \frac{1}{210}\int_{3}^{5}\left(7+y\right)dy = \frac{1}{210}\left(7y + \frac{y^2}{2}\right)\Big|_{y=3}^{y=5} = \frac{22}{210}.$$

5.1.5. Phân phối có điều kiện của cặp đại lượng ngẫu nhiên rời rạc

Cho đại lượng ngẫu nhiên 2 chiều rời rạc (X,Y) có bảng phân phối xác suất

X \ Y	y_1	y_2	$\cdots$	y_n	P_X
x_1	p_{11}	p_{12}	$\cdots$	p_{1n}	p_1
x_2	p_{21}	p_{22}	$\cdots$	p_{2n}	p_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	p_{m1}	p_{m2}	$\cdots$	p_{mn}	p_m
P_Y	q_1	q_2	....	q_n	1

Cố định $X = x_k$ hoặc $Y = y_k$, ta tính các xác suất có điều kiện:

$$P(X = x_i \mid Y = y_k) = \frac{P(X = x_i, Y = y_k)}{P(Y = y_k)} = \frac{p_{ik}}{q_k},\ i = 1,\ldots,m$$
$$P(Y = y_j \mid X = x_k) = \frac{P(X = x_k, Y = y_j)}{P(X = x_k)} = \frac{p_{kj}}{p_k},\ j = 1,\ldots,n$$

.

Suy ra bảng phân phối xác suất có điều kiện của X với điều kiện $Y = y_k$:

X	x_1	x_2	...	x_m
$P(X \mid Y = y_k)$	$\frac{p_{1k}}{q_k}$	$\frac{p_{2k}}{q_k}$	...	$\frac{p_{mk}}{q_k}$

Kỳ vọng có điều kiện của đại lượng ngẫu nhiên rời rạc X với điều kiện $Y = y_k$ là:

$$= E(X \mid Y = y_k) = \frac{x_1 p_{1k} + \ \ldots\ + x_m p_{mk}}{q_k}.$$

Tương tự kỳ vọng có điều kiện của đại lượng ngẫu nhiên rời rạc Y với điều kiện $X = x_k$ là:

$$E(Y \mid X = x_k) = \frac{y_1 p_{k1} + \ldots + y_n p_{kn}}{p_k}.$$

Ví dụ. Cho bảng phân phối xác suất của đại lượng ngẫu nhiên 2 chiều (X,Y), trong đó X là doanh thu và Y là chi phí quảng cáo (triệu đồng/tháng) của các công ty tư nhân kinh doanh cùng mặt hàng như sau

Y \ X	100	150	200	P_Y
0	0,1	0,05	0,05	0,2
1	0,05	0,2	0,15	0,4
2	0	0,1	0,3	0,4
P_X	0,15	0,35	0,5	1,0

a) Tìm doanh thu trung bình của các công ty không quảng cáo?

b) Tìm doanh thu trung bình của các công ty có mức quảng cáo 2 (triệu đồng/tháng)?

Giải. Ta có bảng phân phối của đại lượng ngẫu nhiên X với điều kiện $Y = 0$:

X	100	150	200
$P(X \mid Y = 0)$	0,5	0,25	0,25

Suy ra

$$E(X \mid Y = 0) = 137,5 .$$

Vậy doanh thu trung bình của các công ty không quảng cáo $(Y = 0)$ là 137,5 (triệu/tháng).

b) Tương tự ta có bảng phân phối của X với điều kiện $Y = 2$:

X	100	150	200
$P(X \mid Y = 2)$	0	0,25	0,75

Suy ra

$$E(X \mid Y = 2) = 187,5 .$$

Vậy doanh thu trung bình các công ty có mức quảng cáo $Y = 2$ (triệu/tháng) là 187,5 (triệu/tháng).

5.2. Hàm hồi qui và hệ số tương quan

5.2.1. Mẫu ngẫu nhiên 2 chiều

Giả sử trên cùng một tổng thể phải nghiên cứu đồng thời hai dấu hiệu mà trong đó lần lượt dấu hiệu thứ nhất, thứ hai được cụ thể hóa thành đại lượng ngẫu nhiên X, đại lượng ngẫu nhiên Y. Khi đó việc nghiên cứu hai dấu hiệu trên có thể xem như nghiên cứu đại lượng ngẫu nhiên hai chiều (X,Y) trên tổng thể.

Từ tổng thể ta dự định sẽ lấy ngẫu nhiên phần tử thứ 1, …, thứ n. Giá trị của (X,Y) sẽ đo được ở phần tử thứ i (chưa biết cụ thể) là đại lượng ngẫu nhiên 2 chiều (X_i,Y_i) có cùng qui luật phân phối với (X,Y), $i = 1,...,n$. Ta gọi $((X_1,Y_1),...,(X_n,Y_n))$ là một mẫu ngẫu nhiên 2 chiều của đại lượng ngẫu nhiên (X,Y).

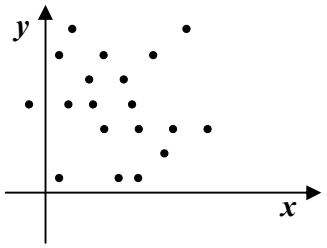
Lấy n phần tử cụ thể của tổng thể, đo giá trị (X,Y) cụ thể trên các phần tử thứ 1, thứ 2, …, thứ n. Giả sử ta quan sát được các số liệu $(x_1,y_1),...,(x_n,y_n)$. Khi đó ta được mẫu cụ thể $((x_1,y_1),...,(x_n,y_n))$.

5.2.2. Đám mây quan sát

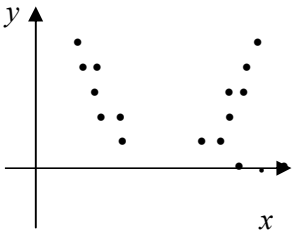
Xét cặp đại lượng ngẫu nhiên (X,Y) trên một tổng thể nào đó. Giả sử ta quan sát được các giá trị của (X,Y) là $(x_1,y_1),(x_2,y_2),...,(x_k,y_k)$. Biểu diễn các điểm (x_i,y_i) lên mặt phẳng Oxy, ta được một tập hợp các điểm gọi là đám mây quan sát. Có thể xảy ra các trường hợp:

a) Các điểm (x_i,y_i) nằm rải rác trong mặt phẳng không theo một qui tắc nào. Khi đó người ta coi giữa X và Y không có tương quan.

b) Các điểm (x_i,y_i) nằm sát xung quanh một đường nào đó. Khi đó người ta sẽ giả định sự tương quan giữa X và Y được biểu diễn bởi một hàm số có đồ thị là đường nói trên.

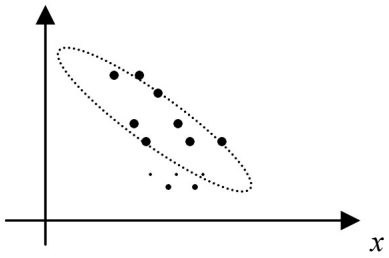


a) Không tương quan.



b) Tương quan bậc hai (coi $Y = aX^2 + bX + c$).

Một trường hợp hay gặp là các điểm (x_i, y_i) nằm trong một elip nào đó. Khi đó người ta nói giữa X và Y có tương quan tuyến tính. Sự tương quan này mạnh hay yếu tùy thuộc vào hình elip là dẹt hay phình.



5.2.3. Hàm hồi qui

Xét cặp đại lượng ngẫu nhiên (X,Y) trên một tổng thể. Với mỗi giá trị x thuộc tập giá trị của X, đặt $f(x) := E(Y \mid X = x)$ là giá trị trung bình của Y với điều kiện $X = x$. Hàm $f(x)$ này được gọi là *hàm hồi qui* (regression function) của Y theo X. Đồ thị của hàm hồi qui $f(x)$ gọi là đường hồi qui của Y theo X.

Nếu hàm hồi qui là hàm bậc nhất (đồ thị hàm hồi qui là đường thẳng) thì ta nói đó là hàm hồi qui tuyến tính (linear regression).

Nhận xét. Khi $y = f(x)$ là hàm hồi qui của Y theo X, để tiện người ta có thể nói hàm hồi qui của Y theo X là $Y = f(X)$.

Với mỗi giá trị y thuộc tập giá trị của Y, đặt $g(y) := E(X \mid Y = y)$. Hàm $g(y)$ này được gọi là hàm hồi qui của X theo Y.

5.2.4. Hệ số tương quan

Hệ số tương quan của hai đại lượng ngẫu nhiên X và Y là

$$\rho = \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Người ta chứng minh được $|\rho| \le 1$ và $\rho = \pm 1$ khi và chỉ khi giữa X và Y có quan hệ tuyến tính

$$Y = aX + b \text{ hay } X = aY + b.$$

Hệ số tương quan ρ là số đo mức độ phụ thuộc tuyến tính của hai đại lượng ngẫu nhiên X,Y. Nếu $|\rho|$ càng gần 1 thì mức độ phụ thuộc tuyến tính giữa chúng càng lớn.

Thực tế, người ta qui ước:

- Nếu $0,7 \le \rho \le 1$ thì X,Y có tương quan tuyến tính mạnh.
- Nếu $|\rho| \le 3$ thì tương quan tuyến tính giữa X và Y là yếu.

- Nếu $\rho > 0$ thì coi X và Y là đồng biến.
- Nếu $\rho < 0$ thì coi X và Y là nghịch biến.

Hệ số tương quan ρ_{XY} có điểm thuận tiện hơn $Cov(X,Y)$ là không phụ thuộc đơn vị đo của X,Y.

5.2.5. Hệ số tương quan mẫu

Xét cặp đại lượng ngẫu nhiên (X,Y) trên cùng một tổng thể và xét sự tương quan giữa X và Y. Giả sử ta lấy được mẫu cụ thể kích thước n gồm n cặp số liệu

$$\{(x_i,y_i), i = 1, 2,\ldots,n\}.$$

Hệ số tương quan mẫu của cặp đại lượng ngẫu nhiên (X,Y) được định nghĩa là trị số

$$r = \frac{n}{n-1}\left(\frac{\overline{xy} - \overline{x}.\overline{y}}{s_X s_Y}\right)$$

với s_X, s_Y là các độ lệch chuẩn mẫu (hiệu chỉnh) cụ thể của biến X và biến Y. Chú ý rằng, khi kích thước mẫu n đủ lớn, ta có $r \approx \rho_{XY}$.

Ví dụ. Tính hệ số tương quan mẫu của mẫu cụ thể 10 cặp số liệu sau

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2	6	7	4	8	8	13	10	14	9

Giải. Ta tính được

$$\sum x_i = 55;\ \sum y_i = 81;\ \sum x_i^2 = 385;\ \sum y_i^2 = 779;\ \sum x_i y_i = 526;$$

$$n = 10;\ \overline{x} = 5,5;\ \overline{y} = 8,1;\ \overline{xy} = 52,656;$$

$$s_X^2 = (385 - 10.5,52)/9 = 9,1667;\ s_Y^2 = (779 - 10.8,12)/9 = 13,65;$$

$$r = \frac{n}{n-1}\left(\frac{\overline{xy} - \overline{x}.\overline{y}}{s_X s_Y}\right) = \frac{10}{9} \times \frac{52,6 - 5,5 \times 8,1}{\sqrt{9,1667 \times 13,6556}} \approx 0,7995.$$

5.3. Hồi qui tuyến tính

5.3.1. Đường hồi qui thực nghiệm

Với mẫu cụ thể kích thước *n* của cặp đại lượng ngẫu nhiên (X,Y) gồm các cặp trị (x_i,y_i) (có thể trùng nhau), $i = 1,\ldots,n$, ta vẽ *đường hồi qui thực nghiệm* của *Y* theo *X* như sau

- Tính $f(x_i) = \overline{y}_{x_i}$ với $\overline{y}_{x_i}$ là trung bình các giá trị của Y trong mẫu ứng với $X = x_i$.
- Vẽ các điểm $(x_i,\ f(x_i)) \equiv (x_i, \overline{y}_{x_i})$ này lên mặt phẳng Oxy.
- Nối các điểm này lại.

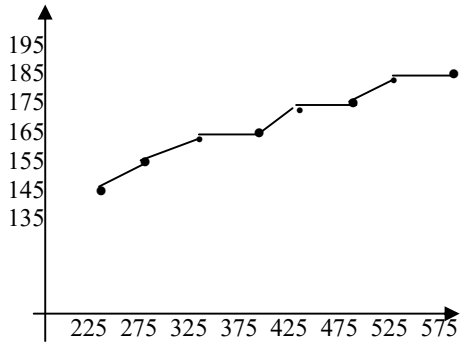
Kết quả ta có đường gấp khúc gọi là đường hồi qui thực nghiệm của Y theo X. Tương tự, ta cũng vẽ được đường hồi qui thực nghiệm của X theo Y.

Ví dụ. Gọi $X(kg)$ là trọng lượng và $Y(cm)$ là chiều dài lồng ngực của bò đực trong đàn bò của trại chăn nuôi. Lấy ngẫu nhiên 300 con bò đực cụ thể và đo được số liệu

Y \ X	225	275	325	375	425	475	525	575	n_i	$\overline{x}_{y_i}$
195								1	1	575
185					1	9	15	2	27	508
175			4	25	35	21	9	1	95	429,7
165		3	40	44	24	8			119	372,5
155	1	17	17	17	1				53	325
145	2	1	1						4	262,5
135	1								1	225
m_j	4	21	62	86	61	38	24	4	300	
$\overline{y}_{x_i}$	145	156	163	166	171	175	181	185		

Hãy vẽ đường hồi qui thực nghiệm của Y theo X.

Giải. Để tiện, dựa vào bảng số liệu, ta tính lên bảng các giá trị của $\overline{y}_{x_i}$, $\overline{x}_{y_i}$.



Các điểm của đường hồi qui thực nghiệm này xấp xỉ thẳng hàng. Đưa đến ta có thể coi hàm hồi qui của Y theo X là hàm hồi qui tuyến tính.

5.3.2. Phương trình đường hồi qui tuyến tính

Xét cặp đại lượng ngẫu nhiên (X,Y) trên một tổng thể. Giả sử coi hàm hồi qui của Y theo X là hàm hồi qui tuyến tính. Cụ thể ta coi $Y = aX + b$, với a,b cần xác định.

Lấy mẫu cụ thể kích thước n gồm các cặp số liệu (x_i,y_i) của cặp đại lượng ngẫu nhiên (X,Y). Ta có tổng các bình phương sai số là

$$S = \sum (ax_i + b - y_i)^2 \equiv S(a,b).$$

Xác định a,b là các giá trị để tổng các bình phương sai số S là bé nhất. Cặp trị (a,b) này sẽ là điểm tới hạn của hàm hai biến $S(a,b)$. Đưa đến giải hệ

$$\begin{cases} S'_a \equiv 2\sum (ax_i + b - y_i)x_i = 0 \\ S'_b \equiv \sum (ax_i + b - y_i) = 0 \end{cases} \Leftrightarrow \begin{cases} a\sum x_i^2 + b\sum x_i = \sum x_i y_i \\ a\sum x_i + bn = \sum y_i \end{cases}.$$

Với điều kiện $s_X^2 \neq 0$ (luôn thỏa khi tồn tại các trị x_i khác nhau), ta giải ra được

$$\begin{cases} a = \dfrac{n\sum x_i y_i - \sum x_i . \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \dfrac{\dfrac{\sum x_i y_i}{n} - \dfrac{\sum x_i}{n} \cdot \dfrac{\sum y_i}{n}}{\dfrac{1}{n}\left[\sum x_i^2 - n\left(\dfrac{\sum x_i}{n}\right)^2\right]} = \dfrac{n}{(n-1)} \cdot \dfrac{\overline{xy} - \overline{x}.\overline{y}}{s_X^2}, \\ b = \dfrac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}. \end{cases}$$

Biến đổi ta đưa đến

$$\begin{cases} a = r.\dfrac{s_Y}{s_X}, \\ b = \overline{y} - r.\dfrac{s_Y}{s_X}.\overline{x}. \end{cases}$$

Đưa đến phương trình đường hồi qui tuyến tính của Y theo X là:

$$Y = r\frac{s_Y}{s_X}(X - \overline{x}) + \overline{y}$$

và dạng khác là

$$Y - \overline{y} = r\frac{s_Y}{s_X}(X - \overline{x})$$

Đổi vai trò các biến, ta có phương trình đường hồi qui tuyến tính của X theo Y là

$$X - \overline{x} = r\frac{s_X}{s_Y}(Y - \overline{y}).$$

Nhận xét. Nói chung hai phương trình trên xác định hai đường thẳng khác nhau. Khi $r = \pm 1$, hai đường hồi qui trên trùng nhau.

Ví dụ 1. Với số liệu ở ví dụ trên, tìm phương trình đường hồi qui tuyến tính của chiều dài lồng ngực $Y\,(cm)$ theo trọng lượng $X\,(kg)$ của bò đực trong đàn bò của trại chăn nuôi.

Giải. Ta có

$$\overline{x} = \frac{\sum m_j x_j}{n} = \frac{117950}{300} = 393{,}1667;\ \overline{y} = \frac{\sum n_i y_i}{n} = \frac{50380}{300} = 167{,}9333;$$

$$\overline{xy} = \frac{\sum n_{ij} x_j y_i}{n} = \frac{19961750}{300} = 66539{,}16667;$$

$$s_X^2 = \frac{1}{n-1}\left(\sum m_j x_j^2 - n(\overline{x})^2\right) = \frac{47962500 - 300 \times 393{,}16672}{299} = 5312{,}6549;$$

$$s_Y^2 = \frac{1}{n-1}\left(\sum n_i y_i^2 - n(\overline{y})^2\right) = \frac{8486900 - 300 \times 167{,}93332}{299} = 88{,}368;$$

$$s_X = 72{,}888;\ s_Y = 9{,}4004;$$

$$r = \frac{n}{n-1}\left(\frac{\overline{xy} - \overline{x}.\overline{y}}{s_X s_Y}\right) = \frac{300}{299}\left(\frac{66539{,}16667 - 393{,}1667 \times 167{,}9333}{72{,}888 \times 9{,}4004}\right) = 0{,}7518.$$

Phương trình đường hồi qui của Y theo X là

$$Y-\overline{y}=r\frac{s_Y}{s_X}(X-\overline{x})\Leftrightarrow Y-167,9333=0,097(X-393,1667)$$
$$\Leftrightarrow Y=0,097X+129,8119$$

Ví dụ 2. Tìm phương trình đường hồi qui tuyến tính của đại lượng ngẫu nhiên Y theo X biết mẫu cụ thể 10 cặp số liệu của (X,Y) ở bảng sau

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2	6	7	4	8	8	13	10	14	9

Giải. Ta có

$$n=10;\ \overline{x}=\frac{\sum x_i}{n}=5,5;\ \overline{y}=\frac{\sum y_i}{n}=8,1;\ \overline{xy}=\frac{\sum x_iy_i}{n}=52,6;$$
$$s_X^2=\frac{\sum x_i^2-n(\overline{x})^2}{n-1}=9,1667;\ s_Y^2=\frac{\sum y_i^2-n(\overline{y})^2}{n-1}=13,6556;$$
$$r=\frac{n}{n-1}\left(\frac{\overline{xy}-\overline{x}.\overline{y}}{s_Xs_Y}\right)=\frac{10}{9}\times\frac{52,6-5,5\times8,1}{\sqrt{9,1667\times13,6556}}\approx0,7995$$

Phương trình đường hồi qui tuyến tính của Y theo X là

$$Y-\overline{y}=r\frac{s_Y}{s_X}(X-\overline{x})\Leftrightarrow Y-8,1=0,7995\times\sqrt{\frac{13,6556}{9,1667}}(X-5,5)$$
$$\Leftrightarrow Y=0,9758X+2,7331$$

5.3.3. Hồi qui bậc hai

Xét cặp đại lượng ngẫu nhiên (X,Y) trên một tổng thể. Giả sử coi mối quan hệ giữa Y và X là bậc hai, cụ thể coi

$$Y=aX^2+bX+c$$

Lấy mẫu cụ thể kích thước *n* gồm các cặp (x_i,y_i). Ta có

$$S=\sum(ax_i^2+bx_i+c-y_i)^2\equiv S(a,b,c)$$

là tổng bình phương các sai số. Xác định a,b,c để tổng bình phương sai số là bé nhất, đưa đến giải hệ

$$\begin{cases}S'_a=0\\S'_b=0\\S'_c=0\end{cases}\Leftrightarrow\ldots\Leftrightarrow\begin{cases}a\sum x_i^4+b\sum x_i^3+c\sum x_i^2=\sum x_i^2y_i\\a\sum x_i^3+b\sum x_i^2+c\sum x_i=\sum x_iy_i\\a\sum x_i^2+b\sum x_i+cn=\sum y_i\end{cases}.$$

Giải hệ ta được a,b,c cần tìm.

Ví dụ. Xét cặp đại lượng ngẫu nhiên (X,Y) trên tổng thể và xét sự tương quan giữa chúng. Giả sử có mẫu cụ thể như sau

x_i	1	2	3	4	5	6	7	8	9
y_i	2	3	6	8	9	9	10	8	7

Giả thiết rằng coi $Y=aX^2+bX+c$. Hãy xác định phương trình đường hồi qui bậc hai này.

Giải. Ta có

$$n = 9; \quad \sum x_i = 45; \quad \sum y_i = 61; \sum x_i^2 = 285;$$
$$\sum x_i^3 = 2025; \ \sum x_i^4 = 15333; \ \sum x_i^2 y_i = 2113; \ \sum x_i y_i = 353.$$

Bằng phương pháp tổng bình phương sai số bé nhất, đưa đến giải hệ

$$\begin{cases} 15333\,a + 2025\,b + 285\,c = 2113 \\ 2025\,a + 285\,b + 45\,c = 353 \\ 285\,a + 45\,b + 9\,c = 61. \end{cases}$$

Giải hệ ta được

$$a \approx -0,3203; \ b \approx 4,003; \ c \approx -3,08$$

Vậy phương trình hồi qui bậc hai của Y và X là

$$Y = -0,3203X^2 + 4,003X - 3,08\,.$$

5.4. Bài tập chương 5

1. Trên một sàn giao dịch chứng khoán có hai loại cổ phiếu KHP và ACB được bán và lãi suất tương ứng của chúng là hai đại lượng ngẫu nhiên X và Y. Giả sử (X,Y) có bảng phân phối xác suất như sau

Y / ***X***	**- 2**	**0**	**5**	**10**
0	0	0,05	0,05	0,1
4	0,05	0,1	0,25	0,15
6	0,1	0,05	0,1	0

a) Để đạt lãi suất kỳ vọng cao nhất thì nên đầu tư vào cả hai loại cổ phiếu theo tỷ lệ nào?

b) Để hạn chế rủi ro lãi suất đến mức thấp nhất thì đầu tư hai loại cổ phiếu theo tỷ lệ nào?

2. Điều tra thu nhập hàng năm (triệu đồng/năm) của các cặp vợ chồng đang làm việc tại một nhà máy, thu được kết quả sau

***Y* (thu nhập của vợ)** / ***X* (thu nhập của chồng)**	**10**	**20**	**30**	**40**
10	0,2	0,04	0,01	0
20	0,1	0,36	0,09	0
30	0	0,05	0,10	0
40	0	0	0	0,05

a) Tìm phân phối xác suất thu nhập của chồng và thu nhập của vợ.

b) Tìm phân phối thu nhập của những người vợ có chồng thu nhập 20 triệu/năm.

c) Tính thu nhập trung bình của các bà vợ có chồng thu nhập ở mức 20triệu/năm.

d) Thu nhập của chồng và vợ có phụ thuộc nhau không và nếu có thì phụ thuộc như thế nào?

3. Một kiện hàng có 10 sản phẩm, trong đó có 4 sản phẩm loại A. Một máy sản xuất sản phẩm với xác suất sản suất ra sản phẩm loại A là 0,2. Lấy ngẫu nhiên không hoàn lại từ kiện

ra 2 sản phẩm và cho máy sản xuất ra 2 sản phẩm. Gọi X là số sản phẩm loại A trong 4 sản phẩm đó. Lập bảng phân phối xác suất của X và tính $E(X), D(X)$.

4. Tiến hành quan sát về hai chỉ tiêu X và Y trên một tổng thể, ta thu được mẫu số liệu

X	0,25	0,37	0,44	0,55	0,60	0,62	0,68	0,70	0,73	0,75	0,92	0,84	0,87	0,88	0,90	0,95	1
Y	2,57	2,31	2,12	1,92	1,75	1,71	1,60	1,51	1,50	1,41	1,33	1,31	1,25	1,20	1,19	1,15	1

a) Tính $\overline{x}, \overline{y}, \overline{xy}, s_X^2, s_Y^2, r$.

b) Viết phương trình đường hồi qui tuyến tính của *Y* theo *X*.

5. Điều tra thu nhập của 10 cặp vợ chồng (triệu đồng/năm) thu được kết quả sau

X (thu nhËp cña chång)	20	30	30	20	20	30	40	30	40	40
Y (thu nhËp cña vî)	15	35	25	25	25	15	25	25	35	25

a) Lập bảng phân phối xác suất đồng thời của (X,Y).

b) Lập bảng phân phối xác suất biên của X. Tính E[*X*] và D[*X*].

c) Lập bảng phân phối xác suất biên của Y. Tính E[*Y*] và D[*Y*].

d) Tính Cov (*X*, *Y*); X và Y có độc lập với nhau không?

e) Giả sử thu nhập sau thuế W của các cặp vợ chồng được xác định bởi biểu thức: $W = 0,6X + 0,8Y$. Tính E[*W*] và D[*W*].

6. Nghiên cứu về thu nhập và tỉ lệ thu nhập chi cho giáo dục của các hộ gia đình ở một vùng, điều tra 400 hộ, ta thu được số liệu

Y \ *X*	10	20	30	40	50
150 - 250	40	20	10		
250 - 350		60	40	10	
350 - 450		20	70	60	
450 - 550			10	20	40

Trong đó $X(\%)$ là tỉ lệ thu nhập chi cho giáo dục, Y (USD/tháng) là thu nhập bình quân của một người trong gia đình.

a) Tìm hệ số tương quan mẫu giữa X và Y.

b) Lập phương trình đường hồi quy tuyến tính của Y theo X.

7. Nghiên cứu về X (ngàn đồng) là thu nhập bình quân tháng mỗi người của hộ gia đình và Y (%) là tỷ lệ thu nhập chi cho ăn uống của hộ gia đình trong một vùng, điều tra 400 hộ gia đình ở vùng đó, ta có bảng số liệu thực nghiệm sau đây

X \ *Y*	10	20	30	40	50
50 - 150			30	30	10
150 - 250		20	80	40	
250 - 350		40	60	20	
350 - 450	10	40	20		

a) Tìm khoảng tin cậy cho tỷ lệ thu nhập trung bình chi cho ăn uống của một gia đình với mức tin cậy 95%.

b) Những hộ gia đình có *thu nhập cao* trong vùng là những hộ có thu nhập bình quân mỗi người/tháng là trên 350.000 đồng. Nếu nói rằng tỷ lệ hộ gia đình có thu nhập cao trong toàn vùng là 20% với mức ý nghĩa 5% thì bạn có chấp nhận được không?

c) Tìm hệ số tương quan mẫu giữa X và Y. Lập phương trình đường hồi quy tuyến tính của Y theo X.

8. Một loại sản phẩm nào đó được đánh giá chất lượng qua hai chỉ tiêu X, Y nào đó. Kiểm tra một số sản phẩm về hai chỉ tiêu ở trên ta có kết quả sau đây

X \ Y	0 - 4	4 - 8	8 - 12	12 - 16	16 - 22
115 - 125	4				
125 - 135	6	8	10		
135 - 145		12	15	2	
145 - 155	10	10	14	7	5
155 - 165				4	3

a) Yêu cầu đạt tiêu chuẩn của chỉ tiêu X là 145. Có người cho rằng chỉ tiêu X trung bình là nhỏ hơn yêu cầu. Với mức ý nghĩa 5%, bạn hãy cho biết ý kiến của mình.

b) Sản phẩm có chỉ tiêu Y lớn hơn 12 là sản phẩm loại I. Hãy ước lượng trung bình chỉ tiêu Y của sản phẩm loại I với mức tin cậy 90%. Cho biết chỉ tiêu Y của sản phẩm loại I tuân theo quy luật chuẩn.

c) Tìm hệ số tương quan mẫu giữa X và Y. Lập phương trình đường hồi quy tuyến tính của Y theo X.

Mục lục

Chương 1. Biến cố và xác suất của các biến cố **3**

1.1. Các khái niệm về phép thử, biến cố, không gian mẫu 3
1.1.1. Phép thử ngẫu nhiên (Random experiment) 3
1.1.2. Biến cố và phân loại các biến cố 3
1.2. Các phép toán về biến cố 4
1.2.1. Biến cố tổng 4
1.2.2. Biến cố tích 4
1.2.3. Biến cố đối 5
1.2.4. Hệ đầy đủ 5
1.3. Các định nghĩa về xác suất 5
1.3.1. Định nghĩa xác suất cổ điển 6
1.3.2. Định nghĩa xác suất theo thống kê 7
1.4. Công thức cộng xác suất và nhân xác suất 7
1.4.1. Công thức cộng xác suất 7
1.4.2. Xác suất có điều kiện và công thức nhân xác suất 8
1.5. Công thức xác suất đầy đủ và công thức Bayes 10
1.5.1. Công thức xác suất đầy đủ 10
1.5.2. Công thức Bayes 10
1.6. Bài tập chương 1 11

Chương 2. Đại lượng ngẫu nhiên **14**

2.1. Đại lượng ngẫu nhiên 14
2.1.1. Khái niệm 14
2.1.2. Phân loại đại lượng ngẫu nhiên 14
2.2. Phân phối xác suất của đại lượng ngẫu nhiên 15
2.2.1. Bảng phân phối xác suất 15
2.2.2. Hàm phân phối xác suất 16
2.2.3. Hàm mật độ xác suất 17
2.3. Các tham số đặc trưng của dlnn 19
2.3.1. Kỳ vọng toán 19
2.3.2. Phương sai 20
2.3.3. Giá trị tin chắc nhất 21
2.4. Các phân phối xác suất thông dụng 22
2.4.1. Phân phối nhị thức 22
2.4.2. Phân phối Poisson 23
2.4.3. Phân phối chuẩn 24
2.4.4. Phân phối khi-bình-phương 27
2.4.5. Phân phối Student 27
2.4.6. Mối quan hệ giữa các phân phối xác suất 28
2.5. Bài tập chương 2 29

Chương 3. Mẫu thống kê và ước lượng tham số **33**

3.1. Mẫu và thống kê mô tả 33
3.1.1. Tổng thể và mẫu nn 33
3.1.2. Biểu diễn số liệu cụ thể 35
3.2. Các tham số đặc trưng của mẫu ngẫu nhiên và tính chất 37
3.2.1. Các tham số đặc trưng tương ứng của tổng thể và mẫu 37
3.2.2. Các tính chất của đặc trưng mẫu 37
3.3. Ước lượng điểm 38
3.3.1. Một số khái niệm về ước lượng tham số 38
3.3.2. Các tính chất của ước lượng điểm 39
3.4. Ước lượng khoảng 42

3.4.1. Bài toán ước lượng khoảng cho trung bình tổng thể 42
3.4.2. Bài toán ước lượng khoảng cho tỉ lệ tổng thể 44
3.4.3. Bài toán ước lượng khoảng cho phương sai tổng thể 45
3.5. Các bài toán liên quan đến bài toán ước lượng 46
3.5.1. Bài toán xác định cỡ mẫu 46
3.5.2. Bài toán xác định độ tin cậy 47
3.6. Bài tập chương 3 48

Chương 4. Kiểm định giả thuyết thống kê 52

4.1. Các khái niệm 52
4.1.1. Giả thuyết thống kê 52
4.1.2. Tiêu chuẩn kiểm định, mức tin cậy, miền bác bỏ, sai lầm loại 1 và sai lầm loại 2 52
4.2. Một số bài toán kiểm định thường gặp 53
4.2.1. Bài toán kiểm định giả thuyết trung bình tổng thể 53
4.2.2. Bài toán kiểm định giả thuyết tỉ lệ tổng thể 56
4.2.3. Bài toán so sánh hai trung bình tổng thể 58
4.2.4. Bài toán so sánh hai tỉ lệ tổng thể 60
4.2.5. Bài toán kiểm định giả thuyết phương sai tổng thể 61
4.2.6. Bài toán kiểm định giả thuyết về tính độc lập 62
4.3. Bài tập chương 4 63

Chương 5. Tương quan và hồi quy 68

5.1. Đại lượng ngẫu nhiên hai chiều 68
5.1.1. Khái niệm 68
5.1.2. Bảng phân phối xác suất của đại lượng ngẫu nhiên rời rạc 68
5.1.3. Hiệp phương sai 69
5.1.4. Hàm mật độ của đại lượng ngẫu nhiên hai chiều liên tục 71
5.1.5. Phân phối có điều kiện của đại lượng ngẫu nhiên rời rạc 72
5.2. Hàm hồi quy và hệ số tương quan 73
5.2.1. Mẫu ngẫu nhiên hai chiều 73
5.2.2. Đám mây quan sát 73
5.2.3. Hàm hồi qui 74
5.2.4. Hệ số tương quan 74
5.2.5. Hệ số tương quan mẫu 75
5.3. Hồi qui tuyến tính 75
5.3.1. Đường hồi qui thực nghiệm 75
5.3.2. Phương trình đường hồi qui tuyến tính 76
5.3.3. Hồi qui bậc hai 78
5.4. Bài tập chương 5 79

Tài liệu tham khảo 84

Tài liệu tham khảo

[1] Đặng Hấn, *Xác suất và thống kê*. NXB Thống kê, 2004.

[2] Hoàng Ngọc Nhậm, *Lý thuyết xác suất và thống kê*. ĐH Kinh tế TP. HCM, 2007.

[3] Lê Bá Phi, *Bài giảng xác suất và thống kê*. ĐH Nha Trang, 2005.

[4] Nguyễn Bác Văn, *Xác suất và xử lý số liệu thống kê*. NXB Giáo dục, 1998.

[5] Nguyễn Cao Văn, *Bài giảng xác xuất thống kê*. ĐH Kinh tế quốc dân Hà nội, 2005.

[6] Đặng Hùng Thắng, *Thống kê và ứng dụng*. NXB Giáo dục, 1999.

[7] Nguyễn Đình Thúc, Đặng Hải Vân, Lê Phong. *Giáo trình thống kê máy tính*. NXB Khoa học và Kỹ thuật, 2010.

[8] Nguyễn Duy Tiến, Vũ Việt Yên, *Lý thuyết xác suất*. NXB Giáo dục, 2000.

[9] Lê Văn Tiến. *Giáo trình lý thuyết xác suất và thống kê toán học*. NXB Đại học và Trung học chuyên nghiệp, 1991.

[10] Harald Cramer, *Phương pháp toán học trong thống kê*. NXB Khoa học, 1969.

[11] Calvin Dytham, *Choosing and Using Statistics: A Biologist's Guide*. Department of Biology, University of York. Blackwell Science, 1999.