Summary MAS

Mục lục

1 Probability

Probability (Xác suất) là một thước đo cho biết mức độ khả năng một sự kiện nào đó sẽ xảy ra. Nó có giá trị từ 0 đến 1.

1.1 Không gian xác suất (Probability Space)

Không gian xác suất là cơ sở để định nghĩa các sự kiện và tính toán xác suất. Nó bao gồm ba thành phần:

- **Sample space** Ω: Tập hợp tất cả các kết quả có thể xảy ra của một thí nghiệm ngẫu nhiên.
- **Events**: Là các tập con của sample space, đại diện cho các kết quả mà chúng ta quan tâm.
- **Probability measure (P)**: Hàm gán xác suất cho các sự kiện, đảm bảo các quy tắc xác suất (axioms) được thỏa mãn.

1.2 Sự kiện (Events)

Sự kiện là bất kỳ tập con nào của sample space. Ví dụ, trong một lần tung đồng xu, sự kiện có thể là "ra mặt sấp" hoặc "ra mặt ngửa". Các sự kiện có thể đơn giản (chỉ có một kết quả) hoặc phức tạp (chứa nhiều kết quả).

1.3 Các định đề xác suất(Axioms of Probability)

Các định đề này là nền tảng của lý thuyết xác suất:

- **Không âm**: Xác suất của mọi sự kiện luôn không âm, $P(A) \geq 0$.
- **Xác suất của sample space**: Tổng xác suất của toàn bộ sample space luôn bằng 1, $P(\Omega) = 1$.
- **Cộng xác suất**: Nếu A và B là hai sự kiện không giao nhau, thì $P(A \cup B) = P(A) + P(B)$.

1.4 Xác suất có điều kiện (Conditional Probability)

Xác suất có điều kiện của một sự kiện xảy ra khi biết một sự kiện khác đã xảy ra, ký hiệu là $P(A|B)$ và được tính bằng:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{với } P(B) > 0$$

1.5 Quy tắc nhân xác suất (Multiplication Rule)

Quy tắc này dùng để tính xác suất đồng thời của hai sự kiện A và B:

$$P(A \cap B) = P(A|B) \cdot P(B)$$

1.6 Quy tắc xác suất toàn phần (Total Probability Rule)

Quy tắc xác suất toàn phần dùng khi chúng ta có một tập hợp các sự kiện rời rạc và phủ hết sample space $B_1, B_2, \ldots, B_n$ với $P(B_i) > 0$. Khi đó, xác suất của một sự kiện A có thể được tính như sau:

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

1.7 Độc lập (Independence)

Hai sự kiện A và B được gọi là độc lập nếu xác suất xảy ra của một sự kiện không bị ảnh hưởng bởi sự kiện kia, tức là:

$$P(A \cap B) = P(A) \cdot P(B)$$

1.8 Bayes's Theorem (Định lý Bayes)

Định lý Bayes giúp tính xác suất có điều kiện của một sự kiện với sự kiện khác đã xảy ra, đặc biệt hữu ích trong các bài toán suy luận ngược:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Định lý này cho phép cập nhật xác suất của giả thuyết A dựa trên bằng chứng B.

2 Khái niệm Biến Ngẫu Nhiên

Biến ngẫu nhiên có thể được phân loại thành hai loại chính: biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục.
Biến ngẫu nhiên rời rạc là biến ngẫu nhiên chỉ nhận các giá trị rời rạc, thường là các số nguyên. Xác suất để biến ngẫu nhiên rời rạc nhận một giá trị cụ thể có thể khác 0, và được mô tả bằng **hàm mật độ** $f(x) = P(X = x)$.
Biến ngẫu nhiên liên tục là biến ngẫu nhiên có thể nhận giá trị trong một khoảng liên tục trên trục số thực. Thay vì xác suất cho một giá trị cụ thể, ta xem xét xác suất để biến này nằm trong một khoảng giá trị thông qua hàm mật độ xác suất (density function).

2.1 Kỳ Vọng và Phương Sai

Kỳ vọng của biến ngẫu nhiên X là giá trị trung bình của X, ký hiệu là $E(X)$, và được tính bởi:
- Với biến rời rạc:

$$E(X) = \sum_x x f(x) = \sum_x x P(X = x)$$

- Với biến liên tục:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

Phương sai của biến ngẫu nhiên X, ký hiệu $Var(X)$, đo độ phân tán của các giá trị X quanh kỳ vọng và được tính bởi:
- Với biến rời rạc:

$$Var(X) = \sum_x (x - E(X))^2 f(x) = \sum_x (x - E(X))^2 P(X = x)$$

- Với biến liên tục:

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\, dx$$

2.2 Hàm Phân Phối Tích Lũy (Cumulative Distribution Function - CDF)

Hàm phân phối tích lũy của biến ngẫu nhiên X, ký hiệu là $F(x)$, cho biết xác suất mà biến ngẫu nhiên X nhận một giá trị nhỏ hơn hoặc bằng x:

$$F(x) = P(X \leq x)$$

Đối với biến ngẫu nhiên rời rạc: hàm phân phối tích lũy $F(x)$ được tính bằng cách cộng các xác suất của các giá trị mà X có thể nhận được và nhỏ hơn hoặc bằng x:

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t) = \sum_{t \leq x} P(X = t)$$

trong đó $f(t) = P(X = t)$ là hàm mật độ xác suất (probability mass function) của X.
Đối với biến ngẫu nhiên liên tục: hàm phân phối tích lũy $F(x)$ được tính bằng cách tích phân hàm mật độ xác suất $f(x)$ từ $-\infty$ đến x:

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

trong đó $f(t)$ là hàm mật độ xác suất (probability density function) của X.

3 Các Biến Ngẫu Nhiên Rời Rạc Phổ Biến

3.1 Phân phối đều rời rạc (Discrete Uniform Distribution)

- **Ký hiệu**: $X \sim U\{a, b\}$
- **Ý nghĩa**: Mô tả một tập hợp hữu hạn các giá trị rời rạc từ a đến b, với mỗi giá trị đều có xác suất như nhau.
- **Hàm mật độ**:
$$f(x) = P(X = x) = \frac{1}{b - a + 1}, \quad x = a, a + 1, \ldots, b$$
- **Kỳ vọng**: $E(X) = \frac{a+b}{2}$
- **Phương sai**: $Var(X) = \frac{(b-a+1)^2-1}{12}$

3.2 Phân phối Nhị thức (Binomial Distribution)

- **Ký hiệu**: $X \sim Bin(n, p)$
- **Ý nghĩa**: Mô tả số lần thành công trong n phép thử độc lập, mỗi phép thử có xác suất thành công là p.
- **Hàm mật độ**:
$$f(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n$$
- **Kỳ vọng**: $E(X) = np$
- **Phương sai**: $Var(X) = np(1 - p)$

3.3 Phân phối Hình học (Geometric Distribution)

- **Ký hiệu**: $X \sim Geo(p)$
- **Ý nghĩa**: Mô tả số lần thử cần thiết để đạt được thành công đầu tiên, với xác suất thành công là p trong mỗi lần thử.
- **Hàm mật độ**:
$$f(x) = P(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, 3, \ldots$$
- **Kỳ vọng**: $E(X) = \frac{1}{p}$
- **Phương sai**: $Var(X) = \frac{1-p}{p^2}$

3.4 Phân phối Nhị thức âm (Negative Binomial Distribution)

- **Ký hiệu**: $X \sim NB(r, p)$
- **Ý nghĩa**: Mô tả số lần thử cần thiết để đạt được r thành công, với xác suất thành công là p trong mỗi lần thử.
- **Hàm mật độ**:
$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, r+2, \ldots$$
- **Kỳ vọng**: $E(X) = \frac{r}{p}$
- **Phương sai**: $Var(X) = \frac{r(1-p)}{p^2}$

3.5 Phân phối Siêu bội (Hypergeometric Distribution)

- **Ký hiệu**: $X \sim Hyp(N, K, n)$
- **Ý nghĩa**: Mô tả số lượng các thành công trong mẫu kích thước n được chọn ngẫu nhiên mà không hoàn lại từ tổng thể N phần tử, trong đó có K phần tử thành công.
- **Hàm mật độ**:
$$f(x) = P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$
- **Kỳ vọng**: $E(X) = n\frac{K}{N}$
- **Phương sai**: $Var(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$

3.6 Phân phối Poisson (Poisson Distribution)

- **Ký hiệu**: $X \sim Poisson(\lambda)$
- **Ý nghĩa**: Mô tả số lần xảy ra của một sự kiện trong một khoảng thời gian cố định với tần suất trung bình là λ.

- **Hàm mật độ**:
$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$
- **Kỳ vọng**: $E(X) = \lambda$
- **Phương sai**: $Var(X) = \lambda$

4 Các Biến Ngẫu Nhiên Liên Tục Phổ Biến

4.1 Phân phối đều liên tục (Continuous Uniform Distribution)

- **Ký hiệu**: $X \sim U(a, b)$
- **Ý nghĩa**: Mô tả các giá trị phân bố đều trong đoạn $[a, b]$, mỗi giá trị trong đoạn có xác suất như nhau.
- **Hàm mật độ**:
$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$
- **Hàm phân phối tích lũy**:
$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$
- **Kỳ vọng**: $E(X) = \frac{a+b}{2}$
- **Phương sai**: $Var(X) = \frac{(b-a)^2}{12}$

4.2 Phân phối Chuẩn (Normal Distribution)

- **Ký hiệu**: $X \sim N(\mu, \sigma^2)$
- **Ý nghĩa**: Mô tả các giá trị xung quanh trung bình μ với độ lệch chuẩn σ, có dạng hình chuông. Phân phối chuẩn thường xuất hiện trong tự nhiên và các dữ liệu lớn.
- **Hàm mật độ**:
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
- **Hàm phân phối tích lũy**: Không có công thức đóng, nhưng có thể tính qua hàm chuẩn hóa.
- **Kỳ vọng**: $E(X) = \mu$
- **Phương sai**: $Var(X) = \sigma^2$

4.3 Biến ngẫu nhiên chuẩn hóa (Standard Normal Random Variable)

- **Ký hiệu**: $X \sim N(0,1)$
- **Ý nghĩa**: Biến ngẫu nhiên chuẩn với trung bình 0 và độ lệch chuẩn 1, được sử dụng rộng rãi để chuẩn hóa các biến khác.
- **Hàm mật độ**:
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$
- **Hàm phân phối tích lũy**: Không có công thức đóng, nhưng có thể tính qua bảng phân phối chuẩn.
- **Kỳ vọng**: $E(X) = 0$
- **Phương sai**: $Var(X) = 1$

4.4 Phân phối Mũ (Exponential Distribution)

- **Ký hiệu**: $X \sim Exp(\lambda)$
- **Ý nghĩa**: Mô tả khoảng thời gian giữa các sự kiện trong một quá trình Poisson, với tần suất trung bình là λ.
- **Hàm mật độ**:
$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$
- **Hàm phân phối tích lũy**:
$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$
- **Kỳ vọng**: $E(X) = \frac{1}{\lambda}$
- **Phương sai**: $Var(X) = \frac{1}{\lambda^2}$

5 Descriptive Statistics

Descriptive Statistics (Thống kê mô tả) là một nhánh của thống kê tập trung vào việc tóm tắt và mô tả đặc điểm của một tập dữ liệu. Nó giúp chúng ta hiểu rõ hơn về dữ liệu bằng cách trình bày các thông tin tổng quan về trung tâm, độ phân tán, hình dạng và các đặc điểm quan trọng khác mà không đi sâu vào phân tích các mối quan hệ hoặc đưa ra dự đoán

5.1 Tóm tắt Dữ liệu bằng Số liệu (Numerical Summaries of Data)

Đây là các số liệu tóm tắt nhằm mô tả trung tâm, độ phân tán và hình dạng của dữ liệu, bao gồm các chỉ số như:

- **Mean (Trung bình)**: Giá trị trung bình của dữ liệu.
- **Median (Trung vị)**: Giá trị ở giữa khi dữ liệu được sắp xếp theo thứ tự.

- **Mode (Mode)**: Giá trị xuất hiện nhiều nhất.
- **Range (Biên độ)**: Hiệu giữa giá trị lớn nhất và nhỏ nhất.
- **Variance và Standard Deviation (Phương sai và Độ lệch chuẩn)**: Đo lường độ phân tán của dữ liệu.

Ví dụ: Giả sử chiều cao của một nhóm học sinh là 160, 165, 170, 160, 155 cm.

- Trung bình = $\frac{160+165+170+160+155}{5} = 162$ cm.
- Trung vị = 160 cm (sau khi sắp xếp dữ liệu: 155, 160, 160, 165, 170).
- **Độ lệch chuẩn**:
  - Tính độ lệch của mỗi giá trị so với trung bình:

$$(160-162)^2, (165-162)^2, (170-162)^2, (160-162)^2, (155-162)^2$$

  - Kết quả là: $4, 9, 64, 4, 49$.
  - Trung bình các độ lệch bình phương: $\frac{4+9+64+4+49}{5} = 26$.
  - Phương sai = 26.
  - Độ lệch chuẩn là căn bậc hai của phương sai: $\sqrt{26} \approx 5.1$ cm.

5.2 Biểu đồ Thân-Lá (Stem-and-Leaf Diagrams)

Biểu đồ thân-lá là một công cụ để tóm tắt và mô tả phân phối của một tập dữ liệu bằng cách chia nhỏ các giá trị.

- **Stem (Thân)** là các chữ số đầu của số liệu, giúp nhóm dữ liệu vào các phạm vi nhất định.
- **Leaf (Lá)** là chữ số cuối, cho thấy giá trị chi tiết trong mỗi nhóm.

Ví dụ: Dữ liệu điểm số: 45, 47, 52, 53, 53, 55, 57, 60.

- Biểu đồ thân-lá:
  - 4 | 5 7
  - 5 | 2 3 3 5 7
  - 6 | 0

  Biểu đồ này cho thấy các nhóm điểm số theo hàng chục và sự phân bố dữ liệu quanh các giá trị trung tâm.

5.3 Phân phối Tần suất và Biểu đồ Tần suất(Frequency Distributions and Histograms)

- **Frequency Distribution (Phân phối Tần suất)** là bảng thể hiện tần suất (số lần xuất hiện) của các giá trị hoặc nhóm giá trị trong tập dữ liệu.
- **Histogram (Biểu đồ Tần suất)** là biểu đồ dạng cột biểu diễn tần suất của các giá trị hoặc nhóm giá trị.

Ví dụ: Giả sử bạn khảo sát số sách mà mỗi sinh viên đọc trong tháng và có các khoảng: 0-2, 3-5, 6-8.

- Phân phối Tần suất:
  - 0-2 sách: 5 sinh viên
  - 3-5 sách: 10 sinh viên
  - 6-8 sách: 5 sinh viên
- Histogram có trục hoành là số sách (0-2, 3-5, 6-8) và trục tung là số sinh viên, giúp dễ dàng thấy được khoảng 3-5 có số sinh viên đọc nhiều nhất.

5.4 Biểu đồ Hộp (Box Plots)

Biểu đồ hộp tóm tắt dữ liệu dựa trên các giá trị như **Minimum** (Giá trị nhỏ nhất), **First Quartile (Q1)**, **Median** (Trung vị), **Third Quartile (Q3)**, và **Maximum** (Giá trị lớn nhất).
Ví dụ: Dữ liệu về thời gian hoàn thành bài kiểm tra (phút): 20, 22, 24, 24, 25, 27, 30, 35, 37, 40.

- **Minimum** = 20 phút.
- **Q1 (Tứ phân vị thứ nhất)**: Trung vị của nửa dưới tập dữ liệu (20, 22, 24, 24, 25) là Q1 = 24 phút.
- **Median (Trung vị)**: Trung vị của toàn bộ tập dữ liệu. Trung bình của 25 và 27 là Median = $\frac{25+27}{2} = 26$ phút.
- **Q3 (Tứ phân vị thứ ba)**: Trung vị của nửa trên tập dữ liệu (27, 30, 35, 37, 40) là Q3 = 35 phút.
- **Maximum** = 40 phút.

Interquartile Range (IQR):

$$\text{IQR} = Q3 - Q1 = 35 - 24 = 11 \text{ phút}$$

Outliers và Extreme Outliers:

- **Outliers**: Các giá trị nằm ngoài khoảng $[Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR}]$.
  - $Q1 - 1.5 \times \text{IQR} = 24 - 1.5 \times 11 = 7.5$
  - $Q3 + 1.5 \times \text{IQR} = 35 + 1.5 \times 11 = 51.5$

  - Do đó, các giá trị ngoài khoảng $[7.5, 51.5]$ được xem là outliers. Trong dữ liệu này, không có giá trị nào ngoài khoảng này, nên không có outlier.

- **Extreme Outliers**: Các giá trị nằm ngoài khoảng $[Q1 - 3 \times \text{IQR}, Q3 + 3 \times \text{IQR}]$.
  - $Q1 - 3 \times \text{IQR} = 24 - 3 \times 11 = -9$
  - $Q3 + 3 \times \text{IQR} = 35 + 3 \times 11 = 68$
  - Do đó, các giá trị ngoài khoảng $[-9, 68]$ được xem là extreme outliers. Trong dữ liệu này, không có extreme outliers vì tất cả các giá trị đều nằm trong khoảng này.

6 Statistical Inference

Thống kê suy diễn (*statistical inference*) sử dụng dữ liệu mẫu để đưa ra các quyết định hoặc kết luận về tổng thể. Hai lĩnh vực chính của thống kê suy diễn là **ước lượng tham số** và **kiểm định giả thuyết**.

6.1 Phân phối mẫu (Sampling Distribution)

Khi lấy nhiều mẫu từ một tổng thể, các giá trị mẫu có thể khác nhau giữa các mẫu, dẫn đến sự biến thiên trong các thống kê mẫu như trung bình mẫu hoặc phương sai mẫu. Phân phối các giá trị thống kê từ nhiều mẫu gọi là **phân phối mẫu**.

6.2 Thống kê (Statistic)

Thống kê là bất kỳ giá trị nào được tính từ dữ liệu mẫu. Ví dụ, trung bình mẫu $\bar{X}$ và phương sai mẫu S^2 là các thống kê và có thể thay đổi theo từng mẫu.

6.3 Ước lượng điểm (Point Estimate)

Ước lượng điểm là một giá trị đơn được tính từ dữ liệu mẫu để ước lượng tham số tổng thể. Ví dụ, trung bình mẫu là ước lượng điểm của trung bình tổng thể.

6.4 Định lý Giới hạn Trung tâm (Central Limit Theorem - CLT)

Định lý Giới hạn Trung tâm (CLT) cho biết rằng khi kích thước mẫu n đủ lớn, phân phối của trung bình mẫu $\bar{X}$ sẽ xấp xỉ phân phối chuẩn, ngay cả khi tổng thể ban đầu không phải là phân phối chuẩn.
Cụ thể, nếu tổng thể có trung bình μ và độ lệch chuẩn σ, thì trung bình mẫu $\bar{X}$ của một mẫu kích thước n sẽ có:

- **Trung bình của trung bình mẫu**: $E(\bar{X}) = \mu$
- **Phương sai của trung bình mẫu**: $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$
- **Độ lệch chuẩn của trung bình mẫu** (hay còn gọi là **sai số chuẩn**): $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Khi n lớn (thường là $n \geq 30$), phân phối của $\bar{X}$ có thể được xấp xỉ bằng phân phối chuẩn với trung bình μ và độ lệch chuẩn $\frac{\sigma}{\sqrt{n}}$:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

6.5 Phân phối mẫu của sự khác biệt giữa trung bình hai mẫu

Khi lấy hai mẫu độc lập từ hai tổng thể với các trung bình và độ lệch chuẩn lần lượt là μ_1, σ_1 và μ_2, σ_2, sự khác biệt giữa trung bình hai mẫu $\bar{X}_1 - \bar{X}_2$ cũng tuân theo một phân phối mẫu.
Nếu kích thước của hai mẫu lần lượt là n_1 và n_2, thì:

- **Trung bình của sự khác biệt**: $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$
- **Phương sai của sự khác biệt**: $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- **Độ lệch chuẩn của sự khác biệt**: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Với kích thước mẫu đủ lớn, phân phối của sự khác biệt giữa trung bình hai mẫu $\bar{X}_1 - \bar{X}_2$ có thể được xấp xỉ bằng phân phối chuẩn:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

7 Statistical Intervals for a Single Sample

Statistical Intervals for a Single Sample là các khoảng tin cậy được xây dựng dựa trên dữ liệu từ một mẫu duy nhất, nhằm ước lượng các tham số của tổng thể như trung bình, phương sai hoặc tỷ lệ của tổng thể đó.
Các ký hiệu

- $Z_{\alpha/2}$: Giá trị tới hạn từ phân phối chuẩn chuẩn hóa ứng với mức ý nghĩa $\alpha/2$.
- $t_{\alpha/2,n-1}$: Giá trị tới hạn từ phân phối t với $n-1$ bậc tự do, ứng với mức ý nghĩa $\alpha/2$.
- $\chi^2_{\alpha/2,n-1}$: Giá trị tới hạn từ phân phối chi-bình phương với $n-1$ bậc tự do, ứng với mức ý nghĩa $\alpha/2$.
- E: Sai số cho phép mong muốn của khoảng tin cậy, xác định độ rộng của khoảng tin cậy.

7.1 Bảng tóm tắt

Nội dung	**Công thức**
Khoảng Tin Cậy cho Trung Bình của Phân Phối Chuẩn, Biết Phương Sai	Hai phía: $\bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ Một phía: Giới hạn trên: $\bar{X} + Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ Giới hạn dưới: $\bar{X} - Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$
Chọn Kích Thước Mẫu	$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E}\right)^2$, với E là sai số cho phép mong muốn của khoảng tin cậy.
Khoảng Tin Cậy cho Trung Bình của Phân Phối Chuẩn, Không Biết Phương Sai	Hai phía: $\bar{X} \pm t_{\alpha/2,\, n-1} \cdot \frac{S}{\sqrt{n}}$ Một phía: Giới hạn trên: $\bar{X} + t_{\alpha,\, n-1} \cdot \frac{S}{\sqrt{n}}$ Giới hạn dưới: $\bar{X} - t_{\alpha,\, n-1} \cdot \frac{S}{\sqrt{n}}$
Khoảng Tin Cậy cho Phương Sai và Độ Lệch Chuẩn của Phân Phối Chuẩn	Hai phía cho phương sai: $\frac{(n-1)S^2}{\chi^2_{\alpha/2,\, n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\, n-1}}$ Một phía cho phương sai: Giới hạn trên: $\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha,\, n-1}}$ Giới hạn dưới: $\sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{1-\alpha,\, n-1}}$
Khoảng Tin Cậy cho Tỷ Lệ của Tổng Thể với Mẫu Lớn	Hai phía: $\hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Một phía: Giới hạn trên: $\hat{p} + Z_{\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Giới hạn dưới: $\hat{p} - Z_{\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Chọn Kích Thước Mẫu	$n = \frac{Z^2_{\alpha/2} \cdot \hat{p}(1-\hat{p})}{E^2}$ hoặc khi chưa biết $\hat{p}$, sử dụng $n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (0.25)$, với E là sai số cho phép mong muốn của khoảng tin cậy.

8 Kiểm định giả thuyết (Hypothesis Testing)

8.1 Summary of Tests on the Mean of a normal distribution, Variance Known

Testing Hypotheses on the Mean, Variance Known (Z-Tests)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level T
$H_1 : \mu \neq \mu_0$	Probability above $\lvert Z_0\rvert$ and below $-\lvert Z_0\rvert$, $P = 2[1 - \Phi(\lvert Z_0\rvert)]$	$Z_0 > Z_{\alpha/2}$ or $Z_0 < -Z_{\alpha/2}$
$H_1 : \mu > \mu_0$	Probability above Z_0, $P = 1 - \Phi(Z_0)$	$Z_0 > Z_\alpha$
$H_1 : \mu < \mu_0$	Probability below Z_0, $P = \Phi(Z_0)$	$Z_0 < -Z_\alpha$

Note: The P-values and critical regions for these situations are shown in Figures 9.10 and 9.11.

8.2 Summary of Tests on the Mean of a normal distribution, Variance unknown

Null hypothesis:	$H_0 : \mu = \mu_0$	
Test statistic:	$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	
Alternative Hypothesis	**P-Value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	Probability above $\lvert t_0\rvert$ and below $-\lvert t_0\rvert$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	Probability above t_0	$t_0 > t_{\alpha,n-1}$
$H_1 : \mu < \mu_0$	Probability below t_0	$t_0 < -t_{\alpha,n-1}$

8.3 Summary of tests on the Variance and Standard deviation of a normal distribution

Null hypothesis:	$H_0 : \sigma^2 = \sigma_0^2$
Test statistic:	$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$
Alternative Hypothesis	**Rejection Criteria**
$H_1 : \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi_{\alpha/2,n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2,n-1}^2$
$H_1 : \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi_{\alpha,n-1}^2$
$H_1 : \sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi_{1-\alpha,n-1}^2$

8.4 Summary of Tests on Population proportion

Null hypotheses:	$H_0 : p = p_0$	
Test statistic:	$Z_0 = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$	
Alternative Hypotheses	**P-Value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : p \neq p_0$	Probability above $\lvert z_0 \rvert$ and below $-\lvert z_0 \rvert$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1 : p > p_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1 : p < p_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

9 Suy luận thống kê cho 2 mẫu (Statistical inference for two samples)

9.1 Tests on the difference in means, Variance known

Null hypothesis:	$H_0 : \mu_1 - \mu_2 = \Delta_0$	
Test statistic:	$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	
Alternative Hypotheses	**P-Value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and below $-\lvert z_0 \rvert$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

9.2 Test statistic for the Difference in Means of Two Normal Distributions, Variances Unknown and Equal

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{10-14}$$

Alternative Hypothesis	**P-Value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert t_0 \rvert$ and below $-\lvert t_0 \rvert$	$t_0 > t_{\alpha/2, n_1+n_2-2}$ or $t_0 < -t_{\alpha/2, n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha, n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha, n_1+n_2-2}$

9.3 Test statistic for the Difference in Means of Two Normal Distributions, Variances Unknown and NOT assumed Equal

If $H_0 : \mu_1 - \mu_2 = \Delta_0$ is true, the statistic

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

is distributed approximately as t with degrees of freedom given by

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

If ν is not an integer, round down to the nearest integer.

9.4 Test statistic for the difference in population proportions

$$\text{Null hypothesis:} \quad H_0 : p_1 = p_2$$

$$\text{Test statistic:} \quad Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Alternative Hypothesis	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1 : p_1 \neq p_2$	$P = 2[1 - \Phi(\lvert Z_0 \rvert)]$	$Z_0 > Z_{\alpha/2}$ or $Z_0 < -Z_{\alpha/2}$
$H_1 : p_1 > p_2$	$P = 1 - \Phi(Z_0)$	$Z_0 > Z_\alpha$
$H_1 : p_1 < p_2$	$P = \Phi(Z_0)$	$Z_0 < -Z_\alpha$

9.5 Confidence Interval on a Difference in Means, Variances known

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

9.6 Confidence Interval on a Difference in Means, Variances unknown and equal

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}$$

9.7 Confidence Interval on a Difference in Means, Variances unknown and not assumed equal

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

9.8 Confidence Interval on the Difference in Population Proportions

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

10 Hồi quy tuyến tính đơn và Sự tương quan (Simple linear regression and Correlation)

10.1 Simple Linear Regression

- Phân tích mối quan hệ giữa biến độc lập x (predictor)và biến phụ thuộc Y (response).

- Các biến được biểu diễn dưới dạng

$$Y = \beta_0 + \beta_1 x + \epsilon$$

trong đó:

  - β_0: Hệ số chặn(Intercept).
  - β_1: Hệ số góc (Slope).
  - ϵ: số hạng sai số ngẫu nhiên.

10.2 Method of least Squares

- Phương pháp tổng bình phương nhỏ nhất được dùng để ước lượng các tham số β_0 and β_1.
- Ý tưởng là cực tiểu tổng bình phương độ chệch giữa giá trị quan sát y_i và giá trị dự đoán $\hat{y}_i$:

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Parameter Estimation:

- Slope (β_1):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i x_i) - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

- Intercept (β_0):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regression Line:

- Estimated regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

10.3 Sum of squares Notations

- S_{xx}: Tổng bình phương sai lệch của x so với giá trị trung bình

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

- S_{xy}: Tổng tích sai lệch giữa x và y

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

- SSE đo lường mức độ sai khác giữa giá trị thực y_i và giá trị dự đoán $\hat{y}_i$:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Giá trị này càng nhỏ, mô hình càng tốt

10.4 Unbiased Estimator of Variance (σ^2)

- The variance of the error term is estimated as:

$$s^2 = \frac{SSE}{n-2}$$

- $n-2$ represents the degrees of freedom since two parameters β_0 and β_1 are estimated.

10.5 Properties of least squares Estimators

- **Unbiasedness**:

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

  This means the expected values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are equal to their true values.

- **Efficiency**: $\hat{\beta}_0$ and $\hat{\beta}_1$ have the smallest variance among all linear unbiased estimators.

- **Linearity**:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- **Variance of the Coefficients**:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad \text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

Coefficient of Determination (R^2):

- Measures the proportion of variance in y explained by x:

$$R^2 = 1 - \frac{SSE}{S_{yy}}$$

  where $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares of y around the mean.

10.6 Hypothesis test for the slope β_1

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_{1,0}$$

The appropriate test statistic is:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

We reject the null hypothesis if:

$$|t_0| > t_{\alpha/2, n-2}$$

10.7 Hypothesis test for the intercept β_0

$$H_0 : \beta_0 = \beta_{0,0} \quad \text{vs} \quad H_1 : \beta_0 \neq \beta_{0,0}$$

The appropriate test statistic is:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}}$$

or equivalently:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\text{se}(\hat{\beta}_0)}$$

We reject the null hypothesis if:

$$|t_0| > t_{\alpha/2,n-2}$$

10.8 Correlation Coefficient ρ

- Đo lường độ mạnh và hướng của mối quan hệ tuyến tính giữa hai biến ngẫu nhiên X và Y.
- Công thức:

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

  Trong đó:

  - $\text{Cov}(X,Y)$: Hiệp phương sai của X và Y.
  - σ_X, σ_Y: Độ lệch chuẩn của X và Y.

- Tính chất:

$$-1 \leq \rho \leq 1$$

  $\rho = 1$ (tương quan dương hoàn hảo), $\rho = -1$ (tương quan âm hoàn hảo), $\rho = 0$ (không có tương q

10.9 Sample correlation coefficient

- Công thức:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  Trong đó:

  - x_i, y_i: Các quan sát của X và Y.
  - $\bar{x}, \bar{y}$: Trung bình mẫu của X và Y.

10.10 Hypothesis test for Correlation Coefficient

- Giả thuyết:

$$H_0 : \rho = 0 \quad \text{(không có mối quan hệ tuyến tính)} \quad \text{vs} \quad H_1 : \rho \neq 0$$

- Thống kê kiểm định:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

  Trong đó r là hệ số tương quan mẫu và n là kích thước mẫu.

- Quy tắc quyết định:

$$\text{Bác bỏ } H_0 \text{ nếu } |t_0| > t_{\alpha/2,n-2}.$$