Chapter 17

Probability & Statistics

Probability and statistics provide the mathematical foundation for modeling uncertainty, analyzing data, and making quantitative decisions under incomplete information. Probability deals with the theoretical laws governing random phenomena, while statistics concerns the extraction of information from data using probabilistic models.

17.1 Basic Concepts of Probability

17.1.1 Random Experiments and Sample Space

A **random experiment** is a process whose outcome cannot be predicted with certainty. The set of all possible outcomes of a random experiment is called the **sample space** and is denoted by Ω. Any subset $A \subseteq \Omega$ is called an **event**.

17.1.2 Axioms of Probability

A probability measure $P(A)$ satisfies the following axioms:

1. $0 \leq P(A) \leq 1$ for any event A,
2. $P(\Omega) = 1$,
3. For mutually exclusive events A_i,

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

17.1.3 Conditional Probability and Bayes' Theorem

The **conditional probability** of A given B is

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

Bayes' theorem is

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

17.2 Random Variables and Probability Distributions

17.2.1 Discrete Random Variables

A **discrete random variable** takes countable values x_i with **probability mass function (PMF)**

$$P(X = x_i) = p(x_i).$$

17.2.2 Continuous Random Variables

A **continuous random variable** has a probability density function (PDF) $f(x)$ such that

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

17.2.3 Cumulative Distribution Function

The CDF is defined as

$$F(x) = P(X \le x).$$

17.3 Mathematical Expectation and Moments

17.3.1 Mean and Variance

The **expectation** of a random variable is

$$E[X] = \sum_{i=1}^{n} x_i p(x_i), \quad \text{(discrete)}$$

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx. \quad \text{(continuous)}$$

The **variance** is

$$\operatorname{Var}(X) = E[(X - E[X])^2].$$

17.3.2 Higher-Order Moments

The n-th moment is

$$\mu_n = E[X^n].$$

17.4 Standard Probability Distributions

Many random phenomena encountered in science and engineering can be modeled using a small collection of fundamental probability distributions. These distributions characterize the statistical behavior of discrete and continuous random variables.

17.4.1 Discrete Distributions

17.4.1.1 Bernoulli Distribution

A random variable X is said to follow a **Bernoulli distribution** with parameter p, denoted by $X \sim \text{Bern}(p)$, if it takes the value

$$X = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1-p. \end{cases}$$

The probability mass function (PMF) is

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}.$$

The mean and variance are

$$E[X] = p, \qquad \text{Var}(X) = p(1-p).$$

This distribution models a **single trial** with two possible outcomes, such as success or failure.

17.4.1.2 Binomial Distribution

A random variable X follows a **Binomial distribution** with parameters n and p, denoted by $X \sim \text{Bin}(n, p)$, if it represents the number of successes in n independent Bernoulli trials. The PMF is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

The mean and variance are

$$E[X] = np, \qquad \text{Var}(X) = np(1-p).$$

This distribution is widely used in quality control, reliability analysis, and sampling theory.

17.4.1.3 Poisson Distribution

A random variable X has a **Poisson distribution** with rate parameter $\lambda > 0$, denoted by $X \sim \text{Poisson}(\lambda)$, if

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

The mean and variance are both equal to

$$E[X] = \lambda, \qquad \text{Var}(X) = \lambda.$$

The Poisson distribution models the **number of events occurring in a fixed interval of time or space**, such as radioactive decay or arrival of customers.

17.4.1.4 Geometric Distribution

A random variable X follows a **Geometric distribution** with parameter p, denoted by $X \sim \text{Geom}(p)$, if it represents the number of trials needed to obtain the first success. The PMF is

$$P(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \ldots$$

The mean and variance are

$$E[X] = \frac{1}{p}, \qquad \text{Var}(X) = \frac{1-p}{p^2}.$$

This distribution exhibits the **memoryless property**.

17.4.2 Continuous Distributions

17.4.2.1 Uniform Distribution

A random variable X follows a **Uniform distribution** on the interval $[a, b]$, denoted by $X \sim U(a, b)$, if the PDF is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

The mean and variance are

$$E[X] = \frac{a+b}{2}, \qquad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

All values in the interval $[a, b]$ are equally likely.

17.4.2.2 Exponential Distribution

A random variable X follows an **Exponential distribution** with parameter $\lambda > 0$, denoted by $X \sim \text{Exp}(\lambda)$, if the PDF is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & \text{otherwise.} \end{cases}$$

The mean and variance are

$$E[X] = \frac{1}{\lambda}, \qquad \text{Var}(X) = \frac{1}{\lambda^2}.$$

This distribution models **waiting times** between successive random events and also exhibits the **memoryless property**.

17.4.2.3 Normal (Gaussian) Distribution

A random variable X follows a **Normal distribution** with mean μ and variance σ^2, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$, if the PDF is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The mean and variance are

$$E[X] = \mu, \qquad \text{Var}(X) = \sigma^2.$$

It plays a central role in probability theory due to the **Central Limit Theorem**.

17.4.2.4 Gamma Distribution

A random variable X follows a **Gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$, denoted by $X \sim \Gamma(\alpha, \beta)$, if the PDF is

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The mean and variance are

$$E[X] = \alpha\beta, \qquad \mathrm{Var}(X) = \alpha\beta^2.$$

The Gamma distribution generalizes both the Exponential distribution and the Erlang distribution and is widely used in queueing theory and reliability analysis.

17.5 Joint Distributions and Independence

17.5.1 Joint Probability Functions

For two continuous random variables X and Y, the joint probability density function (PDF) is denoted by $f_{X,Y}(x, y)$.

17.5.2 Independence

The random variables X and Y are said to be **independent** if their joint PDF factors as

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

17.5.3 Covariance and Correlation

The **covariance** between X and Y is defined as

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])].$$

The **correlation coefficient** is defined as

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

17.6 Law of Large Numbers and Central Limit Theorem

17.6.1 Law of Large Numbers

The sample mean converges to the population mean:

$$\bar{X}_n \to E[X] \quad \text{as } n \to \infty.$$

17.6.2 Central Limit Theorem

If $X_1, \ldots, X_n$ are independent with mean μ and variance σ^2, then

$$\frac{\sum X_i - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1).$$

$\mathcal{N}(0,1)$ is the standard normal distribution with mean 0 and variance 1.

17.7 Statistical Data Analysis

Statistical data analysis is concerned with the collection, organization, presentation, interpretation, and inference of data. It provides quantitative tools for extracting meaningful information from raw observations and for making decisions under uncertainty.

17.7.1 Types of Data

Data may be classified into the following categories:

- **Qualitative (Categorical) Data:** Non-numerical data describing attributes such as color, gender, or category.
- **Quantitative Data:** Numerical data representing measurable quantities.

Quantitative data are further classified as:

- **Discrete Data:** Data taking countable values (e.g., number of defects).
- **Continuous Data:** Data taking values on a continuum (e.g., height, temperature).

17.7.2 Methods of Data Collection

Common methods of data collection include:

- Direct measurement and experimentation
- Surveys and questionnaires
- Observational studies
- Automated data acquisition systems

The reliability of any statistical analysis depends critically on the quality and representativeness of the collected data.

17.7.3 Organization of Data

Raw data are often organized using systematic tabular arrangements in order to simplify analysis and interpretation.

17.7.3.1 Frequency Tables

A **frequency table** lists each distinct data value along with the number of times (frequency) it occurs in the dataset. If the observed data values are $x_1, x_2, \ldots, x_k$ with corresponding frequencies $f_1, f_2, \ldots, f_k$,

then

$$\sum_{i=1}^{k} f_i = n$$

where n is the total number of observations.

17.7.3.2 Grouped Data Tables

When the dataset is large or continuous, the data are grouped into class intervals. A grouped data table consists of:

- Class intervals
- Corresponding class frequencies

Each class interval is usually of equal width, and grouped data allow approximate computation of summary statistics.

17.7.3.3 Cumulative Frequency Distributions

A **cumulative frequency distribution** shows the total number of observations less than or equal to a given value. If F_i denotes cumulative frequency up to the i-th class, then

$$F_i = \sum_{j=1}^{i} f_j$$

These distributions are useful for determining percentiles, quartiles, and medians.

17.7.4 Graphical Methods of Data Presentation

Graphical representation provides visual insight into the nature of the data. The main graphical methods include:

- **Histograms**
- **Bar Charts**
- **Pie Charts**
- **Frequency Polygons**
- **Ogives (Cumulative Frequency Curves)**

These graphs allow rapid visual interpretation of the distribution, spread, symmetry, skewness, and outliers.

17.7.5 Descriptive Statistics

Descriptive statistics summarize large datasets using a small number of numerical indices.

17.7.5.1 Measures of Central Tendency

- **Arithmetic Mean:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median:** The middle value of an ordered dataset
- **Mode:** The most frequently occurring value

17.7.6 Measures of Dispersion

Measures of dispersion quantify the spread of the data.

- **Range:**

$$\text{Range} = x_{\max} - x_{\min}$$

- **Variance:**

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

- **Standard Deviation:**

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

17.7.7 Skewness and Kurtosis

Skewness measures the asymmetry of a distribution:

$$\gamma_1 = \frac{1}{n\sigma^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Kurtosis measures the peakedness of a distribution:

$$\gamma_2 = \frac{1}{n\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3$$

17.7.8 Correlation Analysis

Correlation measures the strength and direction of the relationship between two variables. The **Pearson correlation coefficient** is defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where $-1 \leq r \leq 1$.

17.7.9 Regression Analysis

Regression analysis models the functional relationship between a dependent variable y and an independent variable x. The **simple linear regression model** is

$$y = a + bx$$

where a and b are estimated using the least squares method.

17.7.10 Exploratory Data Analysis

Exploratory Data Analysis (EDA) summarizes datasets using visual and numerical techniques to:

- Discover patterns and trends
- Detect outliers
- Check statistical assumptions
- Assess distributional properties

17.7.11 Applications of Statistical Data Analysis

Statistical data analysis is widely used in:

- Engineering and quality control
- Scientific experimentation
- Economics and financial forecasting
- Medical research and clinical trials
- Machine learning and artificial intelligence

17.8 Estimation Theory

Estimation theory deals with the use of sample data to infer the numerical values of unknown population parameters. Since complete population information is rarely available, statistical inference relies on carefully constructed estimators and confidence intervals to quantify uncertainty in parameter estimation.

17.8.1 Point Estimation

A statistic $\hat{\theta}$ is called a **point estimator** of an unknown population parameter θ if it provides a single numerical estimate of that parameter. Once the estimator is evaluated using sample data, it yields a **point estimate**.

Common point estimators include:

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma}^2 = s^2$$

17.8.1.1 Properties of Good Estimators

A good estimator should satisfy the following properties:

- **Unbiasedness:** An estimator $\hat{\theta}$ is unbiased if

$$E[\hat{\theta}] = \theta$$

- **Consistency:** An estimator is consistent if

$$\hat{\theta} \to \theta \quad \text{as } n \to \infty$$

- **Efficiency (Minimum Variance):** Among all unbiased estimators, the one with the smallest variance is said to be the most efficient.
- **Sufficiency:** A statistic is sufficient if it contains all the information in the sample about the parameter.

17.8.2 Interval Estimation

While point estimation provides a single value, it does not convey the uncertainty associated with the estimate. **Interval estimation** overcomes this limitation by providing a range of values within which the true parameter is expected to lie with a specified level of confidence.

A **confidence interval** for a parameter θ is written as

$$P(\theta_1 \leq \theta \leq \theta_2) = 1 - \alpha$$

where $1 - \alpha$ is called the **confidence level**, and α is the **level of significance**.

17.8.2.1 Confidence Interval for the Population Mean (Known Variance)

If the population variance σ^2 is known and the population is normally distributed, or the sample size is large, then the confidence interval for the population mean μ is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

17.8.2.2 Confidence Interval for the Population Mean (Unknown Variance)

If the population variance is unknown and the sample size is small, the confidence interval is based on the t-distribution:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

where s is the sample standard deviation and $t_{\alpha/2,n-1}$ is the critical value from the t-distribution with $n-1$ degrees of freedom.

17.8.2.3 Confidence Interval for the Population Variance

For a normally distributed population, the confidence interval for the variance σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are chi-square critical values with $n-1$ degrees of freedom.

17.8.3 Sample Size Determination

The required sample size for estimating the population mean with an error bound E and confidence level $1-\alpha$ is given by

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

This ensures that the margin of error in the estimate does not exceed E.

17.8.4 Applications of Estimation Theory

Estimation theory is fundamental in:

- Quality control and industrial inspection
- Parameter estimation in physical and engineering models
- Financial risk estimation
- Medical statistics and clinical trials
- Machine learning model training and validation

17.9 Statistical Hypothesis Testing

Statistical hypothesis testing is a formal inferential procedure used to make decisions about population parameters on the basis of sample data. It provides a structured framework for testing scientific claims under uncertainty using probability theory.

17.9.1 Null and Alternative Hypotheses

A **statistical hypothesis** is a statement concerning the value of a population parameter. Two mutually exclusive hypotheses are formulated:

- **Null Hypothesis** (H_0): The default assumption that no significant effect or difference exists.
- **Alternative Hypothesis** (H_1): A competing claim that contradicts the null hypothesis.

Typical forms of hypotheses are:

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0 \quad \text{(two-tailed test)}$$

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0 \quad \text{(right-tailed test)}$$

$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0 \quad \text{(left-tailed test)}$$

17.9.2 Level of Significance and Errors

The **level of significance** α is the maximum allowable probability of rejecting the null hypothesis when it is actually true. Common choices are

$$\alpha = 0.10, \quad 0.05, \quad 0.01$$

Two types of errors may occur:

- **Type I Error**: Rejecting H_0 when it is true (probability = α)
- **Type II Error**: Failing to reject H_0 when it is false (probability = β)

The **power of a test** is defined as

$$\text{Power} = 1 - \beta$$

It measures the ability of the test to correctly detect a false null hypothesis.

17.9.3 Test Statistics and Decision Rules

A **test statistic** is a numerical function of the sample data whose sampling distribution is known under the null hypothesis. Decisions are made using:

- The **critical region** approach
- The **p-value** approach

Decision rules:

- Reject H_0 if the test statistic lies in the critical (rejection) region
- Reject H_0 if p-value $\leq \alpha$

17.9.4 Common Statistical Tests

17.9.4.1 z-Test

The z-test is used for testing hypotheses about the population mean when the population variance is known and the sample size is large. The test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Under H_0, $z \sim \mathcal{N}(0, 1)$

17.9.4.2 t-Test

The t-test is used when the population variance is unknown and the sample size is small. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Under H_0, t follows Student's t-distribution with $n - 1$ degrees of freedom

17.9.4.3 χ^2-Test

The chi-square test is used for:

- testing population variance
- goodness-of-fit tests
- tests of independence in contingency tables

The test statistic is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

17.9.4.4 *F*-Test

The *F*-test is used to compare two population variances and in analysis of variance (ANOVA). The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

17.9.5 Procedure for Hypothesis Testing

The standard steps in hypothesis testing are:

1. Formulate the null and alternative hypotheses
2. Select the level of significance α
3. Choose the appropriate test statistic
4. Determine the critical region or compute the p-value
5. Make the decision and draw conclusions

17.9.6 Interpretation of Results

Rejecting H_0 does not prove that H_1 is absolutely true, nor does failing to reject H_0 establish its absolute truth. All statistical conclusions are **probabilistic** and subject to sampling uncertainty.

17.10 Regression and Correlation Analysis

Regression and correlation analysis quantify the strength, direction, and functional form of relationships between variables.

17.10.1 Simple Linear Regression

The **simple linear regression model** is

$$y = a + bx$$

where a is the intercept and b is the regression coefficient.

17.10.2 Least Squares Method

The least squares estimates of a and b satisfy the normal equations:

$$\sum y = na + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

17.11 Stochastic Processes (Introduction)

A **stochastic process** is a collection of random variables indexed by time. Common examples include:

- Markov Processes
- Poisson Processes
- Random Walks

17.12 Applications of Probability and Statistics

Probability and statistics play a vital role in numerous scientific and engineering fields:

- Engineering and Signal Processing
- Finance and Risk Analysis
- Physics and Quantum Mechanics
- Machine Learning and Data Science