Descriptive Statistics Summary

Measures of Central Tendency

Mean (Arithmetic Average)

$$\bar{x} = \frac{\sum x_i}{n} \tag{1}$$

Assumptions:

- Data must be numerical.
- Sensitive to outliers.
- Best used when data is symmetrically distributed.

Median (Middle Value)

Calculation: If n is odd, take the middle value; if n is even, take the average of the two middle values.
Assumptions:

- Resistant to outliers.
- Preferred when data is skewed.

Mode (Most Frequent Value)

Assumptions:

- Applicable to both categorical and numerical data.
- Can be unimodal, bimodal, or multimodal.
- Useful for identifying the most common category or value.

Measures of Dispersion

Range

$$\text{Range} = \max(x) - \min(x) \tag{2}$$

Use when: A quick measure of spread, but highly sensitive to outliers.

Variance

Population Variance:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{3}$$

Sample Variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{4}$$

Assumptions:

- Measures data spread from the mean.
- Best used when data follows a normal distribution.

Standard Deviation

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2} \tag{5}$$

Use when: You need dispersion in the same unit as the data.

Interquartile Range (IQR)

$$\text{IQR} = Q_3 - Q_1 \tag{6}$$

Use when: Data is skewed, and you want to ignore outliers.

Shape of the Distribution

Skewness (Measure of Asymmetry)

$$\text{Skewness} = \frac{\sum(x_i - \bar{x})^3/n}{s^3} \tag{7}$$

Use when: Checking whether data is symmetric.

- Positive skew: Right tail longer.
- Negative skew: Left tail longer.
- Zero skew: Symmetric distribution.

Kurtosis (Measure of Tail Heaviness)

$$\text{Kurtosis} = \frac{\sum(x_i - \bar{x})^4/n}{s^4} - 3 \tag{8}$$

Use when: Checking for extreme values (outliers).

- Leptokurtic (> 0): Heavy tails.
- Mesokurtic (≈ 0): Normal-like tails.
- Platykurtic (< 0): Light tails.

Relationship Between Variables

Covariance

$$\text{Cov}(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{9}$$

Use when: Checking the direction of relationship (positive or negative).

Correlation (Pearson's r)

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y} \tag{10}$$

Use when: Measuring the linear relationship between two variables (-1 to 1 scale).

Probability Theory

Basic Probability Rules

Addition Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{11}$$

Explanation: Used to find the probability of either event A or B occurring.
Assumptions:

- Events must be properly defined.
- Probabilities must be known.

When to use: When dealing with the probability of the union of events.

Multiplication Rule

$$P(A \cap B) = P(A)P(B|A) \tag{12}$$

Explanation: Used to find the probability of both events occurring.
Assumptions:

- Events should be dependent or independent as required.

When to use: When calculating joint probabilities.

Conditional Probability and Bayes' Theorem

Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{13}$$

Explanation: This represents the probability of event A occurring given that event B has already occurred. It quantifies how our belief about A changes when we know B has happened.
Assumptions: $P(B) > 0$
When to Use: Use conditional probability when the occurrence of one event (e.g., B) affects the likelihood of another event (e.g., A).
Example: Let A be the event "a person has a fever", and B be the event "a person has the flu". Suppose:

$$P(A \cap B) = 0.08 \quad \text{(probability a person has both fever and flu)}$$

$$P(B) = 0.10 \quad \text{(probability a person has the flu)}$$

Then:

$$P(A \mid B) = \frac{0.08}{0.10} = 0.8$$

So, if a person has the flu, there is an 80% chance they also have a fever.

Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{14}$$

Explanation: Bayes' Theorem allows us to update the probability of a hypothesis A given new evidence B. It connects the prior probability $P(A)$, the likelihood $P(B \mid A)$, and the marginal probability $P(B)$ to compute the posterior probability $P(A \mid B)$.

Assumptions:

- $P(B) > 0$
- All probabilities involved are well-defined and correctly conditioned

When to Use: Use Bayes' Theorem when you want to revise existing beliefs in light of new data or evidence. Common in probabilistic models, spam filtering, diagnostics, and decision-making under uncertainty.

Example: Let A be the event "email is spam", and B be the event "email contains the word 'win'". Suppose:

$$P(A) = 0.2 \quad \text{(prior probability of spam)}$$

$$P(B \mid A) = 0.7 \quad \text{(likelihood of 'win' appearing in spam)}$$

$$P(B) = 0.3 \quad \text{(overall probability of 'win' appearing)}$$

Then:

$$P(A \mid B) = \frac{0.7 \times 0.2}{0.3} = \frac{0.14}{0.3} \approx 0.467$$

So, if an email contains the word "win", there's about a 46.7% chance it's spam.

KL-Divergence (Kullback-Leibler Divergence)

Definition: KL-Divergence measures how one probability distribution Q diverges from a second, expected probability distribution P.

- **Discrete form:**

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- **Continuous form:**

$$D_{\mathrm{KL}}(P \parallel Q) = \int P(x) \log \frac{P(x)}{Q(x)}\, dx$$

Interpretation:

- Measures the information loss when using distribution Q to approximate P.
- Always non-negative: $D_{\mathrm{KL}}(P \parallel Q) \geq 0$, equality holds if $P = Q$.
- Not symmetric: $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$.

How to Interpret Values:

- **KL $\approx$ 0:** The two distributions are very similar. (Good fit.)
- **Small KL:** Low information loss when using Q instead of P. (Acceptable fit.)
- **Large KL:** High divergence — Q is a poor approximation of P. (Bad fit.)
- KL values are **relative** — what is "large" depends on the problem domain.

Real-world Analogy: Suppose you're compressing English text with a model trained on French; KL-divergence quantifies the inefficiency in doing so.

Worked Example (Discrete):
Given:

$$P(x) = [0.5, 0.4, 0.1], \quad Q(x) = [0.3, 0.4, 0.3]$$

$$D_{\mathrm{KL}}(P \parallel Q) = 0.5 \log \frac{0.5}{0.3} + 0.4 \log \frac{0.4}{0.4} + 0.1 \log \frac{0.1}{0.3}$$

$$= 0.5 \cdot \log(1.667) + 0 + 0.1 \cdot \log(0.333) \approx 0.5 \cdot 0.737 - 0.1 \cdot 1.098 \approx 0.259$$

Worked Example (Continuous - Gaussian Distributions):
Let:

$$P(x) = \mathcal{N}(0, 1), \quad Q(x) = \mathcal{N}(0, 2)$$

KL divergence between two normal distributions:

$$D_{\mathrm{KL}}(P \parallel Q) = \log\left(\frac{\sigma_Q}{\sigma_P}\right) + \frac{\sigma_P^2 + (\mu_P - \mu_Q)^2}{2\sigma_Q^2} - \frac{1}{2}$$

Substitute values:

$$\mu_P = \mu_Q = 0, \quad \sigma_P = 1, \quad \sigma_Q = \sqrt{2}$$

$$D_{\mathrm{KL}} = \log(\sqrt{2}) + \frac{1}{2 \cdot 2} - \frac{1}{2} = 0.5 \log(2) + 0.25 - 0.5 \approx 0.3466 + 0.25 - 0.5 = 0.0966$$

Tip: Use Jensen's Inequality to prove the non-negativity: $D_{\mathrm{KL}}(P \parallel Q) \geq 0$.

Expected Value and Variance of Random Variables

Expected Value (Mean)

$$E[X] = \sum x \cdot P(X = x) \tag{15}$$

Explanation: The expected value (or mean) of a random variable X represents the long-run average outcome if an experiment is repeated many times. It provides a measure of central tendency for the distribution of X.

Assumptions:

- The probability distribution of X is well-defined.
- The sum of all probabilities equals 1.

When to Use: Use expected value for making decisions under uncertainty, evaluating long-term returns, and in forecasting models.

Example: Suppose X is the outcome when rolling a fair six-sided die:

$$E[X] = \sum_{x=1}^{6} x \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3.5$$

So, the expected outcome of a fair die roll is 3.5.

Variance

$$\text{Var}(X) = E[(X - E[X])^2] = \sum (x - E[X])^2 \cdot P(X = x) \tag{16}$$

Explanation: Variance measures how much the values of a random variable deviate from the expected value. A higher variance indicates more spread or variability in the outcomes.

Assumptions:

- The expected value $E[X]$ exists and is finite.
- The squared deviations are summable with their probabilities.

When to Use: Use variance to understand the consistency and uncertainty of outcomes, and to compare the spread of different distributions.

Example (continued): For a fair die ($E[X] = 3.5$):

$$\text{Var}(X) = \sum_{x=1}^{6} (x - 3.5)^2 \cdot \frac{1}{6} = \frac{(1-3.5)^2 + (2-3.5)^2 + \cdots + (6-3.5)^2}{6} = \frac{17.5}{6} \approx 2.92$$

Probability Distributions (Discrete & Continuous)

Discrete Probability Distributions

Bernoulli Distribution

$$P(X = x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\} \tag{17}$$

Explanation: The Bernoulli distribution models a random experiment that has exactly two possible outcomes:

- $x = 1$ represents a success (with probability p)
- $x = 0$ represents a failure (with probability $1 - p$)

This is the simplest discrete probability distribution and is a building block for more complex distributions like the Binomial.

Assumptions:

- The outcome of the trial is binary (only success or failure).

- The probability of success p is constant and lies in the range $0 \leq p \leq 1$.
- The trial is independent of other trials.

When to Use: Use the Bernoulli distribution when modeling binary processes such as:

- A coin toss (Heads = 1, Tails = 0)
- A test result (Positive = 1, Negative = 0)
- A customer action (Clicked ad = 1, Didn't click = 0)

Example: Suppose you flip a biased coin that lands heads with probability $p = 0.7$. Then:

$$P(X = 1) = 0.7 \quad \text{(heads)}, \qquad P(X = 0) = 0.3 \quad \text{(tails)}$$

So, $X \sim \text{Bernoulli}(0.7)$ represents this experiment.

Binomial Distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{18}$$

Explanation: The binomial distribution models the probability of getting exactly k successes in n independent trials, where each trial has two possible outcomes (success or failure), and the probability of success is p. The term $\binom{n}{k}$ represents the number of ways to choose k successes out of n trials.

Assumptions:

- There are a fixed number of trials n.
- Each trial is independent of the others.
- Each trial has the same probability of success p.
- The outcome of each trial is binary (success or failure).

When to Use: Use the binomial distribution when you want to model the total number of successes in a fixed number of repeated, independent Bernoulli trials. Common applications include quality control, medical testing, and survey analysis.

Example: Suppose a factory produces light bulbs and the probability that a bulb is defective is $p = 0.1$. If you randomly select $n = 5$ bulbs, what is the probability that exactly $k = 2$ are defective?

$$P(X = 2) = \binom{5}{2} (0.1)^2 (0.9)^3 = 10 \times 0.01 \times 0.729 = 0.0729$$

So, there's approximately a 7.29% chance of getting exactly 2 defective bulbs out of 5.

Poisson Distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{19}$$

Explanation: The Poisson distribution models the probability of observing exactly k events in a fixed interval of time or space, assuming the events happen independently and at a constant average rate λ. It is often used for modeling the frequency of rare or random events.

Assumptions:

- Events occur one at a time and independently of each other.
- The average rate of occurrence λ is constant over the observed interval.
- Two events cannot occur at exactly the same instant (events are discrete).

When to Use: Use the Poisson distribution when modeling the number of times an event happens in a specific time or space interval, especially when the events are rare and randomly occurring.

Example: Suppose a call center receives an average of 4 calls per hour. What is the probability of receiving exactly 2 calls in an hour?

$$P(X = 2) = \frac{4^2 \cdot e^{-4}}{2!} = \frac{16 \cdot e^{-4}}{2} \approx \frac{16 \cdot 0.0183}{2} \approx 0.146$$

So, there's approximately a 14.6% chance of receiving exactly 2 calls in an hour.

Continuous Probability Distributions

Uniform Distribution

$$f(x) = \frac{1}{b-a}, \quad \text{for } a \leq x \leq b \tag{20}$$

Explanation: The continuous uniform distribution assumes that all values within the interval $[a, b]$ are equally likely to occur. The probability density is constant, meaning there is no preference for any particular subinterval within $[a, b]$.

Assumptions:

- The variable x lies within the interval $[a, b]$.
- Every value in the interval has the same likelihood.
- There is no bias toward any subregion in the interval.

When to Use: Used when modeling completely random events within a known range. Common in simulations and randomized algorithms.

Example: Suppose a random number generator picks a real number between 2 and 8. What is the probability that the number falls between 3 and 5?

Since the distribution is uniform:

$$f(x) = \frac{1}{8-2} = \frac{1}{6}$$

The probability of falling between 3 and 5 is the area under the PDF between those limits:

$$P(3 \leq X \leq 5) = (5-3) \cdot \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

So, there's a one-third chance of selecting a number between 3 and 5.

Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{21}$$

Explanation: The normal distribution is a continuous probability distribution characterized by its symmetric, bell-shaped curve. It describes how values of a variable are distributed around the mean (μ), with spread determined by the standard deviation (σ). Approximately 68% of values lie within 1 standard deviation of the mean, 95% within 2, and 99.7% within 3 (empirical rule).

Assumptions:

- The data is symmetrically distributed about the mean.
- No significant skewness or heavy tails (i.e., light tails).
- The underlying phenomenon is influenced by many small, independent effects (central limit theorem).

When to Use: Used for modeling continuous variables like height, blood pressure, IQ scores, test scores, and errors in measurements when the data appears symmetric and bell-shaped.

Example: Suppose the heights of adult males are normally distributed with a mean of 175 cm and a standard deviation of 10 cm. What is the probability that a randomly selected male is between 165 cm and 185 cm?

Using the empirical rule:

$$165 = \mu - \sigma, \quad 185 = \mu + \sigma$$

So, about 68% of values fall within this range.

$$P(165 \leq X \leq 185) \approx 0.68$$

Therefore, there's approximately a 68% chance that a randomly selected individual is between 165 and 185 cm tall.

Exponential Distribution

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0 \tag{22}$$

Explanation: The exponential distribution is a continuous probability distribution used to model the time between events in a Poisson process. It is memoryless, meaning the probability of an event occurring in the next interval is independent of how much time has already elapsed.

Assumptions:

- Events occur independently of each other.
- Events occur at a constant average rate $\lambda > 0$.
- The time between events is continuous and non-negative.

When to Use: Used to model the time or space between random events, such as:

- Time until the next system failure.
- Time between customer arrivals at a service point.
- Duration between phone calls at a call center.

Example: If calls arrive at a call center at an average rate of 3 per hour ($\lambda = 3$), what is the probability that the next call will come in more than 30 minutes from now?

$$P(X > 0.5) = \int_{0.5}^{\infty} 3e^{-3x} dx = e^{-3 \cdot 0.5} = e^{-1.5} \approx 0.2231$$

So, there is approximately a 22.3% chance the next call will come after 30 minutes.

Continuous Probability Distributions

Normal Distribution & Standardization (Z-Scores)

$$Z = \frac{X - \mu}{\sigma} \tag{23}$$

Explanation: Converts a normal distribution into a standard normal distribution (mean 0, variance 1).
Assumptions:

- Data follows a normal distribution.
- Mean (μ) and standard deviation (σ) are known.

When to use: Comparing different normal distributions and hypothesis testing.

Exponential & Gamma Distributions

$$\text{Exponential: } f(x) = \lambda e^{-\lambda x}, \quad x \geq 0 \tag{24}$$

Explanation: Models the time between independent events in a Poisson process.
Assumptions:

- Events occur independently.
- The rate parameter (λ) is constant.

When to use: Modeling wait times like system failures or service rates.

$$\text{Gamma: } f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}, \quad x > 0 \tag{25}$$

Explanation: Generalization of the exponential distribution for modeling the sum of multiple independent exponential variables.
Assumptions: Similar to the exponential distribution.
When to use: Modeling the time until k events occur.

Central Limit Theorem (CLT)

Statement: As the sample size n becomes large, the sampling distribution of the sample mean $\bar{X}$ approaches a normal distribution, regardless of the shape of the population distribution.

Formula:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{26}$$

where μ is the population mean and σ^2 is the population variance.

Explanation: The CLT justifies why many statistical procedures assume normality. It ensures that even if the population distribution is not normal, the distribution of the sample mean will be approximately normal for large enough samples.

Assumptions:

- Samples are drawn independently and identically (i.i.d).
- The sample size is sufficiently large (commonly $n \geq 30$).
- Population variance σ^2 is finite.

When to Use: Used to justify:

- Confidence interval estimation for means.
- Hypothesis testing involving means.
- Normal approximations of sampling distributions.

Example: Suppose the time spent on a website is right-skewed with a mean $\mu = 10$ minutes and standard deviation $\sigma = 5$ minutes. If we take a random sample of $n = 50$ users, by the CLT:

$$\bar{X} \sim \mathcal{N}\left(10, \frac{5^2}{50}\right) = \mathcal{N}(10, 0.5)$$

So, the sample mean will be approximately normally distributed with mean 10 and variance 0.5, despite the original distribution being skewed.

Statistical Inference

Explanation: The process of drawing conclusions about a population based on a sample.
Assumptions:

- The sample is representative of the population.
- Data follows a well-defined probability distribution.

When to use: Making predictions and testing hypotheses.

Sampling Techniques

Random Sampling

Explanation: Each member of the population has an equal chance of being selected.
Assumptions:

- The population is well-defined.
- Each sample is independent.

When to use: Ensuring an unbiased sample for statistical analysis.

Stratified Sampling

Explanation: The population is divided into strata, and random samples are taken from each stratum.
Assumptions:

- The population has distinct subgroups.
- The proportion of each stratum in the sample reflects the population.

When to use: When subgroups need representation in the sample.

Cluster Sampling

Explanation: The population is divided into clusters, and entire clusters are randomly selected.
Assumptions:

- Clusters are representative of the population.
- Sampling within clusters is randomized.

When to use: When studying large, geographically spread populations.

Law of Large Numbers

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = E[X] \tag{27}$$

Explanation: As the sample size increases, the sample mean converges to the population mean.
Assumptions:

- Samples are independent and identically distributed (i.i.d.).

When to use: Ensuring reliability of sample-based estimates.

Confidence Intervals

$$CI = \bar{x} \pm Z \frac{\sigma}{\sqrt{n}} \tag{28}$$

Explanation: Provides a range of values that likely contains the true population parameter.
Assumptions:

- Data follows a normal distribution or sample size is large (CLT applies).
- Standard deviation is known or estimated.

When to use: Estimating population parameters with a degree of certainty.

Hypothesis Testing

Overview

Definition: Hypothesis testing is a statistical method used to make inferences or draw conclusions about a population based on sample data. It determines whether there is enough evidence to reject a null hypothesis.
Key Steps in Hypothesis Testing:

1. Formulate the null and alternative hypotheses.
2. Choose the significance level (α).
3. Select an appropriate test statistic.
4. Compute the test statistic and corresponding p-value.
5. Compare the p-value with α to make a decision.

Null and Alternative Hypothesis

Null Hypothesis (H_0): The default assumption that no effect or relationship exists. It represents the status quo.
Example: "The new drug has no effect on blood pressure compared to the existing drug."

Alternative Hypothesis (H_a): The hypothesis that contradicts the null and suggests a significant effect or difference.
Example: "The new drug significantly lowers blood pressure compared to the existing drug."

One-Tailed vs. Two-Tailed Tests

- **One-tailed test:** Tests for an effect in only one direction (greater or lesser).
- **Two-tailed test:** Tests for an effect in both directions (greater and lesser).

Example of One-Tailed Test: Testing if a new teaching method improves test scores.
Example of Two-Tailed Test: Testing if a new fertilizer increases or decreases crop yield.

Choosing a Significance Level (α)

The significance level represents the probability of rejecting the null hypothesis when it is true. Common values include:

- $\alpha = 0.05$: 5% chance of Type I error (most common).
- $\alpha = 0.01$: 1% chance of Type I error (more stringent).
- $\alpha = 0.10$: 10% chance of Type I error (less stringent).

Lower α values reduce the likelihood of false positives but increase the risk of false negatives.

Test Statistics

Different hypothesis tests use different test statistics, such as:

- **Z-test**: Used when the population variance is known and the sample size is large ($n \geq 30$).
- **T-test**: Used when the population variance is unknown and the sample size is small ($n < 30$).
- **Chi-square test**: Used for categorical data to test for independence.
- **ANOVA (Analysis of Variance)**: Used to compare means across multiple groups.

Decision Rule: p-Value and Critical Region

p-Value: The probability of obtaining a test statistic as extreme as, or more extreme than, the observed one under H_0.
Decision Criteria:

- If $p \leq \alpha$, reject H_0 (sufficient evidence to support H_a).
- If $p > \alpha$, fail to reject H_0 (insufficient evidence to support H_a).

Type I and Type II Errors

- **Type I Error (α)**: Rejecting a true null hypothesis (false positive).
- **Type II Error (β)**: Failing to reject a false null hypothesis (false negative).

Example of Type I Error: A medical test falsely detects a disease in a healthy patient.
Example of Type II Error: A medical test fails to detect a disease in a sick patient.

Power of a Test

$$\text{Power} = 1 - \beta \tag{29}$$

Explanation: The probability of correctly rejecting a false null hypothesis.
Factors Affecting Power:

- Sample size: Larger samples increase power.
- Effect size: Larger effects are easier to detect.
- Significance level: Higher α increases power but also Type I error risk.
- Variability: Lower variability increases power.

Parametric Tests

Overview

Definition: Parametric tests are statistical tests that assume data comes from a specific probability distribution (often normal). These tests use parameters such as mean and variance for inference.
Assumptions:

- Data follows a normal distribution (or large enough sample size for Central Limit Theorem to apply).
- Homogeneity of variance (equal variance among groups).
- Data is measured on an interval or ratio scale.

Common Parametric Tests:

- t-Tests (One-Sample, Independent, Paired)
- Analysis of Variance (ANOVA)
- F-Test

t-Tests (One-Sample, Independent, Paired)

One-Sample t-Test

Purpose: Tests whether the mean of a single sample is significantly different from a known population mean.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \tag{30}$$

Assumptions:

- The sample is randomly drawn from the population.
- Data follows a normal distribution.

Example: Testing if the average test score of students differs from the national average.

Independent t-Test (Two-Sample t-Test)

Purpose: Compares the means of two independent groups.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{31}$$

Assumptions:

- The two groups are independent.
- Data in each group follows a normal distribution.
- Homogeneity of variance (equal variance across groups).

Example: Comparing test scores between students in two different schools.

Paired t-Test

Purpose: Compares means of the same group under different conditions (dependent samples).

$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}} \tag{32}$$

where $\bar{D}$ is the mean difference and s_D is the standard deviation of the differences.
Assumptions:

- The differences between paired observations follow a normal distribution.
- Measurements are dependent (same subjects measured before and after an intervention).

Example: Testing weight loss before and after a fitness program.

Analysis of Variance (ANOVA)

One-Way ANOVA

Purpose: Compares the means of three or more independent groups.

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}} \tag{33}$$

Assumptions:

- Data in each group follows a normal distribution.
- Homogeneity of variance.
- Observations are independent.

Example: Comparing the effectiveness of three different diets on weight loss.

Two-Way ANOVA

Purpose: Examines the effect of two independent variables on a dependent variable. **Example:** Studying the impact of different teaching methods and genders on test scores.

MANOVA (Multivariate ANOVA)

Purpose: Extends ANOVA to multiple dependent variables. **Example:** Examining the impact of different diets on weight loss and cholesterol levels simultaneously.

F-Test

Purpose: Compares variances between two groups.

$$F = \frac{s_1^2}{s_2^2} \tag{34}$$

where s_1^2 and s_2^2 are the variances of the two groups.
Assumptions:

- The data follows a normal distribution.
- The two groups are independent.

Example: Testing if the variance of test scores differs between two schools.

Non-Parametric Tests

Overview

Definition: Non-parametric tests are statistical tests that do not require assumptions about the population distribution. They are useful when data is not normally distributed or when dealing with ordinal or ranked data.
Assumptions:

- Data may not follow a normal distribution.
- Can be used for ordinal, nominal, or skewed interval data.
- Some tests require independence of observations.

Common Non-Parametric Tests:

- Chi-Square Test
- Mann-Whitney U Test
- Wilcoxon Signed-Rank Test
- Kruskal-Wallis Test
- Friedman Test

Chi-Square Test

Purpose: Tests for independence between two categorical variables or goodness-of-fit to an expected distribution.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{35}$$

where O represents observed frequencies and E represents expected frequencies.
Assumptions:

- Data is categorical.
- Observations are independent.
- Expected frequency in each cell should be at least 5.

Example: Testing if gender and preference for a product are independent.

Mann-Whitney U Test

Purpose: Compares differences between two independent groups when the assumption of normality is not met.

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{36}$$

where n_1, n_2 are sample sizes and R_1 is the rank sum of the first sample.
Assumptions:

- Observations are independent.
- Data is ordinal or continuous.

Example: Comparing customer satisfaction scores between two service providers.

Wilcoxon Signed-Rank Test

Purpose: Compares two related samples to determine if their population mean ranks differ.

$$W = \sum \text{Ranks of positive differences} \tag{37}$$

Assumptions:

- Data is paired and comes from the same subjects.
- The differences between pairs are symmetrically distributed.

Example: Comparing test scores before and after a training program.

Kruskal-Wallis Test

Purpose: A non-parametric alternative to ANOVA, comparing three or more independent groups.

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \tag{38}$$

where R_i is the sum of ranks for group i, and n_i is its sample size.
Assumptions:

- Observations are independent.
- Data is ordinal or continuous.

Example: Comparing median customer ratings across different product versions.

Friedman Test

Purpose: A non-parametric alternative to repeated measures ANOVA, used for dependent samples.

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum R_j^2 - 3n(k+1) \tag{39}$$

where R_j is the sum of ranks for treatment j.
Assumptions:

- Data is ordinal or continuous.
- Observations are dependent (repeated measures on the same subjects).

Example: Comparing reaction times of the same individuals under different lighting conditions.

Correlation & Association

Overview

Definition: Correlation and association measure the strength and direction of relationships between variables. While correlation quantifies the linear relationship between numerical variables, association can refer to broader relationships, including categorical variables.
Common Measures:

- Pearson Correlation
- Spearman Correlation
- Covariance vs. Correlation
- Chi-Square Test for Independence

Pearson Correlation

Purpose: Measures the strength and direction of a linear relationship between two continuous variables.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \tag{40}$$

where X_i and Y_i are data points, and $\bar{X}$ and $\bar{Y}$ are their means.
Assumptions:

- The relationship between variables is linear.
- Both variables are continuous and normally distributed.
- Data is free of significant outliers.

Example: Analyzing the correlation between hours studied and exam scores.

Spearman Correlation

Purpose: Measures the strength and direction of a monotonic relationship between two variables using ranked data.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{41}$$

where d_i is the difference between ranks of corresponding values, and n is the number of observations.
Assumptions:

- The relationship is monotonic but not necessarily linear.
- Variables can be ordinal or continuous.
- No requirement for normality.

Example: Analyzing the correlation between customer satisfaction rankings and purchase frequency.

Covariance vs. Correlation

Covariance: Measures the direction of the relationship between two variables but is scale-dependent.

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n} \tag{42}$$

Correlation: Standardized measure of relationship strength, ranging from -1 to 1.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{43}$$

Key Differences:

- Covariance is unbounded, while correlation is between -1 and 1.
- Correlation accounts for the scale of variables, making comparisons easier.

Example: Comparing the correlation between stock returns vs. their absolute covariance values.

Chi-Square Test for Independence

Purpose: Determines if two categorical variables are independent.

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{44}$$

where O represents observed frequencies and E represents expected frequencies.
Assumptions:

- Data is categorical.
- Expected frequency in each cell should be at least 5.
- Observations are independent.

Example: Checking if customer age group influences product preference.

Regression Analysis

Overview

Definition: Regression analysis is a statistical method used to model relationships between a dependent variable and one or more independent variables. It helps in prediction, inference, and understanding the impact of predictors.
Types of Regression:

- Simple & Multiple Linear Regression
- Assumptions of Linear Regression
- Ridge & Lasso Regression
- Logistic Regression
- Generalized Linear Models (GLMs)

Simple & Multiple Linear Regression

Simple Linear Regression: Models the relationship between a single independent variable X and a dependent variable Y.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{45}$$

Multiple Linear Regression: Extends the model to multiple independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{46}$$

where β_0 is the intercept, β_i are coefficients, and ϵ is the error term.
Assumptions:

- Linearity: Relationship between predictors and response is linear.
- Independence: Observations are independent.
- Homoscedasticity: Constant variance of errors.
- Normality: Errors follow a normal distribution.
- No Multicollinearity: Predictors are not highly correlated.

Example: Predicting house prices based on square footage, number of rooms, and location.

Assumptions of Linear Regression

Key Assumptions:

- **Linearity:** The relationship between independent and dependent variables is linear.
- **No Autocorrelation:** Residuals are independent (important in time series data).
- **Homoscedasticity:** Residuals have constant variance.
- **No Multicollinearity:** Predictor variables are not highly correlated.
- **Normality of Residuals:** Residuals follow a normal distribution.

Diagnostics:

- Scatter plots to check linearity.
- Variance Inflation Factor (VIF) to detect multicollinearity.
- Residual plots to assess homoscedasticity.

Ridge & Lasso Regression

Ridge Regression: Adds an L_2 penalty to the loss function to reduce overfitting.

$$\min \sum (Y_i - \beta_0 - \sum \beta_j X_{ij})^2 + \lambda \sum \beta_j^2 \tag{47}$$

Lasso Regression: Adds an L_1 penalty, encouraging sparsity in coefficients.

$$\min \sum (Y_i - \beta_0 - \sum \beta_j X_{ij})^2 + \lambda \sum |\beta_j| \tag{48}$$

Differences:

- Ridge shrinks coefficients but does not set them to zero.
- Lasso forces some coefficients to zero, performing feature selection.

Example: Regularizing high-dimensional financial models to prevent overfitting.

Logistic Regression

Purpose: Used for binary classification problems.
Model: Instead of modeling Y directly, logistic regression models the probability using the sigmoid function.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_j X_j)}} \tag{49}$$

Assumptions:

- The dependent variable is categorical (binary or multinomial).
- No multicollinearity among independent variables.
- Observations are independent.

Example: Predicting customer churn (churn or no churn).

Generalized Linear Models (GLMs)

Purpose: Extends linear regression to handle non-normal distributions.
Model Structure:

$$g(E[Y]) = \beta_0 + \sum \beta_j X_j \tag{50}$$

where $g(\cdot)$ is a link function.
Common GLMs:

- Logistic Regression (for binary outcomes)
- Poisson Regression (for count data)
- Gamma Regression (for skewed continuous data)

Example: Modeling insurance claims using Poisson regression.

Bayesian Statistics

Bayesian Inference

Definition: Bayesian inference is a statistical method that updates the probability of a hypothesis as more evidence becomes available. It is based on Bayes' theorem:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{51}$$

where:

- $P(H|D)$ is the **posterior probability** (updated belief about hypothesis H given data D).
- $P(D|H)$ is the **likelihood** (probability of observing data D given hypothesis H).
- $P(H)$ is the **prior probability** (initial belief about hypothesis H before observing data).
- $P(D)$ is the **marginal likelihood** (overall probability of data D across all possible hypotheses).

Key Features of Bayesian Inference

- Allows the incorporation of prior knowledge into statistical modeling.
- Updates beliefs as new data arrives.
- Provides a full probability distribution rather than point estimates.

Prior, Likelihood, and Posterior Distribution

Prior Distribution: Represents the initial belief about a parameter before data is observed. Common priors include:

- **Uniform Prior:** Assumes all values are equally likely.
- **Gaussian Prior:** Used when prior knowledge suggests a normal distribution.
- **Beta Prior:** Used for probabilities (e.g., success rates in binomial models).

Likelihood Function: Represents how probable the observed data is under different values of the parameter. Given data $D = \{x_1, x_2, ..., x_n\}$, the likelihood is:

$$L(\theta) = P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) \tag{52}$$

Posterior Distribution: The updated probability distribution of the parameter after incorporating data:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \tag{53}$$

where $\propto$ indicates proportionality (since $P(D)$ is constant for a given dataset).

Interpretation of the Posterior

- The posterior distribution reflects updated knowledge after seeing data.
- The shape of the posterior depends on the prior and the data.
- A well-chosen prior combined with enough data leads to more accurate inferences.

Markov Chain Monte Carlo (MCMC)

Definition: MCMC is a class of algorithms used to approximate complex posterior distributions when exact computation is infeasible.

Why MCMC?

- Posterior distributions are often difficult to compute analytically.
- MCMC enables sampling from the posterior, allowing estimation of expectations and credible intervals.

Common MCMC Algorithms

- **Metropolis-Hastings Algorithm:** Iteratively proposes new parameter values and accepts/rejects them based on probability ratios.
- **Gibbs Sampling:** Special case of MCMC where conditional distributions are sampled sequentially.
- **Hamiltonian Monte Carlo (HMC):** Uses gradient information to make more efficient sampling moves.

Multivariate Statistics

Principal Component Analysis (PCA)

Definition: PCA is a dimensionality reduction technique used to transform correlated variables into a smaller set of uncorrelated variables called principal components.

Steps in PCA

Given a dataset with n observations and p variables, PCA follows these steps:

1. **Standardize the Data:**
   - Compute the mean μ_j and standard deviation σ_j of each variable.
   - Transform each variable X_j into a standardized form:

$$Z_j = \frac{X_j - \mu_j}{\sigma_j} \tag{54}$$

2. **Compute the Covariance Matrix:**
   - The covariance between two variables X_i and X_j is given by:

$$\text{Cov}(X_i, X_j) = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) \tag{55}$$

   - This results in a $p \times p$ covariance matrix that captures relationships between variables.

3. **Compute the Eigenvalues and Eigenvectors:**
   - Solve the characteristic equation for the covariance matrix:

$$|\Sigma - \lambda I| = 0 \tag{56}$$

   - The eigenvalues (λ) represent the variance explained by each principal component.
   - The eigenvectors define the directions (principal components) in the feature space.

4. **Select the Top k Principal Components:**
   - Rank eigenvalues in descending order and select the top k components that explain the most variance.
   - The proportion of variance explained (PVE) by each component is:

$$\text{PVE} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{57}$$

5. **Transform the Data:**
   - Project the original dataset onto the new principal component axes:

$$Z = XW \tag{58}$$

   where W is the matrix of selected eigenvectors.

Key Properties

- Principal components are orthogonal to each other.
- The first few components explain most of the variance in the data.
- Helps visualize high-dimensional data in 2D or 3D.

When to Use PCA

- When reducing dimensionality to speed up machine learning models.
- When detecting patterns in high-dimensional data.
- When removing multicollinearity in regression analysis.

Factor Analysis

Definition: Factor Analysis (FA) is a statistical method used to uncover underlying latent factors that explain the observed variables' correlations.

Types of Factor Analysis

- **Exploratory Factor Analysis (EFA):** Identifies underlying factors without prior assumptions.
- **Confirmatory Factor Analysis (CFA):** Tests hypotheses about expected factor structures.

Mathematical Model

FA assumes that observed variables X_i can be expressed as a linear combination of latent factors F_j:

$$X = LF + \epsilon \tag{59}$$

where:

- X is the observed data matrix.
- L is the factor loading matrix.
- F is the matrix of latent factors.
- ϵ is the matrix of unique variances (error terms).

When to Use Factor Analysis

- Identifying underlying constructs in survey data.
- Reducing the number of observed variables while retaining interpretability.
- Psychological and social sciences for measuring latent traits.

Canonical Correlation Analysis (CCA)

Definition: CCA is a technique used to find relationships between two sets of multivariate variables.

Mathematical Formulation

Given two sets of variables, X (with p variables) and Y (with q variables), CCA finds linear combinations:

$$U = a^T X, \quad V = b^T Y \tag{60}$$

such that the correlation between U and V is maximized.

Key Properties

- Finds the most correlated linear combinations of two datasets.
- Can be seen as a generalization of multiple regression.
- Useful in analyzing multi-view datasets (e.g., brain imaging and behavioral data).

Applications of CCA

- Linking genetic markers with clinical traits in bioinformatics.
- Analyzing relationships between different psychological tests.
- Multi-modal data fusion (e.g., image and text data in AI).

Experimental Design

A/B Testing

Definition: A/B testing is a controlled experiment where two or more versions (A and B) of a treatment are compared to determine which performs better.

Hypothesis Formulation

$$H_0 : \mu_A = \mu_B \tag{61}$$

$$H_1 : \mu_A \neq \mu_B \tag{62}$$

where μ_A and μ_B represent the mean performance of versions A and B, respectively.

Statistical Testing

The test statistic for comparing means is:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \tag{63}$$

where:

- $\bar{X}_A, \bar{X}_B$ are the sample means,
- s_A^2, s_B^2 are the sample variances,
- n_A, n_B are the sample sizes.

When to Use:

- Website optimization (e.g., UI changes, pricing strategies).
- Marketing campaigns (e.g., email subject line testing).
- Product development (e.g., new feature adoption).

Randomized Controlled Trials (RCTs)

Definition: RCTs are experiments where subjects are randomly assigned to treatment or control groups to measure the causal effect of an intervention.

Effect Estimation

The average treatment effect (ATE) is computed as:

$$\text{ATE} = E[Y_1] - E[Y_0] \tag{64}$$

where:

- Y_1 is the outcome for the treatment group,
- Y_0 is the outcome for the control group.

Statistical Testing

For large samples, hypothesis testing can be conducted using a two-sample t-test:

$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}} \tag{65}$$

When to Use:

- Medical research (e.g., testing new drugs or vaccines).
- Policy evaluation (e.g., social programs and education initiatives).
- Behavioral economics (e.g., interventions for saving habits).

Factorial Experiments

Definition: A factorial experiment is a study where multiple factors (independent variables) are varied simultaneously to assess their individual and interaction effects.

Factorial Design Model

A two-factor factorial model can be expressed as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{66}$$

where:

- μ is the overall mean,
- α_i represents the effect of factor A (e.g., ad type),
- β_j represents the effect of factor B (e.g., color),
- $(\alpha\beta)_{ij}$ represents the interaction effect,
- ϵ_{ijk} is the random error term.

Analysis Using ANOVA

Factorial experiments are analyzed using Analysis of Variance (ANOVA), where the F-statistic is given by:

$$F = \frac{\text{MS}_{\text{Factor}}}{\text{MS}_{\text{Error}}} \tag{67}$$

where:

- $\text{MS}_{\text{Factor}}$ is the mean square for the factor.
- MS_{Error} is the mean square for the residual error.

When to Use:

- Optimizing manufacturing processes.
- Studying the combined effects of drugs in medicine.
- Evaluating marketing strategies with multiple variables.

Survival Analysis

Kaplan-Meier Estimator

Definition: The Kaplan-Meier (KM) estimator is a non-parametric method used to estimate the survival function from time-to-event data.

Survival Function

The survival function, $S(t)$, represents the probability of surviving beyond time t:

$$S(t) = P(T > t) \tag{68}$$

where T is the time of an event (e.g., failure, death, dropout).

Kaplan-Meier Estimator Formula

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{69}$$

where:

- t_i are the observed event times.
- d_i is the number of events (failures) at time t_i.
- n_i is the number of individuals at risk just before t_i.

When to Use:

- Medical research (e.g., patient survival after treatment).
- Reliability engineering (e.g., failure time of machines).
- Customer retention analysis.

Cox Proportional Hazards Model

Definition: The Cox model is a semi-parametric regression model used to assess the effect of covariates on survival times.

Hazard Function

The hazard function, $h(t)$, represents the instantaneous failure rate at time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{70}$$

Cox Model Formula

$$h(t|X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \tag{71}$$

where:

- $h_0(t)$ is the baseline hazard function.
- $X_1, X_2, ..., X_p$ are covariates (e.g., age, treatment).
- $\beta_1, \beta_2, ..., \beta_p$ are the regression coefficients.

Assumptions:

- The hazard ratios remain constant over time (proportional hazards assumption).
- The covariates have a linear effect on the logarithm of the hazard.

When to Use:

- Clinical trials to compare treatment effects.
- Customer churn prediction.
- Risk assessment in finance and engineering.