Title Cheat Sheet

Relational Database Concepts

Relational Database: A relational database organizes data into tables (relations) consisting of rows and columns, where each table represents a different entity.

Primary Key: A primary key is a unique identifier for each record in a table, ensuring that no two rows have the same primary key value.

Foreign Key: A foreign key is a column or set of columns in one table that uniquely identifies a row of another table, establishing a relationship between the two tables.

Normalization: 1NF: A table is in First Normal Form (1NF) if all its columns contain atomic, indivisible values and each column contains values of a single type.

Database Relations: Relations in a database refer to the logical connections between tables, often established through primary and foreign keys to maintain data integrity.

SQL Basics

Structured Query Language (SQL): SQL is a domain-specific language used in programming and designed for managing data held in a relational database management system (RDBMS).

Data Definition Language (DDL): DDL includes SQL commands such as CREATE, ALTER, and DROP, which are used to define or modify database structures.

Data Manipulation Language (DML): DML consists of SQL commands like SELECT, INSERT, UPDATE, and DELETE, which are used to retrieve and manipulate data in a database.

SQL Statements: SQL statements are used to perform tasks such as updating data on a database, or retrieving data from a database, and include SELECT, INSERT, UPDATE, DELETE, etc.

SQL Joins: Joins are SQL operations that allow you to combine rows from two or more tables based on a related column between them, such as INNER JOIN, LEFT JOIN, RIGHT JOIN, and FULL JOIN.

SQLite

SQLite Architecture: SQLite is a file-based database engine that stores the entire database in a single file, making it lightweight and easy to deploy.

SQL92 Standard Compliance: SQLite supports most of the SQL92 standard, providing a wide range of SQL features for database management.

Atomic Commit and Rollback: SQLite ensures data integrity through atomic commit and rollback capabilities, allowing transactions to be completed fully or not at all.

Data Types in SQLite: SQLite uses dynamic typing, allowing flexibility in storing data types such as INTEGER, REAL, TEXT, BLOB, and NULL.

Concurrency Control: SQLite uses a locking mechanism to manage concurrency, allowing multiple readers or a single writer at any time.

Minimum Spanning Tree Algorithms

Minimum Spanning Tree (MST) Definition: An MST of a graph G is a subset of edges that connects all vertices with the minimum possible total edge weight.

Prim's Algorithm: Prim's algorithm starts with a single vertex and grows the MST by adding the cheapest edge from the tree to a vertex not yet in the tree.

Kruskal's Algorithm: Kruskal's algorithm builds the MST by sorting all edges and adding them one by one, ensuring no cycles are formed, until all vertices are connected.

Edge Weights in MST: Edge weights determine the selection of edges in MST algorithms, where the goal is to minimize the sum of the weights of the edges in the tree.

Cycle Prevention in Kruskal's Algorithm: Kruskal's algorithm uses a union-find data structure to efficiently check for cycles when adding edges to the MST.

Graph Theoretical Concepts

Undirected Graph: An undirected graph $G = (V, E)$ consists of a set of vertices V and a set of edges E, where each edge is an unordered pair of vertices.

Weighted Graph: In a weighted graph, each edge $(u, v) \in E$ has an associated weight $w(u, v)$, representing the cost or distance between vertices u and v.

Vertex: A vertex, also known as a node, is a fundamental part of a graph, representing an entity or a point where edges meet.

Edge: An edge in a graph is a connection between two vertices, and in an undirected graph, it is represented as an unordered pair (u, v).

Graph Connectivity: A graph is connected if there is a path between every pair of vertices; otherwise, it is disconnected.

Standard Normal Distribution

Standard Normal Distribution: The standard normal distribution, denoted as $Z \sim N(0, 1)$, is a normal distribution with a mean of 0 and a standard deviation of 1.

Probability of Z: The probability $P(Z > z)$ represents the area under the standard normal curve to the right of a given z-score.

Area Under the Curve: The area under the $N(0, 1)$ distribution curve between two points gives the probability that Z falls within that interval.

Z-Score Calculation: A z-score is calculated as $z = \frac{X-\mu}{\sigma}$, where X is a value from the dataset, μ is the mean, and σ is the standard deviation.

Applications of Standard Normal Distribution: The standard normal distribution is used to find probabilities and percentiles for normally distributed data by converting to z-scores.

Hypothesis Testing in Statistics

Null Hypothesis (H_0): The null hypothesis H_0 is a statement that there is no effect or no difference, and it is assumed true until evidence indicates otherwise.
Alternative Hypothesis (H_a): The alternative hypothesis H_a is a statement that indicates the presence of an effect or a difference, opposing the null hypothesis.
Population Parameter: A population parameter is a numerical characteristic of a population, such as a mean (μ) or standard deviation (σ), that is estimated using sample data.
Sample Mean ($\bar{x}$): The sample mean $\bar{x}$ is the average of a set of sample data, calculated as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where n is the sample size.
Significance Level (α): The significance level α is the probability of rejecting the null hypothesis when it is true, commonly set at 0.05 or 5%.

Type I and Type II Errors

Type I Error: A Type I error occurs when we reject a true null hypothesis, with probability denoted by α, the significance level.
Type II Error: A Type II error occurs when we fail to reject a false null hypothesis, with probability denoted by β.
Significance Level: The significance level α is the threshold for rejecting the null hypothesis, often set at 0.05 or 0.01.
Power of a Test: The power of a test is $1 - \beta$, representing the probability of correctly rejecting a false null hypothesis.
Trade-off Between Type I and Type II Errors: Reducing α to decrease Type I error increases β, thus increasing the chance of a Type II error.

Confidence Intervals and Rejection Regions

Confidence Intervals: A confidence interval for a parameter θ is an interval $[L, U]$ such that $P(L \leq \theta \leq U) = 1 - \alpha$, where α is the significance level.
Rejection Regions: The rejection region is the set of all values of the test statistic for which the null hypothesis H_0 is rejected in favor of the alternative hypothesis H_1.
Standard Error: The standard error (SE) is the standard deviation of the sampling distribution of a statistic, often estimated as $SE = \frac{s}{\sqrt{n}}$, where s is the sample standard deviation.
Critical Value: The critical value is the threshold value that the test statistic must exceed for the null hypothesis to be rejected, typically found from a statistical distribution table.
Significance Level: The significance level α is the probability of rejecting the null hypothesis when it is true, commonly set at 0.05 or 0.01 in hypothesis testing.

Algorithmic Complexity and Optimization

Time Complexity: Time complexity measures the amount of time an algorithm takes to complete as a function of the length of the input, commonly expressed using Big O notation, e.g., $O(n)$ for linear time.
Optimal Algorithm: An optimal algorithm is one that solves a problem in the least possible time or space complexity, often serving as a benchmark for evaluating other algorithms.
Linear Time Algorithms: An algorithm runs in linear time, $O(n)$, if the time to complete is directly proportional to the size of the input data set.
Quadratic Time Algorithms: Quadratic time complexity, $O(n^2)$, indicates that the time taken is proportional to the square of the size of the input data set, often seen in nested loop scenarios.
Algorithmic Optimization: Algorithmic optimization involves improving an algorithm to reduce its time or space complexity, often by eliminating redundant operations or using more efficient data structures.

SQL Queries

FROM Statement: The FROM statement specifies the table from which to retrieve or delete data, forming the basis of the SQL query.
WHERE Clause: The WHERE clause filters records based on specified conditions, allowing for precise data retrieval in SQL queries.
SELECT Statement: The SELECT statement is used to query the database and retrieve data, specifying columns to be displayed.
Joins in SQL: Joins in SQL are used to combine rows from two or more tables based on a related column, enabling complex queries across multiple datasets.
SQL Query Optimization: Optimizing SQL queries involves using indexes, avoiding unnecessary columns in SELECT, and minimizing subqueries to improve performance.

Experimental Design in Statistics

Sample Size: The sample size n affects the precision of estimates and the power of a statistical test, with larger samples providing more reliable results.
Power of a Test: The power of a test, $1 - \beta$, is the probability of correctly rejecting a false null hypothesis, and it increases with larger sample sizes and effect sizes.
Significance Level: The significance level α is the probability of rejecting the null hypothesis when it is true, commonly set at 0.05 for a 5**Variance:** Variance measures the dispersion of data points around the mean, and in experimental design, controlling variance is crucial for detecting true effects.
Experimental Design: Experimental design involves planning how to collect data efficiently and effectively to answer research questions, often using randomization to reduce bias.