1. Descriptive Statistics

Descriptive statistics is a branch of statistics that deals with summarising and describing the features of a dataset. It involves methods for organising and simplifying large amounts of data.

Key Concepts:

Measures of Central Tendency:

- Mean (Average): The sum of all data points divided by the number of points.
  Formula:
  Mean = (Sum of all data points) / n
  Application in Data Science: The mean is used to find the "average" of data, for example, in predicting house prices or evaluating the average score of students in a class.
- Median: The middle value of a dataset when arranged in ascending or descending order.
  Application in Data Science: The median is useful in skewed distributions (e.g., in salary data, where a few high values may distort the mean).
- Mode: The value that appears most frequently in a dataset.
  Application in Data Science: Mode is used in categorical data to find the most frequent category, such as in customer preferences or popular products.

Measures of Dispersion:

- Range: The difference between the maximum and minimum values.
  Formula:
  Range = Max - Min
  Application in Data Science: The range gives an idea of the spread of data (e.g., in stock prices).
- Variance: Measures the spread of data points from the mean.
  Formula:
  Variance = $(1 / n) * \Sigma(x_i - \mu)^2$
  Application in Data Science: Variance is used to measure the consistency of model predictions in machine learning.
- Standard Deviation: The square root of variance, providing the spread in the same units as the original data.
  Formula:
  Standard Deviation = $\sqrt{\text{Variance}}$
  Application in Data Science: Standard deviation helps evaluate model stability or variability in experimental results.
- Interquartile Range (IQR): The range between the first quartile (Q1) and the third quartile (Q3), representing the middle 50% of data.
  Application in Data Science: Used to identify outliers and assess the spread of data in predictive modelling.
- Skewness: A measure of the asymmetry of the data distribution.
  - Positive Skew: Long right tail, where mean > median.
  - Negative Skew: Long left tail, where mean < median.
    Application in Data Science: Skewness analysis helps in selecting the right transformation for skewed data in machine learning models.
- Kurtosis: Measures the "tailedness" of the data distribution.
  - Leptokurtic: High kurtosis, sharp peak.
  - Platykurtic: Low kurtosis, flat peak.
    Application in Data Science: Identifying the risk of extreme values, such as in financial data or risk modelling.

Application in Data Science:

- Summarising and understanding datasets, such as customer data, website traffic, or financial transactions.
- Data preprocessing: Identifying outliers and ensuring data normalisation.
- Model evaluation: Understanding how well your model fits the data using mean, variance, and standard deviation.

2. Frequency Distribution

Frequency distribution refers to how frequently each value appears in a dataset. It provides a summary of the data by organising the data points into intervals or categories.

Key Concepts:

- Frequency Table: A table that displays the frequency of different values or intervals.
  Application in Data Science: Used to understand the distribution of data before applying machine learning models.
- Relative Frequency: The fraction or percentage of the total number of observations that belong to a particular class.
  Formula:
  Relative Frequency = (Frequency of Class) / (Total Number of Observations)
  Application in Data Science: Analysing market share in business or customer preference by category.
- Cumulative Frequency: The sum of the frequencies up to a certain class.
  Application in Data Science: Used in building histograms and understanding cumulative distributions in data.
- Histogram: A bar chart representation of the frequency distribution, where each bar represents the frequency of an interval.
  Application in Data Science: Histogram is used for data visualisation, especially in understanding the distribution of continuous data (e.g., age distribution in a population).

Application in Data Science:

- Data exploration: Helps in identifying patterns, distributions, and potential outliers.
- Model input: Frequency distributions are essential for deciding how to handle categorical variables, such as encoding techniques (e.g., one-hot encoding).
- Feature engineering: Helps in identifying the most common values to design new features.

3. Probability

Probability is a branch of mathematics that deals with the likelihood of an event occurring. It is foundational in statistical inference, decision-making, and risk management.

Key Concepts:

- Basic Probability: The probability of an event is defined as the number of favourable outcomes divided by the total number of possible outcomes.
  Formula:
  P(A) = (Number of favourable outcomes) / (Total number of possible outcomes)
  Application in Data Science: Used in predictive modelling, such as in classification tasks, where probability helps to predict the likelihood of an event (e.g., churn prediction).
- Conditional Probability: The probability of an event given that another event has occurred.
  Formula:
  $P(A | B) = P(A \cap B) / P(B)$
  Application in Data Science: Important for anomaly detection and recommendation systems.
- Bayes' Theorem: A way to update probabilities based on new evidence.
  Formula:
  $P(A | B) = (P(B | A) * P(A)) / P(B)$
  Application in Data Science: Foundation of Naive Bayes classifiers used for text classification and spam detection.
- Distributions (Normal, Binomial, Poisson, etc.): Understanding different types of distributions helps in making predictions based on known patterns.
  Application in Data Science: Most machine learning algorithms assume data follows a normal distribution (e.g., linear regression, logistic regression).

Application in Data Science:

- Predictive modelling: Probability is central to estimating the likelihood of various outcomes in classification and regression tasks.
- Risk assessment: In fields like finance, probability is used to model potential losses or gains.
- Machine Learning algorithms: For models like Naive Bayes, random forests, etc., where probabilities are calculated for decision-making.

4. Inferential Statistics

Inferential statistics is the process of making conclusions about a population based on a sample of data. It involves hypothesis testing, estimation, and making predictions.

Key Concepts:

- Hypothesis Testing: A method for testing whether a hypothesis about a population parameter is supported by sample data.
    - Null Hypothesis (H0): The hypothesis that there is no effect or difference.
    - Alternative Hypothesis (H1): The hypothesis that there is an effect or difference.
    - P-value: The probability of observing the data, or something more extreme, if the null hypothesis is true.
      Application in Data Science: Used in A/B testing to decide whether a new feature or algorithm improves the current system.
- Confidence Intervals: A range of values, derived from the sample, that is likely to contain the population parameter with a certain level of confidence (e.g., 95% confidence interval).
  Application in Data Science: Helps quantify the uncertainty of model predictions.
- Regression and Correlation:
    - Correlation measures the strength and direction of the relationship between two variables.
    - Regression models the relationship between a dependent variable and one or more independent variables.
      Application in Data Science: Regression is used in predictive modelling, such as predicting sales, and correlation is used to identify relationships between variables, such as customer behaviour and marketing spend.

Application in Data Science:

- A/B Testing: Used for experimentation to test product or feature changes (e.g., test two versions of a website to see which one performs better).
- Model Evaluation: Confidence intervals are used to understand the uncertainty in model predictions.
- Feature selection: Identifying which features are statistically significant for model training.