Statistics (DSC 3: Introductory Statistics for Economics) - BA Economics Hons, DU

Statistics (DSC 3: Introductory Statistics for Economics)

BA Economics Hons, DU - Semester I: Exam-Focused Notes (NEP 2022)

This document provides a comprehensive and detailed overview of essential concepts in statistics, aligned with the Delhi University BA Economics Honours (NEP 2022) Semester I syllabus for DSC 3: Introductory Statistics for Economics.

Course Objectives and Learning Outcomes

- Objectives: To familiarize students with methods of summarizing and describing important features of data, introduce the basics of probability theory, and lay a foundation for Inferential Statistical Theory and Econometrics.
- Outcomes: Students will grasp concepts of probability, random variables, their distributions, and common discrete/continuous distributions, enabling real-life data analysis.

Unit 1: Introduction and Overview

This unit lays the groundwork for understanding statistical concepts, focusing on data types, data representation, and fundamental measures for describing datasets.

1.1 Introduction to Statistics

- Definition and Scope: Statistics is the science concerned with the collection, organization, summarization, analysis, interpretation, and presentation of data.
  - Descriptive Statistics: Methods for organizing, summarizing, and presenting data in an informative way. This includes calculating measures of central tendency and dispersion, and creating graphs.
  - Inferential Statistics: Methods used to draw conclusions or make predictions about a larger group (population) based on data gathered from a smaller subset of that group (sample). It involves hypothesis testing and estimation.
- Importance in Economics: Statistics is indispensable in economics for:
  - Forecasting: Predicting economic variables like GDP, inflation, unemployment, and stock prices.

  - Policy Formulation: Providing empirical evidence for designing and evaluating economic policies (e.g., fiscal, monetary, trade policies).
  - Market Research: Analyzing consumer behavior, market trends, and industry performance to inform business strategies.
  - Economic Modeling: Building and testing economic models using statistical techniques.

1.2 Populations, Samples, Parameters, and Statistics

- Population: The entire group of individuals or objects under consideration in a study about which information is desired (e.g., all households in Delhi, all registered voters).
- Sample: A subset or a smaller, representative group selected from the population (e.g., a survey of 500 households in Delhi, 1000 randomly selected voters).
- Parameter: A numerical characteristic or measure that describes a population (e.g., population mean μ, population standard deviation σ, population proportion P).
- Statistic: A numerical characteristic or measure that describes a sample. Statistics are calculated from sample data and are used to estimate population parameters (e.g., sample mean $\overline{x}$, sample standard deviation s, sample proportion $\hat{p}$).

1.3 Pictorial Methods in Descriptive Statistics

Graphical methods are essential for visualizing data distributions and patterns.

- Frequency Distribution: A table that lists categories of data along with the number of occurrences for each category.
- Bar Graphs/Charts: Used for displaying and comparing categorical (qualitative) data. Each bar represents a category, and its height indicates the frequency or proportion.
- Pie Charts: Used to show the proportion of each category within a whole. The entire pie represents 100% of the data.
- Histograms: Used for displaying the distribution of numerical (quantitative) data. Data are grouped into intervals (bins), and the height of each bar represents the frequency of observations falling into that interval. Adjacent bars touch, indicating continuous data.
- Frequency Polygons: A line graph that connects the midpoints of the tops of the bars in a histogram.
- Ogive (Cumulative Frequency Curve): A graph that plots cumulative frequencies against the upper class boundaries. It is useful for determining percentiles and other measures of relative standing.
- Stem-and-Leaf Plots: A method of organizing numerical data that shows the shape of the distribution while preserving the individual data values.
- Box-and-Whisker Plots (Box Plots): A standardized way of displaying the distribution of data based on five key numbers: minimum, first quartile (Q1), median (Q2), third

quartile (Q3), and maximum. It's useful for comparing distributions across different groups.

1.4 Measures of Location (Central Tendency)

These statistics aim to describe the center or typical value of a dataset.

- Mean ($\overline{x}$): The arithmetic average, calculated by summing all values and dividing by the number of observations.
  - Formula: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Merits: Uses all data points; provides a unique value.
  - Demerits: Highly sensitive to outliers or extreme values; not suitable for highly skewed distributions or categorical data.
- Median: The middle value of a dataset when it is arranged in ascending or descending order. If the number of observations (n) is odd, the median is the value at the $((n+1)/2)^{th}$ position. If n is even, it's the average of the two middle values.
  - Merits: Robust to extreme values (outliers); suitable for skewed distributions.
  - Demerits: Does not consider all data points; requires sorting the data.
- Mode: The value(s) that appear most frequently in a dataset. A dataset can be unimodal (one mode), bimodal (two modes), multimodal (more than two modes), or have no mode.
  - Merits: Applicable to both numerical and categorical data; easy to determine.
  - Demerits: May not be unique; may not exist; does not use all data points.

1.5 Measures of Variability (Dispersion)

These statistics describe the spread or dispersion of data points around the central value.

- Range: The simplest measure of dispersion, calculated as the difference between the maximum and minimum values in a dataset.
  - Merits: Easy to calculate.
  - Demerits: Only uses two extreme values, making it highly sensitive to outliers and not reflective of the spread of intermediate values.
- Quartile Deviation (Semi-Interquartile Range): Half the difference between the third quartile (Q3) and the first quartile (Q1). It measures the spread of the middle 50% of the data.
  - Formula: $QD = \frac{Q_3 - Q_1}{2}$
  - Merits: Less affected by extreme values than the range.
- Mean Deviation: The average of the absolute differences between each data point and the mean (or median).
  - Formula (from mean): $MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|$

  - Merits: Considers all data points.
  - Demerits: The use of absolute values makes it less amenable to further mathematical treatment compared to variance/standard deviation.
- Variance (σ^2 or s^2): The average of the squared deviations from the mean. It quantifies how much the individual data points typically vary from the mean.
  - Population Variance (σ^2): $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
  - Sample Variance (s^2): $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (using $n-1$ for unbiased estimation)
  - Merits: Uses all data points; mathematically tractable; forms the basis for many inferential statistics.
  - Demerits: Units are squared, making interpretation difficult; sensitive to outliers.
- Standard Deviation (σ or s): The square root of the variance. It returns the measure of dispersion to the original units of the data, making it more interpretable.
  - Formula (Population SD): $\sigma = \sqrt{\sigma^2}$
  - Formula (Sample SD): $s = \sqrt{s^2}$
  - Preference over Range: SD is generally preferred over range because it provides a more robust and comprehensive measure of variability by taking into account all observations in the dataset, rather than just the two extreme values.
- Coefficient of Variation (CV): A relative measure of dispersion, expressed as a percentage, which allows for the comparison of variability between datasets with different units or vastly different means.
  - Formula: $CV = \frac{\sigma}{\overline{x}} \times 100\%$ (or $s/\overline{x}$)
  - Merits: Unitless; useful for comparing variability across different scales.

Unit 2: Elementary Probability Theory

This unit introduces the fundamental concepts and rules governing probability, which is essential for understanding uncertainty and building inferential statistical models.

2.1 Sample Spaces and Events

- Experiment (Random Experiment): A process that leads to one of several possible outcomes, where the outcome cannot be predicted with certainty beforehand (e.g., flipping a coin, rolling a die, drawing a card).
- Outcome: A single possible result of a random experiment.
- Sample Space (S or Ω): The set of all possible distinct outcomes of a random experiment.
  - Discrete Sample Space: A sample space with a finite or countably infinite number of outcomes (e.g., {Heads, Tails} for a coin flip; {1, 2, 3, 4, 5, 6} for a die roll).

  - Continuous Sample Space: A sample space with an uncountably infinite number of outcomes, often represented by an interval (e.g., all real numbers between 0 and 1, the time it takes for a bulb to fuse).
- Event (A, B, C, ...): Any subset of the sample space. An event is a collection of one or more outcomes.
  - Simple Event: An event consisting of a single outcome.
  - Compound Event: An event consisting of more than one outcome.
  - Mutually Exclusive Events (Disjoint Events): Two events A and B are mutually exclusive if they cannot occur at the same time; their intersection is empty ($A \cap B = \emptyset$).
  - Collectively Exhaustive Events: A set of events such that at least one of the events must occur, and together they cover the entire sample space (their union is the sample space).

2.2 Probability Axioms and Properties (Kolmogorov's Axioms)

These are the fundamental rules that any valid probability assignment must satisfy.

1. Non-negativity: The probability of any event A is a non-negative real number.
   - $0 \leq P(A) \leq 1$ for any event A.
2. Normalization: The probability of the sample space (the certain event) is 1.
   - $P(S) = 1$
3. Additivity (for Mutually Exclusive Events): If $A_1, A_2, A_3, \ldots$ are a sequence of pairwise mutually exclusive events (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then the probability of their union is the sum of their individual probabilities.
   - $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$
4. For a finite number of mutually exclusive events, say A and B: $P(A \cup B) = P(A) + P(B)$.

Derived Properties:

- Complement Rule: $P(A') = 1 - P(A)$, where A' is the complement of event A (all outcomes in S that are not in A).
- General Addition Rule: For any two events A and B (not necessarily mutually exclusive): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- If $A \subseteq B$, then $P(A) \leq P(B)$.
- $P(\emptyset) = 0$.

2.3 Counting Techniques

These techniques are crucial for determining the number of possible outcomes in a sample space or an event, especially in situations where outcomes are equally likely (for classical probability).

- Multiplication Rule (Fundamental Principle of Counting): If an operation can be performed in n_1 ways, and if for each of these ways a second operation can be performed in n_2 ways, and so on, then the sequence of k operations can be performed in $n_1 \times n_2 \times \ldots \times n_k$ ways.
- Permutations: The number of ways to arrange n distinct objects in a specific order. The order matters.
  - Permutations of n distinct objects: $n! = n \times (n-1) \times \ldots \times 1$
  - Permutations of k objects chosen from n distinct objects (P_k^n or nPk):

$$P(n,k) = \frac{n!}{(n-k)!}$$

- Combinations: The number of ways to select k objects from a set of n distinct objects, where the order of selection does not matter.
  - Combinations of k objects chosen from n distinct objects (C_k^n or nCk or $\binom{n}{k}$):

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

2.4 Conditional Probability and Bayes' Rule

- Conditional Probability: The probability of an event A occurring, given that another event B has already occurred. It changes the sample space to event B.
  - Formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$, provided $P(B) > 0$.
  - Multiplication Rule for Dependent Events: $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$.
- Bayes' Theorem: A fundamental theorem that describes how to update the probability of a hypothesis based on new evidence.
  - Formula: For events $A_1, A_2, \ldots, A_n$ that form a partition of the sample space (mutually exclusive and collectively exhaustive), and any event B with $P(B) > 0$:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{k=1}^{n} P(A_k)P(B|A_k)}$$

  - Interpretation: $P(A_i)$ is the prior probability of A_i, $P(B|A_i)$ is the likelihood, and $P(A_i|B)$ is the posterior probability.

2.5 Independence of Events

- Two events A and B are considered independent if the occurrence of one does not affect the probability of the other.
- Conditions for Independence:
  - $P(A|B) = P(A)$ (if $P(B) > 0$)
  - $P(B|A) = P(B)$ (if $P(A) > 0$)
  - Product Rule for Independent Events: $P(A \cap B) = P(A)P(B)$

Unit 3: Random Variables and Probability Distributions

This unit introduces the concept of random variables, which are numerical outcomes of random experiments, and their associated probability distributions.

3.1 Defining Random Variables

- Random Variable (RV): A numerical quantity whose value is determined by the outcome of a random experiment. It's a function that maps outcomes from the sample space to real numbers. Random variables are typically denoted by capital letters (e.g., X, Y, Z).
- Types of Random Variables:
  - Discrete Random Variable: A random variable that can take on a finite or countably infinite number of distinct values (e.g., number of heads in 3 coin flips {0, 1, 2, 3}, number of cars passing a point in an hour).
  - Continuous Random Variable: A random variable that can take on any value within a given interval or range of real numbers (e.g., height, weight, time taken to complete a task, temperature).

3.2 Probability Distributions

A probability distribution describes the possible values that a random variable can take and the probability associated with each of those values.

- Probability Mass Function (PMF) (for Discrete RVs):
  - The PMF, $f_X(x)$ or $P(X = x)$, gives the probability that a discrete random variable X takes on a specific value x.
  - Properties:
    1. $0 \leq f_X(x) \leq 1$ for all possible values of x.
    2. $\sum_x f_X(x) = 1$ (the sum of all probabilities for all possible values must be 1).
  - Example: For a fair die roll, X = outcome. $f_X(x) = 1/6$ for $x \in \{1, 2, 3, 4, 5, 6\}$.
- Probability Density Function (PDF) (for Continuous RVs):
  - The PDF, $f_X(x)$, describes the relative likelihood for a continuous random variable X to take on a given value. For continuous variables, the probability of X taking any exact single value is 0.
  - Properties:
    1. $f_X(x) \geq 0$ for all x.
    2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (the total area under the curve is 1).
  - Calculating Probability: $P(a \leq X \leq b) = \int_a^b f_X(x)dx$
- Cumulative Distribution Function (CDF) (for both Discrete and Continuous RVs):

- The CDF, $F_X(x)$, gives the probability that a random variable X takes on a value less than or equal to a given value x.
- Definition: $F_X(x) = P(X \leq x)$
- Properties:
  1. $0 \leq F_X(x) \leq 1$.
  2. $F_X(x)$ is a non-decreasing function of x.
  3. $\lim_{x \to -\infty} F_X(x) = 0$
  4. $\lim_{x \to \infty} F_X(x) = 1$
  5. $F_X(x)$ is right-continuous.
- For Discrete RVs (from PMF): $F_X(x) = \sum_{t \leq x} f_X(t)$ (a step function).
- For Continuous RVs (from PDF): $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ (a continuous function).
- Relation: For a continuous random variable, $f_X(x) = \frac{d}{dx} F_X(x)$.

3.3 Expected Values and Functions of Random Variables

- Expected Value (Mean or Expectation) E[X] or μ_X: The weighted average of all possible values a random variable can take, where the weights are the probabilities of those values.
  - For Discrete RV: $E(X) = \sum_x xf(x)$
  - For Continuous RV: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$
- Expected Value of a Function of a Random Variable ($E[g(X)]$):
  - For Discrete RV: $E(g(X)) = \sum_x g(x)f(x)$
  - For Continuous RV: $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$
- Properties of Expectation:
  - $E(c) = c$ (for a constant c)
  - $E(cX) = cE(X)$
  - $E(X + Y) = E(X) + E(Y)$
- Variance Var(X) or σ_X^2: Measures the spread or dispersion of the random variable's values around its expected value.
  - Formula: $Var(X) = E[(X - \mu_X)^2] = E(X^2) - [E(X)]^2$
  - Properties of Variance:
    - $Var(c) = 0$
    - $Var(cX) = c^2 Var(X)$
    - $Var(X + c) = Var(X)$
    - For independent random variables X and Y: $Var(X + Y) = Var(X) + Var(Y)$

Unit 4: Sample Distributions (Commonly Used Probability Distributions)

This unit delves into specific types of probability distributions that frequently appear in statistical modeling and economic applications.

4.1 Discrete Distributions

- Uniform Discrete Distribution: Each possible outcome has an equal probability of occurring.
  - PMF: $f(x) = \frac{1}{n}$ for $x \in \{x_1, x_2, \ldots, x_n\}$ (where n is the number of possible outcomes).
  - Mean: $E(X) = \frac{x_{min}+x_{max}}{2}$
  - Variance: $Var(X) = \frac{(n^2-1)}{12}$ (for consecutive integers starting from 1)
- Binomial Distribution: Models the number of successes in a fixed number of independent Bernoulli trials.
  - Parameters: n (number of trials), p (probability of success on each trial).
  - PMF: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $k = 0, 1, ..., n$
  - Mean: $E(X) = np$
  - Variance: $Var(X) = np(1-p)$
- Poisson Distribution: Models the number of events occurring in a fixed interval of time or space, given a constant average rate of occurrence (λ) and independence of events.
  - Parameter: λ (average rate of events, $\lambda > 0$).
  - PMF: $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, for $k = 0, 1, 2, ...$
  - Mean: $E(X) = \lambda$
  - Variance: $Var(X) = \lambda$
- Hypergeometric Distribution: Models the number of successes in a sample drawn without replacement from a finite population.
  - Parameters: N (total population size), K (number of successes in the population), n (sample size).
  - PMF: $P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, for $k = \max(0, n-(N-K)), ..., \min(n, K)$
  - Mean: $E(X) = n\frac{K}{N}$
  - Variance: $Var(X) = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}$

4.2 Continuous Distributions

- Uniform Continuous Distribution: Describes a scenario where all values within a given interval $[a, b]$ are equally likely.
  - Parameters: a (minimum value), b (maximum value).

  - PDF: $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$
  - Mean: $E(X) = \frac{a+b}{2}$
  - Variance: $Var(X) = \frac{(b-a)^2}{12}$
- Normal Distribution (Gaussian Distribution): The most important continuous distribution in statistics due to its frequent occurrence in natural phenomena and its role in the Central Limit Theorem.
  - Parameters: μ (mean), σ (standard deviation).
  - PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $-\infty < x < \infty$
  - Properties:
    * Symmetry: Symmetric around its mean μ.
    * Bell-shaped: The graph of the PDF is a bell-shaped curve.
    * Asymptotic to x-axis: The tails extend infinitely, approaching but never touching the x-axis.
    * Empirical Rule (68-95-99.7 Rule):
      · $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6826$
      · $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9544$
      · $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$
    * Standard Normal Distribution (Z-distribution): A normal distribution with $\mu = 0$ and $\sigma = 1$. Any normal RV X can be standardized: $Z = \frac{X-\mu}{\sigma}$.

Unit 5: Random Sampling and Jointly Distributed Random Variables

This unit extends probability concepts to situations involving multiple random variables and introduces the basics of sampling.

5.1 Random Sampling

- Random Sample: A sample chosen from a population in such a way that every individual or set of individuals has an equal chance of being selected.
  - Simple Random Sampling (SRS): Each possible sample of a given size has an equal chance of being selected.
- Sampling Distribution: The probability distribution of a statistic (e.g., sample mean, sample proportion) derived from all possible samples of a given size drawn from a population.
- Central Limit Theorem (CLT): For a sufficiently large sample size ($n \geq 30$), the sampling distribution of the sample mean ($\overline{X}$) will be approximately normally distributed, regardless of the population distribution.

- If X_i are i.i.d. random variables with mean μ and variance σ^2, then for large n, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

5.2 Jointly Distributed Random Variables

- Joint Probability Distribution: Describes the probabilities of combinations of values for multiple random variables.
  - Joint PMF for Discrete RVs (X, Y):

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

    * Properties: $0 \leq f_{X,Y}(x, y) \leq 1$ and $\sum_x \sum_y f_{X,Y}(x, y) = 1$.
  - Joint PDF for Continuous RVs (X, Y):

$$f_{X,Y}(x, y)$$

    * Properties: $f_{X,Y}(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$.
    * $P((X, Y) \in R) = \iint_R f_{X,Y}(x, y)dxdy$ for some region R.
- Marginal Distributions:
  - Marginal PMF for X: $f_X(x) = \sum_y f_{X,Y}(x, y)$
  - Marginal PDF for X: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$
- Expected Values of Jointly Distributed Random Variables:
  - Expected Value of a Function: $E(g(X, Y)) = \sum_x \sum_y g(x, y)f_{X,Y}(x, y)$ (Discrete) or $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$ (Continuous)
  - Expectation of Sums: $E(X + Y) = E(X) + E(Y)$
- Conditional Distributions and Expectations:
  - Conditional PMF (Discrete): $f_{Y|X}(y|x) = P(Y = y|X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$, provided $f_X(x) > 0$.
  - Conditional PDF (Continuous): $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$, provided $f_X(x) > 0$.
  - Conditional Expectation:
    * Discrete: $E(Y|X = x) = \sum_y y f_{Y|X}(y|x)$
    * Continuous: $E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$
- Covariance and Correlation:
  - Covariance ($\text{Cov}(X, Y)$): Measures the extent to which two random variables vary together.

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

    * $\text{Cov}(X, Y) > 0$: Tend to move in the same direction.
    * $\text{Cov}(X, Y) < 0$: Tend to move in opposite directions.

    - $\text{Cov}(X, Y) = 0$: No linear relationship.
    - If X and Y are independent, then $\text{Cov}(X, Y) = 0$.
    - $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
  - Correlation Coefficient ($\rho_{X,Y}$ or r): A standardized measure of the linear relationship between two random variables, ranging from -1 to +1.

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

    - $\rho_{X,Y} = +1$: Perfect positive linear relationship.
    - $\rho_{X,Y} = -1$: Perfect negative linear relationship.
    - $\rho_{X,Y} = 0$: No linear relationship.

Additional Concepts

Skewness and Kurtosis

- Skewness: Measures the asymmetry of the probability distribution of a real-valued random variable about its mean.
  - Positive Skew (Right-skewed): The tail on the right side is longer or fatter; Mean > Median > Mode.
  - Negative Skew (Left-skewed): The tail on the left side is longer or fatter; Mean < Median < Mode.
  - Symmetric: Mean = Median = Mode.
- Kurtosis: Measures the "tailedness" of the probability distribution.
  - Mesokurtic: Similar peakedness to a normal distribution (excess kurtosis = 0).
  - Leptokurtic: More peaked (or heavier tails) than a normal distribution (excess kurtosis > 0).
  - Platykurtic: Flatter than a normal distribution (or lighter tails) (excess kurtosis < 0).

Correlation and Regression Analysis

- Correlation: Pearson's 'r' is the sample estimate of $\rho_{X,Y}$.
- Regression Analysis: Models the relationship between a dependent variable and one or more independent variables.
  - Simple Linear Regression: Models the linear relationship between two variables.
    - Equation: $Y = \beta_0 + \beta_1 X + \epsilon$
    - $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ (estimated regression line)
    - Least Squares Method: Minimizes $\sum (Y_i - \hat{Y}_i)^2$.
- Correlation vs. Regression:

– Correlation: Measures strength and direction of a linear association.
– Regression: Describes the nature of the relationship for prediction.