DATA COLLECTION, ANALYSIS AND DECISION MAKING

Chapter 1: Introduction of Data in Decision Making

a) Importance of Data

Data is the **lifeblood** of modern decision-making. It provides the **raw facts, figures, and information** needed to move beyond guesswork and intuition.

- **Evidence-Based Decisions:** Data offers concrete evidence, replacing subjective opinions with objective facts.
- **Identification of Trends:** Analyzing data helps spot patterns, market shifts, and emerging customer behavior, allowing businesses to stay ahead.
- **Problem Solving:** Data pinpoints the root cause of issues, making problem resolution more targeted and effective.
- **Operational Efficiency:** Detailed data on processes helps in identifying bottlenecks and optimizing resource allocation, leading to cost savings.

b) Importance of Decision Making

Decision-making is the process of choosing a course of action among several alternatives. In a business context, good decisions determine **success, growth, and survival**.

- **Goal Achievement:** Strategic decisions ensure that organizational efforts are aligned with overarching goals.
- **Resource Allocation:** Decisions guide how limited resources (time, money, personnel) are best utilized for maximum impact.
- **Risk Management:** Thoughtful decisions, informed by data, help in anticipating potential risks and developing mitigation strategies.
- **Competitiveness:** The speed and quality of decisions often differentiate market leaders from followers.

c) Implication of Data-Driven Decisions on Key Performance Indicators (KPIs)

Data-driven decisions mean using facts to guide strategy, operations, and actions. This approach has a direct and positive implication on **Key Performance Indicators (KPIs)**.

- **Improved Accuracy:** Decisions based on solid analysis are less likely to fail, directly improving KPIs like **Success Rate** or **Conversion Rate**.
- **Enhanced Efficiency:** Analyzing operational data allows for optimization, leading to better KPIs such as **Cycle Time** or **Cost Per Unit**.
- **Increased Customer Satisfaction:** Data on customer behavior and feedback guides product/service improvements, boosting KPIs like **Net Promoter Score (NPS)** or **Customer Retention Rate**.
- **Higher Profitability:** Better forecasting and risk management, driven by data, positively affect financial KPIs like **Return on Investment (ROI)** and **Revenue Growth**.

d) Transforming Data into Actionable Insight

Raw data, by itself, is not enough. It must be processed and analyzed to become **actionable insight**. An insight is a realization that leads to a clear action or strategy.

- **Collection and Cleaning:** The process begins with gathering relevant, high-quality data and cleaning it to remove errors or inconsistencies.
- **Analysis:** Statistical and analytical techniques are applied to explore the data, look for correlations, and identify significant findings.
- **Interpretation:** The results of the analysis are interpreted in the context of the business problem to extract meaningful observations.
- **Visualization and Communication:** Insights are often presented using **data visualization** to communicate complex findings simply and clearly to decision-makers. The final step is translating the insight into a **clear, measurable action plan**.

e) Impact on Business Strategy

Data-driven insights profoundly shape and redefine **business strategy**.

- **Market Strategy:** Data informs target audience definition, pricing, and optimal marketing channels.
- **Product Development:** Customer usage and feedback data guide feature prioritization and new product development roadmaps.
- **Operational Strategy:** Supply chain and logistics data help in building more resilient and efficient operations.
- **Strategic Direction:** Data can reveal entirely new market opportunities or signal the need to exit underperforming areas, directly influencing the long-term **strategic direction** of the enterprise.

Chapter 2: Sampling Concepts

a) Sample Design

Sample design is a definitive plan for obtaining a sample from a given population. It must be carefully crafted to ensure the sample is representative of the whole population.

i. Defining Objectives

- The first and most critical step is to clearly define the **research question** or **objective** the sample is intended to answer. This determines what data needs to be collected.

ii. Target Population

- This refers to the entire group of individuals, objects, or items that the research is interested in drawing conclusions about. It must be clearly and precisely defined.

iii. Sampling Units

- A **sampling unit** is the basic unit of observation that is selected for the sample. This could be an individual person, a household, a company, or a geographic area.

iv. Size of Sample

- The **sample size (n)** is the number of units to be included in the sample. It is determined by factors like the desired level of **precision**, the variability within the population, and the available budget.

v. Parameters of Interest

- These are the **characteristics** of the target population that the study aims to estimate or make inferences about, such as the population **mean** (μ) or **proportion** (P).

vi. Data Collection

- This involves determining the method for gathering information from the selected sampling units, which could include surveys, interviews, or observation.

b) Sampling Errors

Sampling errors are inaccuracies that arise because only a **subset** (sample) of the population is studied rather than the entire population.

- **Definition:** The difference between the value calculated from the sample (the statistic) and the true value for the population (the parameter).
- **Cause:** They are inherent in the sampling process, especially due to **random chance** in selecting the sample.
- **Control:** They can be reduced by increasing the **sample size** and using a robust, appropriate sampling method. They are *not* caused by mistakes, unlike non-sampling errors.

c) Sample Survey vs Census Survey

- **Sample Survey:**
  - **Involves:** Collecting data from only a **representative subset** of the population.
  - **Advantages:** Faster, less expensive, and feasible for large populations.
  - **Disadvantages:** Results are subject to sampling error.
- **Census Survey:**
  - **Involves:** Collecting data from **every single unit** in the entire population.
  - **Advantages:** Provides a true measure of the population parameter (no sampling error).
  - **Disadvantages:** Extremely costly, time-consuming, and often impractical for very large populations.

d) Design Effect

The **design effect (de ff)** is a statistical measure used to quantify the loss of precision (increase in variance) that results from using a complex sampling design (like cluster sampling) instead of **Simple Random Sampling (SRS)**.

- **Formula:** de ff = Variance of the estimate under the actual sampling design / Variance of the estimate under Simple Random Sampling
- **Interpretation:** A de ff value **greater than 1** indicates that the complex design is less efficient than SRS, requiring a larger sample size to achieve the same precision. It helps researchers adjust the required sample size.

e) Sampling Methods

Probability Sampling

Probability sampling methods ensure that **every unit in the population has a known, non-zero chance** of being selected. This allows for statistical inferences about the population.

i. Simple Random Sampling (SRS)

- **Method:** Every possible sample of a given size has an equal chance of being selected. Selection is typically done using a random number generator.
- **Characteristics:** Easiest to understand, provides unbiased estimates, but requires a complete list (sampling frame) of the entire population.

ii. Systematic Sampling

- **Method:** Selects units at regular intervals from a sampling frame, after a random start. For example, selecting every kth unit.
- **Characteristics:** Simpler to execute than SRS, but can be biased if there is a periodicity or pattern in the population list that coincides with the sampling interval.

iii. Stratified Sampling

- **Method:** The population is first divided into mutually exclusive and exhaustive subgroups called **strata** (e.g., age groups, gender). Then, a sample is randomly selected from *each* stratum.
- **Characteristics:** Ensures representation of key subgroups, leading to more precise estimates if the strata are internally homogeneous and externally heterogeneous.

iv. Cluster Sampling

- **Method:** The population is divided into groups called **clusters** (e.g., geographic areas, schools). A random sample of clusters is selected, and *all* units within the selected clusters are then surveyed.
- **Characteristics:** Cost-effective for geographically dispersed populations, but generally results in less precise estimates compared to SRS because units within a cluster tend to be similar.

Non-Probability Sampling

Non-probability sampling methods do **not** rely on random selection, meaning some units have no chance of being selected, and the probability of selection cannot be determined. These methods are common in qualitative research or pilot studies where statistical

generalization is not the primary goal. Examples include convenience sampling, quota sampling, and judgmental sampling.

Chapter 3: Introduction to Applied Statistics

a) Identifying the Dependent and Independent Variable

In statistical modeling and research, variables are classified based on their role in a potential cause-and-effect relationship.

- **Independent Variable (IV):**
  - The variable that is **manipulated or chosen** to predict or cause a change in another variable.
  - It is often denoted as X and is the **predictor** or **explanatory** variable.
  - *Example:* The amount of fertilizer given to a plant.
- **Dependent Variable (DV):**
  - The variable that is **measured** or observed to see if it is affected by the independent variable.
  - It is often denoted as Y and is the **outcome** or **response** variable.
  - *Example:* The height of the plant.

Understanding this relationship is fundamental to setting up most statistical analyses.

b) Hypothesis Testing

Hypothesis testing is a formal procedure for investigating a claim about a population using data from a sample. It helps determine if there is enough evidence to reject a presumption (the null hypothesis).

i. Characteristics of a Hypothesis

A good statistical hypothesis must be:

- **Testable:** It must be possible to collect data that can either support or refute it.
- **Falsifiable:** It must be stated in such a way that it can be proven wrong.
- **Clear and Concise:** Stated simply and precisely, defining the variables and population of interest.
- **Specific:** It should specify the relationship or difference being investigated.

ii. Null Hypothesis (H0) & Alternative Hypothesis (Ha or H1)

- **Null Hypothesis (H0):**
  - A statement of **no effect or no difference**. It represents the status quo or what is currently believed to be true.
  - The goal of the test is often to **reject** H0.
  - *Example:* "There is no difference in average sales between the old and new advertisements."
- **Alternative Hypothesis (Ha or H1):**
  - A statement that **contradicts the null hypothesis**. It is the claim the researcher is trying to find evidence to support.

- *Example:* "The average sales are higher with the new advertisement."

iii. Procedure of Hypothesis Testing

The general procedure follows these steps:

1. **State the Hypotheses:** Formulate H0 and Ha.
2. **Determine Significance Level (α):** Choose the acceptable risk of rejecting a true H0 (e.g., $\alpha=0.05$).
3. **Calculate the Test Statistic:** Compute the statistic (like t, Z, or F) using the sample data.
4. **Determine the P-value or Critical Value:** Find the probability of observing the data if H0 were true (p-value) or define the rejection region (critical value).
5. **Make a Decision:**
    - **If P-value ≤α: Reject H0**. Conclude there is statistically significant evidence to support Ha.
    - **If P-value >α: Fail to reject H0**. Conclude there is not enough evidence to reject the status quo.

c) Confidence Levels

The **confidence level** expresses the degree of certainty that a study's results would be the same if repeated.

- **Definition:** It is the long-run percentage of times that an interval estimation (called a **confidence interval**) will contain the true population parameter.
- **Common Levels:** Most commonly, the confidence level is **95%**, which corresponds to an α (significance level) of $1-0.95=0.05$.
- **Confidence Interval:** For a 95% confidence level, the corresponding confidence interval means that we are 95% confident that the true population parameter lies within that range.

d) Math that Manipulates Data

The "math that manipulates data" is essentially the core of **statistics** and **mathematical modeling**. It involves a range of techniques and formulas to process raw data into meaningful results.

- **Summary Statistics:** Calculating measures like mean, median, and standard deviation to summarize data features.
- **Probability Theory:** The mathematical foundation for dealing with uncertainty and randomness.
- **Estimation:** Using sample statistics to estimate unknown population parameters.
- **Regression and Modeling:** Using mathematical equations (like $Y=\beta0+\beta1X+\epsilon$) to model the relationship between variables and make predictions.
- **Inferential Formulas:** Equations used in hypothesis tests (e.g., t-statistic formula) to measure the evidence against H0.

Chapter 4: Descriptive Statistics 📈

a) Summarizing and Describing a Collection of Data

Descriptive statistics are used to summarize and describe the main features of a collection of data in a meaningful way. They provide simple summaries about the sample and observations without drawing conclusions beyond the data itself.

- **Central Tendency:** Describes the center point of the data distribution (Mean, Median, Mode).
- **Variability (Dispersion):** Describes how spread out the data is (Range, Standard Deviation, Variance).
- **Distribution Shape:** Describes the pattern of the data (Skewness, Kurtosis).

b) Univariate and Bivariate Analysis

- **Univariate Analysis:**
  - Involves the analysis of **a single variable** at a time.
  - **Purpose:** To describe the distribution, central tendency, and dispersion of that one variable.
  - *Examples:* Calculating the average age of customers, creating a frequency table of product preferences.
- **Bivariate Analysis:**
  - Involves the analysis of **two variables simultaneously** to determine the empirical relationship between them.
  - **Purpose:** To see if a relationship, association, or correlation exists.
  - *Examples:* Examining the relationship between advertising spending and sales (using correlation or scatter plots), or comparing the mean score for two different groups.

c) Mean, Median, Mode & Standard Deviation

These are key measures of central tendency and dispersion.

- **Mean:** The **arithmetic average** of a dataset. It is calculated by summing all values and dividing by the count of observations. Highly affected by outliers.
- **Median:** The **middle value** in an ordered dataset. 50% of the data falls below the median. It is preferred when the data contains extreme outliers or is heavily skewed.
- **Mode:** The **most frequently occurring value** in a dataset. It is most useful for nominal (categorical) data.
- **Standard Deviation (σ or s):** A measure of the **amount of variation or dispersion** of a set of values. A low standard deviation indicates that the values tend to be close to the mean, while a high standard deviation indicates the values are spread out over a wider range.

d) Percentages and Ratios

- **Percentages:** Used to express a fraction of a whole as a number out of 100.
  - *Formula:* Percentage= Part/Whole ×100
  - They are crucial for understanding proportions, such as market share or completion rates.

- **Ratios:** A relationship between two numbers, indicating how many times the first number contains the second.
    - *Formula:* Ratio=A/B (expressed as A:B or A to B)
    - They are used to compare the magnitude of two different quantities, such as debt-to-equity ratio or male-to-female ratio.

e) Histograms

A **histogram** is a graphical representation of the distribution of **numerical data**.

- **Appearance:** It looks similar to a bar chart, but the bars represent ranges of data (called **bins**) and are typically drawn touching each other to show the data is continuous.

- **Purpose:** They show the **shape** of the distribution, allowing analysts to quickly grasp where values are concentrated, the extent of variation, and whether the distribution is symmetrical or skewed.

f) Identifying Randomness and Uncertainty in Data

Data inherently contains **randomness** (unpredictable variation) and **uncertainty** (doubt about the true value or outcome).

- **Randomness:** Can be identified through:
    - **Residuals:** In modeling, truly random data will have residuals (errors) that are randomly scattered with no discernible pattern.
    - **Probability Distributions:** Recognizing that data often follows known patterns like the Normal, Poisson, or Binomial distribution, which are built on probability and randomness.
- **Uncertainty:** Is addressed through:
    - **Inferential Statistics:** Using techniques (Chapter 5) like confidence intervals to quantify the range of plausible values for a parameter.
    - **Standard Error:** A measure of the variability of a sample statistic, which quantifies the uncertainty in using a sample to estimate a population parameter.

Chapter 5: Inferential Statistics

a) Drawing Inference from Data

Inferential statistics uses data from a **sample** to draw conclusions or make **inferences** about a larger **population**.

- **Core Goal:** To go beyond the immediate data and make generalizations, test hypotheses, and make predictions.
- **Key Concept:** Inferences are always made with a measurable degree of **uncertainty** (expressed via p-values and confidence intervals).

b) Modelling

Statistical modeling is the process of using mathematical equations to represent the relationships between variables in a dataset.

- **Purpose:** To explain, predict, and control outcomes by identifying which independent variables influence a dependent variable.
- **Forms:** Models can be simple (like a linear regression equation) or complex (like a time series model or a neural network). All models are simplifications of reality.

c) Assumptions

Most inferential statistical tests rely on a set of **mathematical assumptions** about the data. If these assumptions are significantly violated, the results of the test may be invalid or misleading.

- **Common Assumptions:**
  - **Normality:** The data (or the error term/residuals) follows a normal distribution.
  - **Homoscedasticity (Equal Variance):** The variance of the dependent variable is equal across all levels of the independent variable.
  - **Independence:** Observations are independent of one another.
  - **Linearity:** For regression, the relationship between the independent and dependent variables is linear.

d) Identifying Patterns

Inferential techniques are critical for moving beyond simple descriptions to **identifying underlying patterns** that are not visible to the naked eye.

- **Relationships:** Discovering whether two variables are related (correlation, regression).
- **Differences:** Finding statistically significant differences between groups (t-tests, ANOVA).
- **Clustering:** Grouping similar data points together (used in machine learning).

e) Regression Analysis

Regression analysis is a powerful statistical technique used to **model the relationship** between a dependent variable and one or more independent variables.

- **Linear Regression:** Fits a straight line to the data to model the relationship, with the simplest form being $Y=\beta0+\beta1X+\epsilon$.
  - $\beta0$ is the intercept, $\beta1$ is the slope coefficient, and ϵ is the error term.
- **Purpose:** Prediction, forecasting, and determining the strength and direction of a relationship.

f) T-test

The **t-test** is an inferential statistic used to determine if there is a **significant difference** between the means of **two groups**.

- **Types:**
    - **One-Sample T-test:** Compares a sample mean to a known value or population mean.
    - **Independent Samples T-test:** Compares the means of two completely separate groups (e.g., male vs. female scores).
    - **Paired Samples T-test:** Compares the means of the same group at two different times (e.g., before and after a training program).

g) Analysis of Variance (ANOVA)

ANOVA is used to determine if there are any statistically significant differences between the means of **three or more independent groups**.

- **Core Logic:** It examines the **ratio of variance between groups** to the **variance within groups** (the F-ratio). A large F-ratio suggests the difference between groups is larger than the random variation within groups.
- **Types:** One-way ANOVA (one independent variable) and Two-way ANOVA (two independent variables).

h) Correlations

Correlation measures the **strength and direction** of the linear relationship between two quantitative variables.

- **Pearson's *r*:** The most common measure, which ranges from **−1 to +1**.
    - **+1:** Perfect positive linear relationship.
    - **−1:** Perfect negative linear relationship.
    - **0:** No linear relationship.
- **Important Note: Correlation does not imply causation**. It only measures association.

i) Chi-square Test (χ2 Test)

The **Chi-square test** is used primarily with **categorical data** (nominal or ordinal) to check for association or goodness-of-fit.

- **Test of Independence:** Determines whether there is a statistically significant relationship between two categorical variables (e.g., Is there a relationship between gender and preferred color?).
- **Goodness-of-Fit Test:** Compares an observed frequency distribution to a theoretical or expected distribution.

Important questions for full subject:

Chapter 1: Introduction of Data in Decision Making

1. **Fundamental Role:** Why is relying solely on managerial intuition and experience risky in today's business environment, and how does **data** mitigate this risk?

2. **Actionable Insight:** Explain the critical difference between **raw data**, **information**, and **actionable insight**. Provide a practical example of how raw sales data is transformed into an actionable insight.
3. **KPI Impact:** Choose a common business KPI (e.g., Customer Churn Rate or Website Conversion Rate). Describe how a specific data-driven decision could directly and quantitatively **improve** that KPI.
4. **Strategic Shift:** How does the adoption of a data-driven culture force an organization to rethink its **long-term business strategy** compared to a traditional, gut-driven approach?
5. **Data Quality:** Data's importance is tied to its quality. Discuss the **implications of using poor-quality, inconsistent data** for a major business decision.

Chapter 2: Sampling Concepts

1. **Census vs. Sample:** Under what specific conditions would a **Census Survey** be the superior choice over a **Sample Survey**, despite the higher cost and time commitment?
2. **Sample Design Justification:** You are asked to survey university students. Justify the definition of the following elements for your study: **Target Population**, **Sampling Unit**, and the **Parameter of Interest**.
3. **Sampling Error:** Explain the concept of **Sampling Error**. Can this error be completely eliminated in a sample survey? If so, how? If not, why not?
4. **Probability vs. Non-Probability:** Why is **Probability Sampling** preferred for making statistical inferences about a population, while **Non-Probability Sampling** is generally not suitable for this purpose?
5. **Sampling Method Choice:**
    - Explain how **Stratified Sampling** is fundamentally different from **Cluster Sampling**.
    - In which scenario would you use **Stratified Sampling** (e.g., polling diverse income groups), and in which scenario would you use **Cluster Sampling** (e.g., surveying people across a large geographic area)?
6. **Design Effect:** A study uses cluster sampling and finds a **Design Effect (de ff) of 1.5**. Explain what this value means for the efficiency of the sampling method and the necessary **sample size**.

Chapter 3: Introduction to Applied Statistics

1. **Variable Roles:** Identify the **Independent** and **Dependent** variables in the following statement, and explain the expected relationship: "The number of hours a student studies is related to their exam score."
2. **Hypothesis Formulation:** A company believes its new website layout increases the average time spent on site (currently 4.5 minutes). State the appropriate **Null Hypothesis (H0)** and **Alternative Hypothesis (Ha)** for testing this claim.
3. **Type I and Type II Errors:** Explain the meaning of the **Significance Level (α)** in the context of hypothesis testing. How does this relate to the risk of committing a **Type I Error** (rejecting a true null hypothesis)?
4. **Inference and Confidence:** If a study reports a **99% Confidence Interval** for the mean salary of a population, what does this actually tell us about the true population mean, and how does it relate to the chance of error?

5. **Statistical Procedure:** Briefly outline the essential four-step **procedure for hypothesis testing**, beginning with the statement of hypotheses and ending with the decision.

Chapter 4: Descriptive Statistics

1. **Outlier Impact:** You have a small dataset of household incomes. Explain how the presence of a single, extremely high income (**outlier**) would affect the **Mean** compared to the **Median**. Which measure would be more representative of the typical income?
2. **Dispersion Importance:** Why is the **Standard Deviation** considered a more robust and informative measure of data variability than the simple **Range**?
3. **Histogram Interpretation:** Describe three key pieces of information (related to shape, center, and spread) that you can immediately infer about a dataset by observing its **Histogram**.

4. Bivariate Purpose: When analyzing data, why is it necessary to move from Univariate Analysis (analyzing one variable) to Bivariate Analysis (analyzing two variables)? Give an example of a relationship that only bivariate analysis can uncover.

5. Describing Uncertainty: In the context of descriptive statistics, how do measures like skewness and kurtosis help in understanding the randomness and uncertainty present in a dataset's distribution?

Chapter 5: Inferential Statistics

1. **Inference vs. Description:** What is the primary functional difference between **Descriptive Statistics** and **Inferential Statistics**? What is the role of **probability theory** in this difference?
2. **Model Assumptions:** Why is checking the **assumptions** (such as normality or homoscedasticity) critical before interpreting the results of an inferential test like **Regression** or **ANOVA**? What happens if an assumption is violated?
3. **Regression Interpretation:** In a simple linear regression model, the slope coefficient ($\beta 1$) is calculated as **+0.75**. Interpret this coefficient in the context of the relationship between the Independent Variable (X) and the Dependent Variable (Y).
4. **T-test vs. ANOVA:** Both the **t-test** and **ANOVA** are used to test for differences in means. Explain the structural difference between these two tests and when you would choose one over the other.
5. **Correlation and Causation:** A study finds a very strong **positive correlation (r = 0.92)** between the sale of ice cream and the number of drowning incidents in a city. Critically discuss why this correlation **does not** imply causation, and what the likely hidden variable (confounder) is.
6. **Chi-square Application:** Describe a scenario in which the **Chi-square Test of Independence** would be the appropriate statistical tool to use. What specific type of data (variable type) is required for this test?