DESCRIPTIVE STATISTICS

1 UNDERSTANDING DATA

- **Numerical Data**:
  - **Continuous**: Can take any value within a range (e.g., height, weight).
  - **Discrete**: Countable, finite values (e.g., number of students).
- **Categorical Data**:
  - **Nominal**: No order (e.g., gender, color).
  - **Ordinal**: Ordered categories (e.g., rating scales).

2 MEASURES OF CENTRAL TENDENCY

- **Mean** ($\bar{x}$): $\bar{x} = \frac{\sum x}{n}$, *prone to outliers*
- **Median**
  - Odd n: Middle element in sorted data
  - Even n: Average of two middle elements
- **Mode**: Most frequent value
- **Weighted Mean:** $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$
- **Moving Avg (Sliding Window):** Average over last k data points

3 DATA SCATTERNESS

- **Range:** $\text{Range} = \text{Max} - \text{Min}$
- **Quartiles (IQR):** $\text{IQR} = Q3 - Q1$ Outliers:

: Below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$.

Percentiles:

- **Calculating Position:** $I = \frac{(n-1) \times P}{100}$

$$\text{Value at } I = \text{Value at } \lfloor I \rfloor + (I - \lfloor I \rfloor)$$

$$\times (\text{Value at } \lceil I \rceil - \text{Value at } \lfloor I \rfloor)$$

4 PERCENTILE CALCULATION

Percentile Calculation:

- **Excluded:** $\frac{\text{No. of data points} < P}{\text{Total no. of data points}} \times 100$
- **Included:** $\frac{\text{No. of data points} <= P}{\text{Total number of data points}} \times 100$
- **Mid-Point Adjustment:**

$$\frac{\text{No. data points} < P + \frac{1}{2} \times \text{No. of data points} = P}{\text{Total no. of data points}} \times 100$$

5 MEASURES OF DISPERSION

MAD (Mean Absolute Deviation): Average distance from the mean; $\frac{\sum |x_i - \bar{x}|}{n}$

Variance: Average squared deviations; $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$

Standard Deviation: Square root of variance; $\sigma = \sqrt{\sigma^2}$

6 SHAPE OF SPREAD

Skewness:

- **Left-Skewed**: Long tail on the left
- **Right-Skewed**: Long tail on the right

Kurtosis:

- **Platykurtic** (<3): Flatter distribution
- **Mesokurtic** (=3): Normal distribution-like
- **Leptokurtic** (>3): Peaked distribution with heavy tails

7 RELATION BETWEEN TWO VARIABLES

Covariance: Measures the directional relationship; (

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

Correlation:

- **Pearson**: Standard measure of correlation; $\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$
- **Spearman**: Rank-based correlation; $r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$

PROBABILITY THEORY

1 BASIC CONCEPTS

- **Sample Space (Ω):** All possible outcomes.
- **Event (E):** A subset of sample space.
- **Core Properties:** 0 ≤ P(E) ≤ 1, P(Ω) = 1.
- **Empirical Probability:** Based on observations.

2 PROBABILITY RULES

- **Sample Space (S)**: All possible outcomes.
- **Event (E)**: Subset of sample space.
- **Probability of Equally Likely Outcomes**: $P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total outcomes}}$
- **Empirical Probability**: $P(E) = \frac{\text{Frequency of E}}{\text{Total trials}}$

Set Operations

- **Union ($A \cup B$)**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Intersection ($A \cap B$)**: $P(A \cap B)$
- **Complement (A')**: $P(A') = 1 - P(A)$
- **De Morgan's Laws**: $(A \cup B)' = A' \cap B'$, $(A \cap B)' = A' \cup B'$

Events

- **Mutually Exclusive**: $P(A \cap B) = 0$
- **Mutually Exhaustive**: $P(A \cup B \cup \dots) = 1$
- **Inclusion-Exclusion Principle**: Adjust for overcounting/undercounting.

3 COUNTING TECHNIQUES

- **Permutations**: $P(n,k) = \frac{n!}{(n-k)!}$
- **Combinations**: $C(n,k) = \frac{n!}{k!(n-k)!}$
- **Multiplication Rule**: Use when events are sequential.
- **Addition Rule**: Use when events are mutually exclusive.

4 CONDITIONAL PROBABILITY

- **Conditional Probability**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- **Independent Events**: $P(A \cap B) = P(A) \times P(B)$
- **Law of Total Probability**: $P(A) = \sum P(A|B_i) P(B_i)$
- **Bayes Theorem**: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

EXPLORATORY DATA ANALYSIS WITH PANDAS & SEABORN

1 DATAFRAME OVERVIEW

- **Basic Info:** `df.head()`, `df.tail()`, `df.shape`, `df.info()`, `df.describe()`
- **Column Operations:**
  - Access: `df["col"]`, `df[["col1", "col2"]]`
  - Aggregation: `df["col"].mean()`, `.sum()`, `.count()`
  - Unique Values: `df["col"].unique()`, `.nunique()`
  - String Operations: `df["col"].str.contains("pattern")`
  - Creating Columns: `df["new_col"] = df["col1"] + df["col2"]`

2 ROW OPERATIONS

- Access: `df.loc[]`, `df.iloc[]`
- Filtering: `df[(df['col1'] > 10) & (df['col2'] < 5)]`
- Sorting: `df.sort_values(by='col', ascending=False)`

3 MISSING VALUES ETC

Handling Missing Data

- **Identification:** `df.isnull().sum()`
- **Imputation:** `df.dropna()`, `df.fillna(value)`

Grouping & Aggregation

- `df.groupby("col")["num_col"].mean()`

Correlation Analysis

- `df.corr(method="pearson")` (default), `method="spearman"`

4 DATA VISUALIZATION

- **Univariate:**
  - Categorical: `sns.barplot()`, `plt.pie()`
  - Numerical: `sns.boxplot()`, `plt.hist()`, `sns.kdeplot()`
- **Bivariate:**
  - Categorical-Categorical: `sns.countplot()`, stacked/dodged barplots.
  - Categorical-Numerical: `sns.boxplot()`
  - Numerical-Numerical: `sns.scatterplot()`, `sns.lineplot()`

MATHS 2
CHEATSHEET

Newton School
of Technology