Beginner's Handout
(Statistics for Data Science)

1. What is Statistics?

Statistics is the science of collecting, analyzing, and understanding data.
In Data Science, it helps us answer questions like:
- What is the average sales in a store?
- How many people prefer tea over coffee?
- What is the chance of rain tomorrow?

👉 Think of statistics as the GPS for decision making – it helps you navigate through raw numbers to meaningful insights.

2. Types of Statistics

2.1 Descriptive Statistics – Summarizing data

- Mean (average)
- Median (middle value)
- Mode (most frequent value)
- Range (difference between max and min)
- Standard Deviation (how spread out the data is)

👉 Example: Exam scores = 70, 80, 85, 90, 95
- Mean = (70+80+85+90+95) ÷ 5 = 84
- Median = 85 (middle value)
- Mode = None (no repeats)
- Range = 95 – 70 = 25

2.2 Inferential Statistics – Making predictions

- Based on a small sample, we predict about the whole population.
- Example: Surveying 100 voters to predict results for 10,000 voters.

3. Types of Data

- Categorical Data (labels): Gender, City, Yes/No
- Numerical Data (numbers): Age, Salary, Marks
- Discrete Data (countable): Number of children
- Continuous Data (measurable): Height, Weight, Temperature

4. Probability – The Language of Uncertainty

- Probability = Chance of an event happening
- Always between 0 and 1
  - 0 = Impossible
  - 1 = Certain

👉 Example: Tossing a coin
- Probability of Heads = 0.5
- Probability of Tails = 0.5

5. Distributions (How Data is Spread)

1. Normal Distribution (Bell Curve)
   - Most values are around the average.
   - Example: Heights of people in a class.

2. Skewed Distribution
   - Data is not balanced – pulled to one side.
   - Example: Income distribution (few rich people pull the average).

6. Correlation & Causation

- Correlation → Relationship between two variables.
  - Example: Height ↑, Weight ↑ (positive correlation).
  - Example: Temperature ↑, Sales of sweaters ↓ (negative correlation).
- Causation → One variable actually causes the other.
  - Example: More practice → Better exam scores.

👉 Remember: Correlation ≠ Causation

7. Hypothesis Testing (Making Decisions)

- Used to test assumptions with data.
- Steps:
  1. State a hypothesis (e.g., "More than 60% of people like tea").
  2. Collect sample data.
  3. Use statistical test (p-value, t-test).
  4. Decide whether to accept or reject.

👉 Example: A company tests if a new ad campaign increased sales compared to the old one.

8. Key Statistics in Data Science

- Mean, Median, Mode → Understanding central tendency
- Variance & Standard Deviation → Understanding spread
- Probability & Distributions → Predicting outcomes
- Correlation & Regression → Finding relationships
- Hypothesis Testing → Validating assumptions

9. Example: Student Marks Dataset

Student	Marks
Aisha	85
Rahul	90
Meena	70
Arjun	95
Ali	80

From this dataset:

- Mean = 84
- Median = 85
- Range = 25
- Standard Deviation = shows how spread marks are around the mean

10. Tips for Beginners

- Focus on concepts, not formulas at first.
- Use Excel, Python, or calculators for calculations.
- Connect to real-life examples (shopping bills, sports scores, surveys).
- Practice small datasets.

✅ By now, you should know:

✓ Descriptive vs Inferential statistics

✓ Mean, Median, Mode, Range, Standard Deviation

✓ Probability basics

✓ Normal distribution and correlation

✓ Hypothesis testing