MATHS

By Rishabh Dubey

AGENDA

- What is stats
- Types of stats
- variables and types of variables
- Population vs Sampling
- Measures of central tendecies
- Measures of dispersions
- Percentiles and Quartiles
- Five number summary and boxplot
- Gaussian And Normal Distribution

Quiz Time

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

1 What is the primary data type in Pandas?

- A) List
- B) Dictionary
- C) DataFrame
- D) Tuple

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

2 Which function is used to read a CSV file into Pandas?

- A) pd.load_csv()
- B) pd.read_csv()
- C) pd.import_csv()
- D) pd.open_csv()

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

3 What does df.head(5) do?

- A) Displays the last 5 rows
- B) Displays the first 5 rows
- C) Shows column names
- D) Returns the shape of the DataFrame

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

4 Which method removes missing values from a DataFrame?

- A) dropna()
- B) fillna()
- C) replace_na()
- D) remove_na()

Descriptive Statistics: Understanding & Summarizing Data

WHAT IS STATISTICS?

📌 Statistics = The Science of Data

✔️ Collecting, analyzing, and interpreting data
✔️ Helps in decision-making & pattern recognition

🔹 **Example**: Businesses use statistics to understand customer behavior!

TYPES OF STATISTICS

- 📊 **Descriptive Statistics** → Summarizes data (Mean, Median, Mode, Variance)
- 📈 **Inferential Statistics** → Makes predictions from data (Hypothesis Testing, P-values)

📌 **Example:**
  - **Descriptive:** "Average height of students in a class is 5'6"**."**
  - **Inferential:** "Based on a sample, we predict 60% of voters prefer candidate A."

REAL-WORLD EXAMPLE

📺 Netflix & Statistics

- Netflix collects watch history (Descriptive Stats)
- It predicts your next favorite show (Inferential Stats)

💡 Where else do we see statistics in daily life?

POPULATION VS SAMPLE?

📌 **What is Population & Sample?**

👥 Population = Entire group (e.g., All students in a country)

👤 Sample = Subset of the group (e.g., 500 students surveyed)

📌 Why Sampling?

- Saves time & cost
- Helps in making predictions

SAMPLING TECHNIQUES

🎯 **How do we select samples?**

1️⃣ **Random Sampling –** Equal chance for everyone

2️⃣ **Stratified Sampling –** Group-based selection

3️⃣ **Cluster Sampling –** Selecting entire groups

💡 Example: Surveys, Elections, Market Research

PRACTICE TASK 2

💻 **Hands-on Task**

Use Python to create a random sample from a dataset!

```
import random
data = [10, 20, 30, 40, 50]
sample = random.sample(data, 3)
print(sample)
```

WHAT ARE VARIABLES?

◆ **Categorical Variables (Qualitative)**

✔️ Nominal – No order (Gender, Colors)

✔️ Ordinal – Ranked order (Education Level, Satisfaction Ratings)

◆ **Numerical Variables (Quantitative)**

✔️ Interval – No true zero (Temperature, Years)

✔️ Ratio – True zero (Salary, Age)

EXAMPLE - VARIABLE TYPES

📝 **Survey Data Example**
- Name – Nominal
- Satisfaction Rating (1-5) – Ordinal
- Age – Ratio
- Temperature in City – Interval

TASK - IDENTIFY VARIABLE TYPES

Variable	Example Data	Type?
Gender	Male/Female	?
Temperature	23°C, 30°C	?
Income	$50,000, $70,000	?

WHAT ARE MEAN, MEDIAN, AND MODE?

📊 Central Tendency = Finding the Center of Data

✓ Mean – Average value

✓ Median – Middle value

✓ Mode – Most frequent value

EXAMPLE - TEST SCORES

📌 **Test Scores: [60, 70, 75, 85, 90, 95, 100]**

- Mean: (Sum ÷ Total Numbers)
- Median: (Middle Value: 85)
- Mode: (Most Frequent: None in this case)

📌 **When to use Median?**

When data has outliers (e.g., income data!)

HANDS-ON TASK - COMPUTE CENTRAL TENDENCY

```python
import numpy as np
scores = [60, 70, 75, 85, 90, 95, 100]
print("Mean:", np.mean(scores))
print("Median:", np.median(scores))
print("Mode:", max(set(scores), key=scores.count))
```

WHY DISPERSION MATTERS?

📊 **Dispersion = How spread out data is**

✔️ Low dispersion → Data is consistent

✔️ High dispersion → Data is variable

📌 Example: Salaries in a company (Interns vs. Executives!)

VARIANCE & STANDARD DEVIATION

✓ **Variance (σ²)** – Measures data spread

✓ **Standard Deviation (σ)** – Square root of variance

📌 Formula for Variance:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

HANDS-ON TASK - COMPUTE SD

```python
import numpy as np
data = [10, 20, 30, 40, 50]
print("Standard Deviation:", np.std(data))
```

PERCENTILES & BOXPLOT

📊 **Percentiles help compare data points!**

✔️ 50th Percentile = Median

✔️ Boxplot = Visual representation of percentiles & outliers

📌 Example: Height percentiles in a school

HANDS-ON TASK - BOXPLOT IN PYTHON

```python
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(y=[10, 20, 30, 40, 50, 100])
plt.show()
```

NORMAL DISTRIBUTION & BELL CURVE

📊 **The 68-95-99.7 Rule**

✔️ **68%** of data falls within **1 SD**

✔️ **95%** of data falls within **2 SDs**

✔️ **99.7%** of data falls within **3 SDs**

📌 Example: Heights of People

SUMMARY & NEXT STEPS

We Learned:

- Mean, Median, Mode
- Variability, Standard Deviation
- Normal Distribution

WHY DISPERSION MATTERS?

📊 **Dispersion = How spread out data is**

✔️ Low dispersion → Data is consistent

✔️ High dispersion → Data is variable

📌 Example: Salaries in a company (Interns vs. Executives!)

THANK YOU