📊 STATISTICS FOR DATA SCIENCE

Section 2: Types of Statistics - Descriptive vs Inferential

🎯 Two Main Branches of Statistics

Statistics is divided into two main branches: Each answers different questions about data and serves different purposes in data analysis.

📈 Descriptive Statistics

Answers: "What does this data look like?"

📊 Inferential Statistics

Answers: "Can I trust this finding? Can I apply it to a larger group?"

📈 Descriptive Statistics

Definition: Descriptive statistics summarize and describe the main features of a dataset using numbers and visualizations. It answers: "What does this data look like?"

Why is this important & what is the use?

Before making predictions or conclusions, you need to understand your data first. Descriptive statistics gives you this understanding quickly.

Beginner Example: Imagine you have exam scores of 100 students. Descriptive statistics helps you answer:

- What's the average score? (Mean)
- What score did most students get? (Mode)
- What's the highest and lowest score? (Range)
- How spread out are the scores? (Standard Deviation)

What it tells about data:

- What's the typical value? (central tendency)
- How spread out are the values? (dispersion)
- Is the data symmetric or skewed?
- Where do most values fall?

Practical Example: *You're analyzing monthly sales data:*

- ***Mean sales:*** *$50,000 per month*
- ***Range:*** *$30,000 - $70,000*
- ***Most frequent sales amount:*** *$45,000 (Mode)*
- ***Middle value when sorted:*** *$52,000 (Median)*

This gives you a complete picture of your sales pattern without any predictions.

How it helps in decision-making:

If a company wants to know their typical customer spending, descriptive statistics quickly shows the average, range, and typical pattern without needing advanced analysis.

More Beginner Examples:

- **Restaurant Owner:** "What are my busiest hours?" (Descriptive statistics shows peak times)
- **Teacher:** "How did my class perform on the test?" (Average, highest, lowest scores)
- **Social Media Manager:** "What's the average engagement on my posts?" (Mean likes/comments)
- **Fitness Tracker:** "What's my average daily steps?" (Simple descriptive stat)

📊 Inferential Statistics

Definition: Inferential statistics uses sample data to make predictions or conclusions about an entire population. It answers: "Can I trust this finding? Can I apply it to a larger group?"

Why is this important & what is the use?

You can't always test every person (the entire population), so you test a smaller group (sample) and infer conclusions about the whole population.

Beginner Example: Imagine you want to know which phone brand is most popular in your city (population: 1 million people). You can't ask everyone, so you:

- Survey 500 people (sample)
- Find 60% prefer Brand A
- Use inferential statistics to say: "We're 95% confident that 55-65% of ALL city residents prefer Brand A"

What it tells about data:

- Is this finding real or just random luck?
- Does this pattern apply to the entire population?
- How confident can we be in our predictions?
- Is there a real difference between two groups?

Practical Example: *A company tests a new website design:*

- *Test with 1,000 visitors (sample)*
- *New design gets 15% more signups*
- *Inferential statistics says: "We're 95% confident the new design will increase signups by 10-20% for ALL visitors"*

How it helps in decision-making:

A pharmaceutical company tests a drug on 1000 people and finds 80% improvement. Inferential statistics tells them: "We can be 95% confident that this drug will help 75-85% of the entire population."

More Beginner Examples:

- **Movie Rating:** Based on 1000 reviews (sample), predict how ALL viewers will rate it
- **Election Poll:** Survey 2000 voters to predict election results for millions
- **Quality Control:** Test 100 products from a batch of 10,000 to ensure quality
- **Medical Study:** Test new treatment on 500 patients to recommend for millions

📋 Descriptive vs Inferential: Key Differences

Descriptive Statistics

- **Purpose:** Describe what's in your data
- **Question:** "What happened?"
- **Data Used:** Entire dataset
- **Tools:** Mean, median, mode, graphs
- **Output:** Summary of existing data
- **Example:** "Average salary is $65,000"
- **When to use:** First step in any analysis

Inferential Statistics

- **Purpose:** Make predictions about larger groups
- **Question:** "What will happen?"
- **Data Used:** Sample from population
- **Tools:** Hypothesis testing, confidence intervals
- **Output:** Predictions with confidence levels
- **Example:** "95% confident salary will be $60,000-$70,000"
- **When to use:** When you can't measure everyone

Simple Analogy:

Descriptive Statistics = Looking at a photo (You see exactly what's there)
Inferential Statistics = Looking through a keyhole (You see part and guess the whole room)

Complete Example - Restaurant Business:

- **Descriptive:** "Last month, we served 5000 customers with average bill of $25. Busiest day was Saturday."
- **Inferential:** "Based on last month's data, we predict next month will have 5200-5500 customers with 95% confidence."

📝 Key Takeaways:

- **Always start with descriptive statistics** to understand your data
- **Use inferential statistics** when you need to make predictions
- **Descriptive = Facts about your data**
- **Inferential = Predictions about larger groups**
- **You need descriptive stats before doing inferential stats**

Python Examples

Descriptive Statistics in Python:

```
import pandas as pd
import numpy as np

# Sample data: Monthly sales (in thousands)
sales = [45, 52, 48, 55, 50, 53, 47, 51, 49, 54, 46, 52]

# Descriptive Statistics
print("=== DESCRIPTIVE STATISTICS ===")
print(f"Mean Sales: ${np.mean(sales):.1f}K")
print(f"Median Sales: ${np.median(sales):.1f}K")
print(f"Minimum Sales: ${np.min(sales)}K")
print(f"Maximum Sales: ${np.max(sales)}K")
print(f"Sales Range: ${np.max(sales)-np.min(sales)}K")
print(f"Standard Deviation: ${np.std(sales):.1f}K")
```

Inferential Statistics in Python (Simple Example):

```
from scipy import stats

# Sample data: Test scores from two teaching methods
method_A = [78, 82, 85, 79, 83, 80, 81]  # 7 students
method_B = [75, 79, 77, 74, 78, 76, 80]  # 7 students

# Inferential: t-test to see if there's a real difference
t_stat, p_value = stats.ttest_ind(method_A, method_B)
```

```
print("=== INFERENTIAL STATISTICS ===")
print(f"T-statistic: {t_stat:.3f}")
print(f"P-value: {p_value:.3f}")

if p_value < 0.05:
    print("Conclusion: There IS a significant difference between methods")
else:
    print("Conclusion: No significant difference found")
```

Section 2 of Statistics for Data Science - Types of Statistics
Descriptive Statistics vs Inferential Statistics

Data Science with Vamsi

Next: Section 3 - Descriptive Statistics Deep Dive