Measure of Central Tendency

Q1. What is the difference between population mean and sample mean? Provide a real world example where each is used.

Population Mean (μ): Average of ALL individuals in the entire population.

Sample Mean (x̄): Average of a subset drawn from the population.

Examples:

- **Population Mean:** Average height of all 500 students in a school (you measure everyone)
- **Sample Mean:** Average height of 50 randomly selected students from that school (used to estimate the population mean)

The sample mean estimates the population mean when measuring everyone is impractical.

Q2. Given the dataset, Calculate the mean, median, and mode.

[7, 9, 5, 13, 5, 6, 8, 5, 9]

Answer: Mean: Sum = 7 + 9 + 5 + 13 + 5 + 6 + 8 + 5 + 9 = 67 Mean = 67 ÷ 9 = **7.44**

Median: Ordered dataset: [5, 5, 5, 6, 7, 8, 9, 9, 13] Middle value (5th position) = **7**

Mode: 5 appears 3 times (most frequent) Mode = **5**

Q3. A dataset contains these salaries (in $1,000s): [35, 42, 38, 40, 39, 41, 120] a) Calculate the mean, median, and mode. b) Explain which measure best represents the central tendency and why.

Answer: a) Calculations:

Mean: Sum = 35 + 42 + 38 + 40 + 39 + 41 + 120 = 355 Mean = 355 ÷ 7 = **$50.7k**

Median: Ordered dataset: [35, 38, 39, 40, 41, 42, 120] Middle value (4th position) = **$40k**

Mode: No value repeats Mode = **None**

b) Best measure:

The **median ($40k)** best represents central tendency because:

- The dataset has an extreme outlier ($120k) that heavily skews the mean upward

- Six of the seven salaries fall between $35-42k, making $40k more representative of typical salaries
- The mean ($50.7k) is misleadingly high due to one unusually high salary
- In skewed distributions with outliers, the median is more robust and better reflects where most data points cluster

Q4. Explain when the median is a better measure than the mean in data analysis.

Answer: Median is better than mean when:

1. Outliers present - Extreme values skew the mean but don't affect median **2. Skewed data** - Mean gets pulled toward the tail, median stays central **3. Ordinal data** - Rankings where differences aren't meaningful

Examples:

- House prices (few expensive homes skew mean up)
- Income data (wealthy individuals create skew)
- Customer satisfaction ratings

Key point: Median shows the "typical" value when extreme values make the mean misleading.

Q5. A school has the following ages of students:
[13, 14, 13, 15, 13, 14, 16, 14, 13, 15]
Find the mode and explain what it indicates in this context.

Answer: Dataset: [13, 14, 13, 15, 13, 14, 16, 14, 13, 15]

Finding the Mode:

- 13 appears **4 times**
- 14 appears 3 times
- 15 appears 2 times
- 16 appears 1 time

Mode = 13

What it indicates: The mode of 13 tells us that **13 years old is the most common age** among students in this school. This suggests:

- The largest group of students are 13-year-olds
- This might indicate the school's primary grade level or most represented class
- Useful for planning age-appropriate activities, resources, or curriculum focus
- Helps understand the school's demographic composition

In educational contexts, knowing the modal age helps administrators make decisions about programs, facilities, and resources that will benefit the largest number of students.

Q6. Write your own logic (without using any built-in function or library) to calculate the mean of a list of 10 numbers.

Answer: Logic Breakdown:

1. **Initialize sum to 0**
2. **Loop through each number** (0 to 9 indices)
3. **Add each number to running sum**
4. **Divide total sum by 10** to get mean

Key Points:

- No built-in `sum()`, `len()`, or `mean()` functions used
- Manual iteration through all 10 elements
- Basic arithmetic operations only (addition and division)
- Shows step-by-step calculation process

The code demonstrates two approaches (for loop and while loop) and includes a detailed example showing each step of the summation process.

Q7. Write your own logic to calculate the median of a list of 7 numbers sorted in ascending order.

Answer: Logic Breakdown for 7 Sorted Numbers:

1. **Identify middle position**: For 7 numbers, middle = 4th number (index 3)
2. **Extract middle value**: `median = numbers[3]`
3. **Return result**

Key Concepts:

- **Odd count (7)**: Median is exactly the middle number
- **Position formula**: Middle index = (count-1) ÷ 2 = (7-1) ÷ 2 = 3
- **0-based indexing**: 7 numbers have indices 0,1,2,3,4,5,6

Why this works:

- 3 numbers below the median: positions 0, 1, 2
- 1 number at median: position 3
- 3 numbers above the median: positions 4, 5, 6

Since the list is already sorted, we simply access the element at the calculated middle index without any sorting logic needed.

Q8. Write your own logic to find the mode of a list of 10 numbers.

Answer: Logic Breakdown for Finding Mode:

Step 1: Count Frequencies

- Loop through each number in the list
- For each number, check if already counted
- If new: add to frequency table with count = 1
- If exists: increment its count

Step 2: Find Maximum Frequency

- Loop through frequency table
- Track highest count and corresponding number
- Return the number with highest frequency

Key Points:

- **No built-in functions** like `count()`, `max()`, or `Counter()`
- **Manual frequency counting** using nested loops
- **Handles multiple occurrences** correctly
- **Shows step-by-step process** for understanding

Algorithm Logic:

1. Create empty frequency table
2. For each number, search if already in table
3. If found: increment count, if not: add with count 1
4. Find entry with maximum count
5. Return that number as the mode

The code includes both a compact version and a detailed version that shows each counting step for educational purposes.

Q9. Given this data: [10, 20, 30, 40, 100], calculate:
- Mean
- Median
- Mode (if any)
Which measure best represents the data and why?

Answer: Dataset: [10, 20, 30, 40, 100]

Calculations:

Mean: Sum = 10 + 20 + 30 + 40 + 100 = 200 Mean = 200 ÷ 5 = **40**

Median: Already sorted: [10, 20, 30, 40, 100] Middle value (3rd position) = **30**

Mode: No value repeats Mode = **None**

Best Measure:

The **median (30)** best represents this data because:

- The dataset has an outlier (100) that pulls the mean upward
- Four of five values (10, 20, 30, 40) cluster in the lower range
- The mean (40) suggests the "typical" value is 40, but only one data point actually reaches 40
- The median (30) better reflects where most of the data is concentrated
- In right-skewed distributions like this, the median provides a more accurate picture of the central tendency

The large gap between 40 and 100 creates skewness that makes the mean misleading, while the median stays anchored to the center of the bulk of the data.

Which measure best represents the data and why?
10. Q10. A set of test scores are: [45, 48, 52, 60, 100]. Calculate:
a) Mean
b) Median
c) Comment on how outliers affect the mean.

Answer: Dataset: [45, 48, 52, 60, 100]

a) Mean: Sum = 45 + 48 + 52 + 60 + 100 = 305 Mean = 305 ÷ 5 = **61**

b) Median: Already sorted: [45, 48, 52, 60, 100] Middle value (3rd position) = **52**

c) How outliers affect the mean:

The outlier (100) significantly **inflates** the mean:

- **Without outlier [45, 48, 52, 60]**: Mean = 205 ÷ 4 = 51.25
- **With outlier [45, 48, 52, 60, 100]**: Mean = 305 ÷ 5 = 61

Impact:

- The single high score (100) pulls the mean up by nearly 10 points
- Mean (61) suggests average performance, but 4 out of 5 students scored below 61
- Median (52) remains stable and better represents typical performance
- The outlier creates a **misleading impression** of overall class performance

Key takeaway: Outliers disproportionately affect the mean because every value contributes equally to the calculation, making the median more reliable when extreme values are present.

Q11. You are analyzing customer purchase amounts on an e-commerce platform. The data is highly skewed. Which measure (mean, median, mode) would you report as the central value and why?

Answer: Answer: Median

Reasoning:

In highly skewed e-commerce purchase data, the **median** is the best measure because:

1. Outlier Resistance

- A few high-value purchases (luxury items, bulk orders) create extreme outliers
- Mean gets pulled toward these expensive purchases, inflating the "typical" purchase amount
- Median stays at the true center where most customers actually spend

2. Business Reality

- Most customers make small-to-moderate purchases
- Few customers make very large purchases
- Median reflects what the "typical customer" actually spends

3. Practical Example:

- Purchase amounts: [15, 25, 30, 35, 40, 45, 50, 2000]
- Mean = $280 (misleadingly high due to one $2000 purchase)
- Median = $37.50 (represents typical customer spending)

4. Business Decision Making

- Marketing budgets and inventory planning should target typical customers
- Median helps identify the customer segment that drives volume
- More actionable for pricing strategies and promotions

When to mention mean: Include it as supplementary information to show the impact of high-value customers on total revenue, but emphasize that median represents typical customer behavior.

The median tells you what most customers actually spend, while the mean tells you about total revenue distribution.

Q12. In a dataset of student exam scores, the mean is 55 but the median is 70. What might be the reason? What does this indicate about the distribution?

Answer: When the mean (55) is significantly lower than the median (70) in a dataset of student exam scores, this indicates a **left-skewed or negatively skewed distribution**. This pattern occurs when there are unusually low scores that pull the mean downward while the median remains at the center of the bulk of the data. The most likely reason is that a small number of students performed very poorly on the exam, creating outliers on the low end of the score distribution. For example, while most students might have scored between 60-85, a few students may have scored 10-30, which dramatically reduces the mean but doesn't affect the median since it represents the middle value when all scores are arranged in order.

This distribution pattern indicates several important things about the class performance. First, the majority of students actually performed reasonably well, as evidenced by the median of 70, suggesting that most students understood the material adequately. However, there's a concerning subset of students who struggled significantly, creating a achievement gap within the class. This could result from various factors such as students missing key foundational concepts, attendance issues, different preparation levels, or the exam containing some extremely difficult questions that only affected certain students. From a teaching perspective, this suggests the need for targeted intervention for the lower-performing students while recognizing that the overall class understanding is actually better than the mean would suggest. The median of 70 provides a more accurate picture of typical student performance in this scenario.

Q13. A company wants to decide the average number of support tickets handled per day by employees. What measure should be used and why?

Answer: Mean

A company should use the **mean** for support tickets because:

- Represents total workload across all employees
- Every ticket handled contributes equally to company productivity
- Needed for resource planning and staffing decisions
- Support ticket data is typically normally distributed without extreme outliers
- Shows true organizational capacity rather than just middle performance
- Essential for setting realistic daily targets and performance standards

The mean captures the complete picture of work output, which is crucial for operational planning and workforce management.

Q14. A dataset contains a bimodal distribution. What does mode reveal in this case? Give an example.

Answer: Bimodal distribution mode reveals: Two distinct peaks indicating **two separate groups or populations** in the data.

Example: Employee commute times: [15, 16, 17, 45, 46, 47] minutes

- Mode 1: ~16 minutes (employees living nearby)
- Mode 2: ~46 minutes (employees from distant suburbs)

What it shows:

- Two different employee populations based on location
- Single average would miss this important pattern
- Need different strategies for each group (flexible hours for long commuters)

The mode reveals natural clustering that helps identify distinct subgroups within the dataset.

Q15. You have a dataset where mean = 30, median = 45, and mode = 50.
a) Sketch a rough distribution shape (left-skewed, right-skewed, symmetric)
b) Interpret what this says about the data spread.

Answer: a) Distribution Shape: Left-skewed (Negatively skewed)

```
  Frequency
     |
     |     *
     |    ***
     |   *****  ← Mode = 50 (peak)
     |  *******
     | *********  ← Median = 45
     |***********
     |*************
     |***************
     +------------------→ Values
   Mean=30       Mode=50
```

b) Data Spread Interpretation:

This distribution shows that **most data points cluster around the higher values** (near the mode of 50), but there are some **extremely low outliers** that pull the mean down to 30. The pattern Mean < Median < Mode is characteristic of left-skewed data.

What this reveals:

- **Majority of values** are concentrated in the 45-50 range
- **Few extreme low values** (outliers) drag the mean significantly downward
- **Typical value** is better represented by median (45) or mode (50) than mean (30)
- **Long tail extends leftward** toward lower values

Real-world example: This could represent income data where most people earn decent wages (mode = 50k), but a few unemployed or part-time workers (creating the low outliers) bring the average down to 30k, while the median stays at 45k representing typical earning.