Seminar 2

Calculating the Mode and the Median

Mode

The **mode** is the value or values that appear most frequently in a dataset. It is the most common value(s) in the data.

Steps to calculate the mode:

1. **Sort the data** (optional): Sorting the data helps in visualizing the frequency of each value.
2. **Count the frequency** of each unique value in the dataset.
3. **Identify the value(s) that occur the most frequently**: The mode is the value(s) with the highest frequency.

Example:
Consider the following data representing the number of books read by students:

$$3, 4, 4, 6, 7, 4, 8, 6, 3, 9$$

- **Step 1:** Sort the data (optional): $3, 3, 4, 4, 4, 6, 6, 7, 8, 9$
- **Step 2:** Count the frequency of each value:
  - 3 appears twice
  - 4 appears three times
  - 6 appears twice
  - $7, 8, 9$ each appear once
- **Step 3:** The value 4 appears the most (three times), so the **mode** is 4.

Special Cases:

- If all values occur with the same frequency or only once, the dataset has **no mode**.
- If more than one value occurs with the highest frequency, the dataset is **multimodal**.

Median

The **median** is the middle value of a dataset when it is ordered from smallest to largest. It separates the dataset into two equal halves.

Steps to calculate the median:

1. **Order the data** from smallest to largest.
2. If the number of observations n is **odd**, the median is the value at position $\frac{n+1}{2}$ in the ordered data.
3. If the number of observations n is **even**, the median is the average of the two middle values at positions $\frac{n}{2}$ and $\frac{n}{2} + 1$.

Example 1 (Odd number of observations):
Data: $5, 3, 9, 6, 4$

- **Step 1:** Order the data: $3, 4, 5, 6, 9$
- **Step 2:** The number of observations $n = 5$, which is odd. The median is at position $\frac{5+1}{2} = 3$.
- **Step 3:** The median is the 3rd value, which is 5.

Example 2 (Even number of observations):
Data: $2, 7, 4, 9, 3, 6$

- **Step 1:** Order the data: $2, 3, 4, 6, 7, 9$
- **Step 2:** The number of observations $n = 6$, which is even. The median is the average of the 3rd and 4th values: 4 and 6.
- **Step 3:** The median is:
$$\text{Median} = \frac{4 + 6}{2} = 5$$

Special Cases:
If there are repeated values, the steps for finding the median remain the same. You just use the positions in the ordered data, regardless of whether some values are repeated.

Summary

- **Mode**: The most frequent value(s) in the dataset.
- **Median**: The middle value in an ordered dataset. For an odd number of observations, it is the middle value, and for an even number, it is the average of the two middle values.

Problem 1: Basic Descriptive Statistics

You are given the following dataset representing the ages of participants in a study:

$$18, 21, 24, 20, 19, 23, 25, 18, 22, 20, 24, 26, 19, 21, 22$$

1. **Mean**:
$$\text{Mean} = \frac{18 + 21 + 24 + 20 + 19 + 23 + 25 + 18 + 22 + 20 + 24 + 26 + 19 + 21 + 22}{15} = 21.47$$
2. **Median**:
$$\text{Ordered data} : 18, 18, 19, 19, 20, 20, 21, 21, 22, 22, 23, 24, 24, 25, 26$$
$$\text{Median} = 21 \quad \text{(middle value of ordered data)}$$
3. **Mode**:
$$\text{Mode} = 18, 20, 21, 22 \quad \text{(multiple modes)}$$
4. **Range**:
$$\text{Range} = 26 - 18 = 8$$
5. **Interpretation**: The range of 8 years indicates that the ages of participants span 8 years, which gives an initial idea of the spread of the data.

6. **Variance**:

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

We now compute the squared deviations from the mean for each data point:

$$(18 - 21.47)^2 = 12.05, \quad (21 - 21.47)^2 = 0.22, \quad (24 - 21.47)^2 = 6.39$$

$$(20 - 21.47)^2 = 2.16, \quad (19 - 21.47)^2 = 6.09, \quad (23 - 21.47)^2 = 2.34$$

$$(25 - 21.47)^2 = 12.45, \quad (18 - 21.47)^2 = 12.05, \quad (22 - 21.47)^2 = 0.28$$

$$(20 - 21.47)^2 = 2.16, \quad (24 - 21.47)^2 = 6.39, \quad (26 - 21.47)^2 = 20.57$$

$$(19 - 21.47)^2 = 6.09, \quad (21 - 21.47)^2 = 0.22, \quad (22 - 21.47)^2 = 0.28$$

Adding these up:

$$\sum(x_i - \bar{x})^2 = 89.45$$

Finally, divide by $n - 1 = 14$:

$$\text{Variance} = \frac{89.45}{14} = 6.39$$

7. **Interpretation**: A higher variance indicates a greater spread in the data. A variance of 6.39 suggests that the data points tend to deviate from the mean age (21.47 years) to a certain extent.

8. **Standard Deviation**:

$$\text{Standard Deviation} = \sqrt{6.39} = 2.53$$

9. **Interpretation**: The standard deviation of 2.53 years indicates that, on average, the ages of participants deviate from the mean age of 21.47 years by approximately 2.53 years. This shows moderate variability in the ages of the participants.

10. **Overall interpretation**: The mean age is 21.47 years, the data is spread across a range of 8 years, and the standard deviation shows moderate variability.

Problem 2: Percentiles and Quartiles

Consider the following test scores of 20 students:

$$45, 67, 78, 88, 54, 72, 61, 80, 92, 53, 77, 84, 69, 65, 70, 89, 90, 74, 81, 73$$

1. **Q1 (25th Percentile)**: 69
   Median (50th Percentile): 74.5
   Q3 (75th Percentile): 84.5

2. **Interquartile Range (IQR)**:

$$\text{IQR} = Q3 - Q1 = 84.5 - 69 = 15.5$$

3. **Outliers**: Using 1.5 × IQR rule, outliers are values lower than $69 - 1.5 \times 15.5 = 45.75$ or higher than $84.5 + 1.5 \times 15.5 = 107.75$. No outliers are detected.

Percentiles are used to understand the relative standing of a value within a dataset. The k-th percentile is the value below which $k\%$ of the data falls. Given the dataset:

$$45, 67, 78, 88, 54, 72, 61, 80, 92, 53, 77, 84, 69, 65, 70, 89, 90, 74, 81, 73$$

Step 1: Sort the Data

First, we need to sort the dataset in ascending order:

$$45, 53, 54, 61, 65, 67, 69, 70, 72, 73, 74, 77, 78, 80, 81, 84, 88, 89, 90, 92$$

Step 2: Calculate the Median (Q2)

The median (Q2) is the middle value of the dataset. If the number of observations (n) is odd, the median is the middle number. If n is even, the median is the average of the two middle numbers.

For our dataset with $n = 20$ (even):

$$\text{Median} = \frac{\text{Value at position 10} + \text{Value at position 11}}{2} = \frac{73 + 74}{2} = 73.5$$

Step 3: Calculate the First Quartile (Q1)

The first quartile (Q1) is the 25th percentile, which separates the lowest 25% of the data from the rest. To find the position of Q_1 in a dataset with $n = 20$ observations, we calculate the position P_{Q_1} using the formula:

$$P_{Q_1} = \frac{(n+1) \times 25}{100}$$

Substituting $n = 20$:

$$P_{Q_1} = \frac{(20+1) \times 25}{100} = \frac{21 \times 25}{100} = 5.25$$

Since the position is not a whole number, we average the values at the 5th and 6th positions in the sorted dataset: - The sorted dataset is:

$$45, 53, 54, 61, 65, 67, 69, 70, 72, 73, 74, 77, 78, 80, 81, 84, 88, 89, 90, 92$$

- The 5th value is 65 and the 6th value is 67:

$$Q_1 = \frac{65 + 67}{2} = \frac{132}{2} = 66$$

Step 4: Calculate the Third Quartile (Q3)

The third quartile Q_3 is the 75th percentile, which separates the lowest 75% of the data from the highest 25%. To find the position of Q_3, we use the formula:

$$P_{Q_3} = \frac{(n+1) \times 75}{100}$$

Substituting $n = 20$:

$$P_{Q_3} = \frac{(20+1) \times 75}{100} = \frac{21 \times 75}{100} = 15.75$$

Since the position is not a whole number, we average the values at the 15th and 16th positions in the sorted dataset: - The sorted dataset is:

$$45, 53, 54, 61, 65, 67, 69, 70, 72, 73, 74, 77, 78, 80, 81, 84, 88, 89, 90, 92$$

- The 15th value is 81 and the 16th value is 84:

$$Q_3 = \frac{81 + 84}{2} = \frac{165}{2} = 82.5$$

Percentile Calculation Methods

1. **Using $n + 1$ Method**

   The formula:

$$P_k = \frac{(n+1) \cdot k}{100}$$

is commonly used when calculating percentiles in statistics, particularly in some textbooks and statistical software. The idea behind using $n + 1$ is to account for the fact that we want to create a more evenly distributed dataset, especially for smaller datasets. The +1 adjusts the rank position to ensure that the percentile calculation represents the entire range of data.

Example: For the 25th percentile (Q1) in a dataset of 20 observations, the calculation would be:

$$P_{Q_1} = \frac{(20+1)\cdot 25}{100} = \frac{21\cdot 25}{100} = 5.25$$

This means that the 25th percentile falls between the 5th and 6th values of the sorted dataset.

2. **Using n Method**

The formula:

$$P_k = \frac{n\cdot k}{100}$$

is often used in practice, particularly when you are directly looking for the rank of a percentile without the additional adjustment for small datasets. This formula tends to yield results that are similar but may not provide the same precision for smaller datasets.

Example: Using the same dataset of 20 observations for the 25th percentile, we would calculate:

$$P_{Q_1} = \frac{20\cdot 25}{100} = 5$$

This suggests that Q1 is at the 5th position in the sorted dataset. If we want to be precise, we would still need to interpolate between values if necessary.

3. **Using $n(k+1)$ Method**

This method adjusts the percentile formula by multiplying n, the sample size, by $(k+1)$, where k is the desired percentile. The formula is as follows:

$$P_k = \frac{n(k+1)}{100}$$

This method provides a slight modification that is helpful for obtaining a rank position that takes both the dataset size and the percentile into account. It's especially useful in some domains like economics and social sciences.

Example: For the 25th percentile (Q1) in a dataset of 20 observations:

$$P_{Q_1} = \frac{20\cdot(25+1)}{100} = \frac{20\cdot 26}{100} = 5.2$$

This suggests that the 25th percentile falls between the 5th and 6th values of the sorted dataset.

- $n+1$: More traditional, often yields a better representation of the dataset for smaller samples.
- n: Simpler, more direct calculation, useful for larger datasets.
- $n(k+1)$: A modified version for specific use cases, providing an adjusted percentile rank that considers both sample size and the next percentile.

Practical Implications

The choice between these formulas can lead to slightly different results, especially in small datasets. In practical applications, either method can be used, but it's essential to stay consistent within your analysis. Different fields or texts might prefer one method over the other, so it's good to clarify which method you are using when reporting or interpreting percentiles.

Summary of Results

- Q1 (25th Percentile): 66
- Median (50th Percentile): 73.5
- Q3 (75th Percentile): 82.5

Interquartile Range (IQR)

The IQR is calculated as:

$$\text{IQR} = Q3 - Q1 = 82.5 - 66 = 16.5$$

Identifying Outliers

To determine outliers, we use the 1.5 times IQR rule:

- Lower Bound: $Q1 - 1.5 \times \text{IQR} = 66 - 1.5 \times 16.5 = 66 - 24.75 = 41.25$
- Upper Bound: $Q3 + 1.5 \times \text{IQR} = 82.5 + 1.5 \times 16.5 = 82.5 + 24.75 = 107.25$

Since all test scores are between 41.25 and 107.25, there are no outliers in this dataset.

Problem 3: Grouped Data

The table above shows the frequency distribution of the number of books read by a group of students over the summer:

Books Read	Frequency
0 - 2	5
3 - 5	8
6 - 8	12
9 - 11	6
12 - 14	3

Table 1: Books Read by Students

1. **Mean**: Midpoints are $1, 4, 7, 10, 13$.

$$\text{Mean} = \frac{(1 \times 5) + (4 \times 8) + (7 \times 12) + (10 \times 6) + (13 \times 3)}{5 + 8 + 12 + 6 + 3} = 6.55$$

2. **Mode**: 7 (the class $6 - 8$ has the highest frequency).
3. **Median**: The cumulative frequency crosses 50% at class $6 - 8$, so the median lies in that range.
4. **Variance**: Use formula for grouped data, results in Variance = 8.09 and Standard Deviation = $\sqrt{8.09} = 2.84$

Cumulative Frequency Summary

Cumulative frequency is a way to summarize the distribution of data points within a dataset. It represents the total number of observations that fall within or below a certain category or class interval.

The cumulative frequency CF for a given class can be calculated using the formula:

$$CF_i = \sum_{j=1}^{i} f_j$$

where: CF_i is the cumulative frequency for the i-th class, f_j is the frequency of the j-th class, i is the class index (1, 2, 3, ...).

The cumulative frequency for the above dataset is calculated as follows:

Books Read	Frequency	Cumulative Frequency
$0 - 2$	5	5
$3 - 5$	8	13
$6 - 8$	12	25
$9 - 11$	6	31
$12 - 14$	3	34

The cumulative frequency shows the total number of students who read a certain number of books or fewer. For instance, the cumulative frequency of 25 for the class $6 - 8$ indicates that 25 students read between 0 and 8 books in total. Cumulative frequency is particularly useful for calculating percentiles and medians in a dataset.

Problem 4: Box Plot and Interpretation

The following data represents the weekly hours spent studying by a group of students:

$$4, 6, 8, 3, 5, 7, 9, 10, 2, 6, 8, 5, 6, 11, 7$$

1. **Box plot**: The data would give us:
   - Q1 = 5
   - Median = 6
   - Q3 = 8
   - No outliers
2. **Interpretation**: The distribution is slightly skewed to the right with no significant outliers.

To determine the skewness of a distribution based on a box plot and summary statistics, we analyze the quartiles and median.

1. Box Plot Structure:
- A box plot shows quartiles (Q1, median, Q3), the interquartile range (IQR), and potential outliers.

2. Skewness Determination:
- Right (Positive) Skew: Median is closer to Q1; the upper whisker is longer than the lower whisker.
- Left (Negative) Skew: Median is closer to Q3; the lower whisker is longer than the upper whisker.
- Symmetrical: Median is centered between Q1 and Q3; whiskers are of similar length.

For the dataset:

$$4, 6, 8, 3, 5, 7, 9, 10, 2, 6, 8, 5, 6, 11, 7$$

- Q1 = 5, Median = 6, Q3 = 8.

Since: - The distance from Q1 to the median (1 unit) is less than the distance from the median to Q3 (2 units), and if the upper whisker is longer than the lower whisker, the distribution is slightly skewed to the right.

Problem 5: Visualization and Summary Measures

The following dataset represents the heights (in cm) of students in a class:

$$162, 167, 170, 155, 165, 160, 168, 172, 163, 159, 166, 169, 173, 161, 164$$

1. **Histogram**: Would show a fairly normal distribution.
2. **Stem-and-leaf plot**: Simple plot showing the structure of the data (not feasible to depict in this text).
3. **Mean**:
$$\text{Mean} = \frac{162 + 167 + 170 + \cdots + 164}{15} = 165.87$$
4. **Standard Deviation**:
$$\text{Standard Deviation} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = 5.11$$

Problem 6: Comparison of Two Data Sets

Two different groups of students take two versions of a test. Their scores are as follows:

Group A: 85, 78, 92, 88, 75, 90, 95, 80, 85, 87
Group B: 72, 65, 80, 75, 82, 68, 78, 83, 70, 76

1. **Group A**: Mean = 85.5, Variance = 55.39, Standard Deviation = 7.44
   Group B: Mean = 74.9, Variance = 27.21, Standard Deviation = 5.22
2. **Variability**: Group A has more variability because of its higher standard deviation.
3. **Performance**: Group A performed better on average.

Problem 7: Real-World Data Analysis

The monthly income (in dollars) of employees at a small company is given below:

$$3000, 3500, 4000, 3800, 4200, 3900, 4500, 4300, 5000, 5500, 6000, 7000$$

1. **Mean**:
$$\text{Mean} = \frac{3000 + 3500 + \cdots + 7000}{12} = 4650$$
2. **Median**: The middle two values are 4300 and 4500, so the median is:
$$\text{Median} = \frac{4300 + 4500}{2} = 4400$$
3. **Outlier Effect**: If the CEO earning $20,000 is added:
$$\text{New Mean} = \frac{3000 + 3500 + \cdots + 7000 + 20000}{13} = 5653.85$$
The median remains the same at 4400, but the mean is heavily influenced by the outlier.

Discussion Questions

1. Skewness affects the relationship between the mean and median because in a skewed distribution, the mean is pulled in the direction of the tail, while the median is more resistant to extreme values.
2. The median is a better measure of central tendency when there are outliers or a skewed distribution, as it is not influenced by extreme values.
3. The range only considers the two most extreme values and ignores the distribution of data points, while the standard deviation provides a more comprehensive measure of spread around the mean.