Seminar 1 - 2025

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.

Problem 1. Compute the mean and the median for the following data set:

20	34	45	60	25	38	40	28	32	10
55	22	47	50	33	15	26	30	42	18
35	39	54	29	46	27	12	24	31	14

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.

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**.

Problem 2. Compute the median for the following two continuous distributions:

a. $Unif(a, b)$

b. $Exp(\lambda)$.

Problem 3.

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

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

Find the mode.

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

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

Compute the mode(s) and the variance.

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.

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.

Problem 4 Consider the following test scores of 20 students:

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

Compute $Q1, Q2, Q3, IQR$ and check for outliers.

Problem 5

A survey was conducted among 50 individuals, asking them how many hours per week they spend on social media. The data were recorded in grouped form, with the frequency distribution given below:

Class interval (hours/week)	**Frequency**
0–4	6
5–9	10
10–14	12
15–19	14
20–24	6
25–29	2

On the basis of these data, compute the following:

1. The sample mean $\bar{x}$ (use class midpoints).
2. The median (via grouped data interpolation).
3. The mode (using the grouped data formula).
4. The variance s^2 and standard deviation s.
5. The 70th percentile.

Working with frequencies and grouped data is slightly different:

Class Interval (hours/week)	Frequency f_i
0–4	6
5–9	10
10–14	12
15–19	14
20–24	6
25–29	2

Total frequency: $N = 50$

For grouped data, the mean is approximated using *class midpoints*:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

where x_i is the midpoint of the i-th class:

$$x_i = \frac{\text{lower limit} + \text{upper limit}}{2}$$

Class Interval	Midpoint x_i	Frequency f_i	$f_i x_i$
0–4	2	6	12
5–9	7	10	70
10–14	12	12	144
15–19	17	14	238
20–24	22	6	132
25–29	27	2	54

$\sum f_i x_i = 650$

$$\bar{x} = \frac{650}{50} = 13$$

For grouped data, the median is estimated using:

$$\text{Median} = L + \frac{\frac{N}{2} - CF}{f_m} \cdot h$$

where:

- L = lower boundary of the median class
- CF = cumulative frequency before the median class
- f_m = frequency of the median class
- h = class width

Identify the median class

$$\frac{N}{2} = \frac{50}{2} = 25$$

Class Interval	f	Cumulative CF
0–4	6	6
5–9	10	16
10–14	12	28
15–19	14	42
20–24	6	48
25–29	2	50

The median class is 10–14 with $L = 10$, $CF = 16$, $f_m = 12$, $h = 5$.

$$\text{Median} = 10 + \frac{25 - 16}{12} \cdot 5 = 10 + 3.75 = 13.75$$

Mode for grouped data:

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \cdot h$$

where:

- L = lower boundary of modal class
- f_1 = frequency of modal class
- f_0 = frequency of previous class
- f_2 = frequency of next class
- h = class width

Identify modal class
Modal class: 15–19, $f_1 = 14$, previous class $f_0 = 12$, next class $f_2 = 6$, $h = 5$.

$$\text{Mode} = 15 + \frac{14 - 12}{2 \cdot 14 - 12 - 6} \cdot 5 = 15 + \frac{2}{10} \cdot 5 = 16$$

Variance and standard deviation:

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{N - 1}, \quad s = \sqrt{s^2}$$

Class Interval	x_i	f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
0–4	2	6	-11	121	726
5–9	7	10	-6	36	360
10–14	12	12	-1	1	12
15–19	17	14	4	16	224
20–24	22	6	9	81	486
25–29	27	2	14	196	392

$\sum f_i(x_i - \bar{x})^2 = 2200$

$$s^2 = \frac{2200}{49} \approx 44.90, \quad s \approx 6.70$$

Standard deviation: $s \approx 6.70$
Percentile:

$$Q_k = L + \frac{\frac{k}{100} N - CF}{f} \cdot h$$

Identify percentile class

$$\frac{70}{100} \cdot 50 = 35$$

Cumulative frequencies show 35 lies in 15–19, with $L = 15$, $CF = 28$, $f = 14$, $h = 5$.

$$Q_{70} = 15 + \frac{35 - 28}{14} \cdot 5 = 15 + 2.5 = 17.5.$$