To build a Frequency Table:

Example:

From below data construct frequency table and draw histogram:

39, 38, 47, 44, 42, 65, 49, 55, 49, 36, 34, 46, 31, 53, 60, 48, 33, 38, 40.

1. **Arrangement data from smallest to largest value**
   31, 33, 34, 36, 38, 38, 39, 40, 42, 44, 46, 47, 48, 49, 49, 53, 55, 60, 65

2. **Determine the range of data.**
   R=largest value – smallest value

31, 33, 34, 36, 38, 38, 39, 40, 42, 44, 46, 47, 48, 49, 49, 53, 55, 60, 65

$$Range = 65 - 31$$

3. Determine the number of class interval

Using Sturge's Rule: A rule for determining number of classes to use in a histogram or frequency distribution table.

$$k = 1 + 3.322(\log_{10} n)$$

$$k = 1 + 3.322(\log_{10} 19)$$

$$k = 1 + 3.322(\log_{10} n)$$

$$k = 1 + 3.322(1.28)$$

$$k = 5.24 \approx 5$$

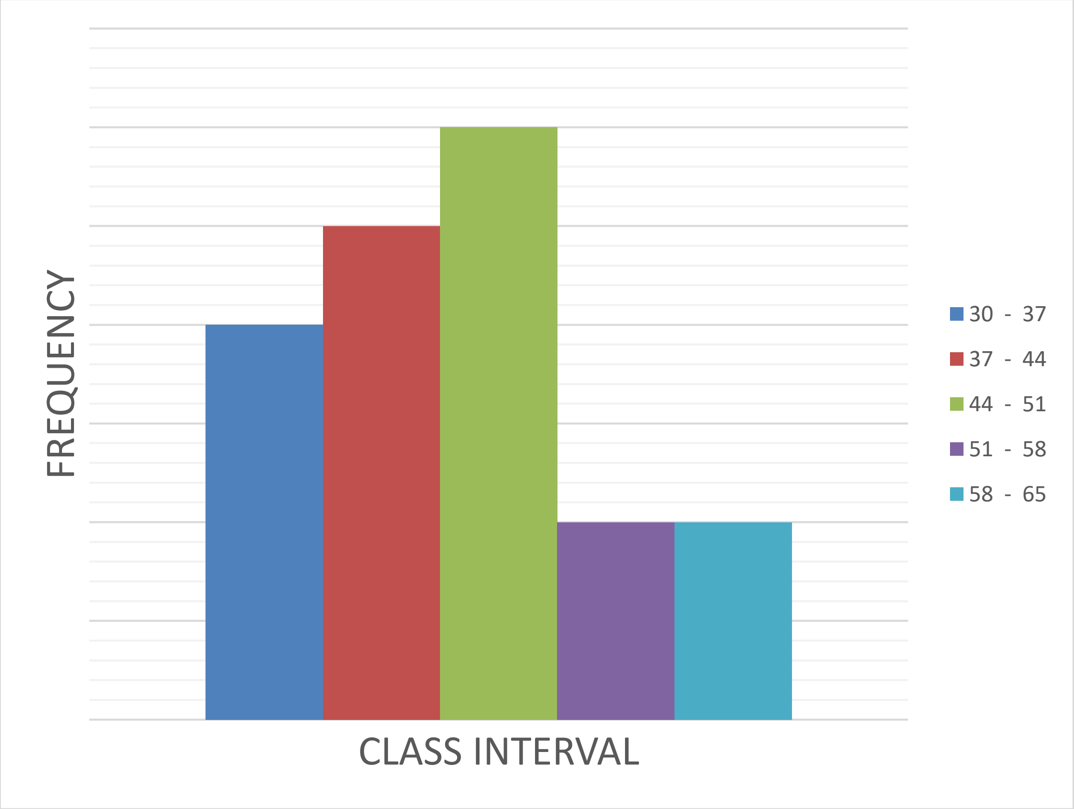
4. Width of class interval

$$\textit{width of class interval}$$

$$(w) = \frac{range\ (R)}{no.of\ class\ interval\ (k)}$$

$$w = \frac{34}{5} = 6.8 \approx 7$$

Class interval	frequency
30 - 37	4
37 - 44	5
44 - 51	6
51 - 58	2
58 - 65	2

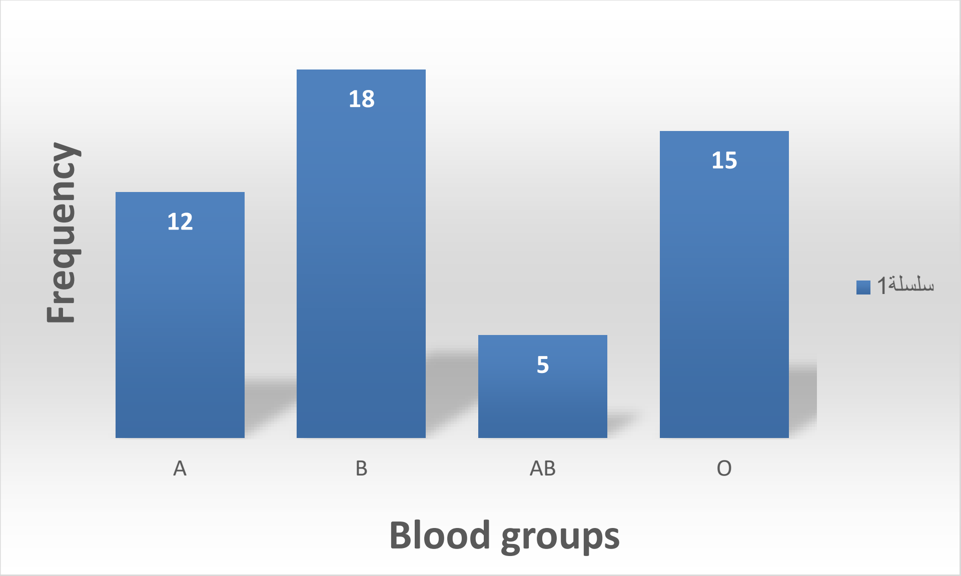


Example:

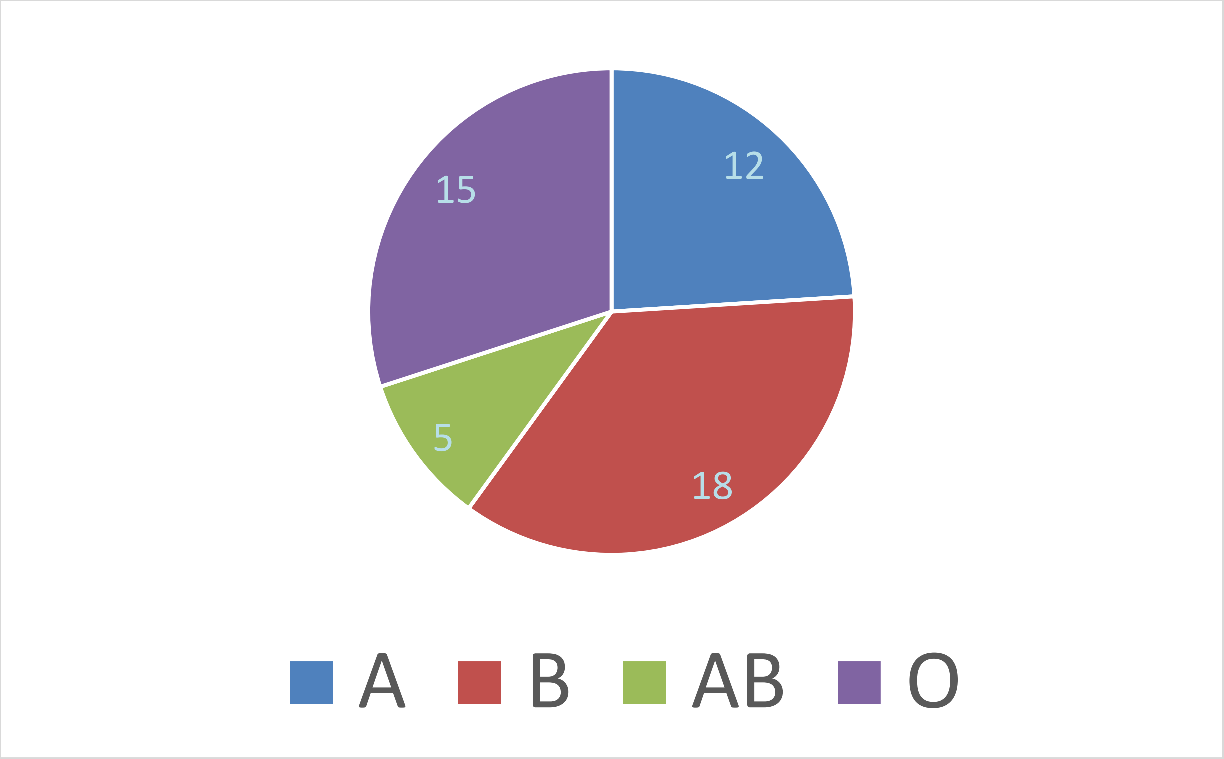
Distribution of **50 patients** at the surgical department in hospital according to their **ABO** blood groups.

Blood groups	Frequency
A	12
B	18
AB	5
O	15
Total	50

Draw bar chart.



Draw Pie chart



$$\text{Area of each sector} = \frac{\text{F}}{\text{n}} \text{x } 360^{\text{o}}$$

$$Area\ for\ A\ group = \frac{12}{50} x360 = 86.4$$

$$Area\ forB\ group = \frac{18}{50} x360 = 129.6$$

$$Area\ for\ AB\ group = \frac{5}{50} x360 = 36$$

$$Area\ for\ O\ group = \frac{15}{50} x360 = 108$$

A descriptive measure computed from the values of a sample is called a "statistic".

A descriptive measure computed from the values of a population is called a "parameter".

For the variable of interest there are:

(1) "N" population values.

(2) "n" sample of values.

- Let $X_1, X_2, \ldots, X_N$ be the population values (in general, they are unknown) of the variable of interest.
  The population size $= N$

- Let $x_1, x_2, \ldots, x_n$ be the sample values (these values are known).
  The sample size $= n$.

(i) A **parameter** is a measure (or number) obtained from the population values: $X_1, X_2, \ldots, X_N$.

- Values of the parameters are unknown in general.
- We are interested to know true values of the parameters.

(ii) A **statistic** is a measure (or number) obtained from the sample values: $x_1, x_2, \ldots, x_n$.

- Values of statistics are known in general.
- Since parameters are unknown, statistics are used to approximate (estimate) parameters.

Measures of Central Tendency:
(or measures of location):

The most commonly used measures of central tendency are: the mean – the median – the mode.

- The values of a variable often tend to be concentrated around the center of the data.
- The center of the data can be determined by the measures of central tendency.
- A measure of central tendency is considered to be a typical (or a representative) value of the set of data as a whole.

Mean:

(1) The Population mean (μ):

If $X_1, X_2, \ldots, X_N$ are the population values, then the population mean is:

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \text{(unit)}$$

- The population mean μ is a parameter (it is usually unknown, and we are interested to know its value)

(2) The Sample mean ($\bar{x}$):

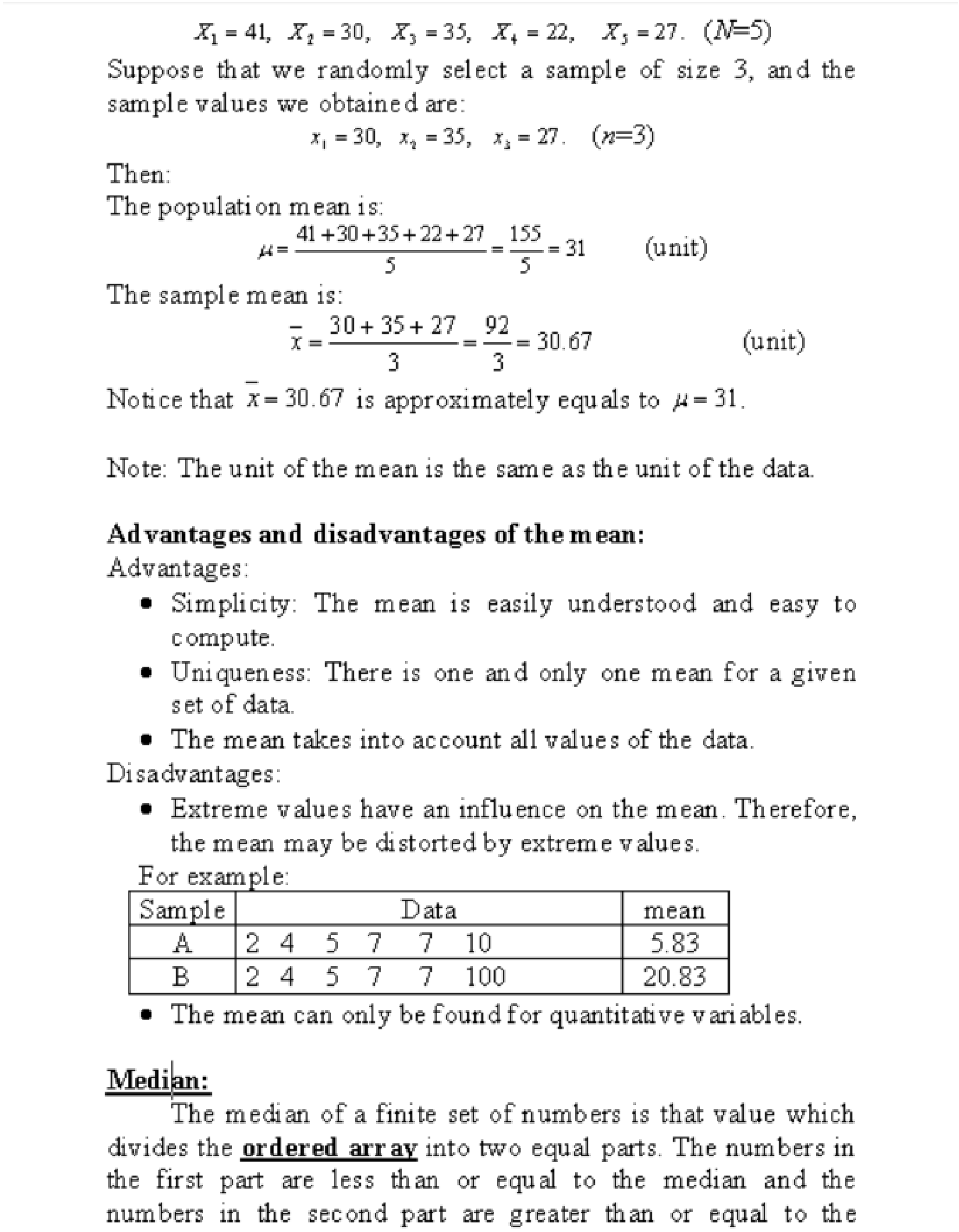
If $x_1, x_2, \ldots, x_n$ are the sample values, then the sample mean is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(unit)}$$

- The sample mean $\bar{x}$ is a statistic (it is known – we can calculate it from the sample).
- The sample mean $\bar{x}$ is used to approximate (estimate) the population mean μ.

Example:

Suppose that we have a population of 5 population values:



$$X_1 = 41,\quad X_2 = 30,\quad X_3 = 35,\quad X_4 = 22,\quad X_5 = 27.\quad (N=5)$$

Suppose that we randomly select a sample of size 3, and the sample values we obtained are:

$$x_1 = 30,\quad x_2 = 35,\quad x_3 = 27.\quad (n=3)$$

Then:

The population mean is:

$$\mu = \frac{41+30+35+22+27}{5} = \frac{155}{5} = 31 \qquad \text{(unit)}$$

The sample mean is:

$$\bar{x} = \frac{30+35+27}{3} = \frac{92}{3} = 30.67 \qquad \text{(unit)}$$

Notice that $\bar{x} = 30.67$ is approximately equals to $\mu = 31$.

Note: The unit of the mean is the same as the unit of the data.

Advantages and disadvantages of the mean:

Advantages:

- Simplicity: The mean is easily understood and easy to compute.
- Uniqueness: There is one and only one mean for a given set of data.
- The mean takes into account all values of the data.

Disadvantages:

- Extreme values have an influence on the mean. Therefore, the mean may be distorted by extreme values.

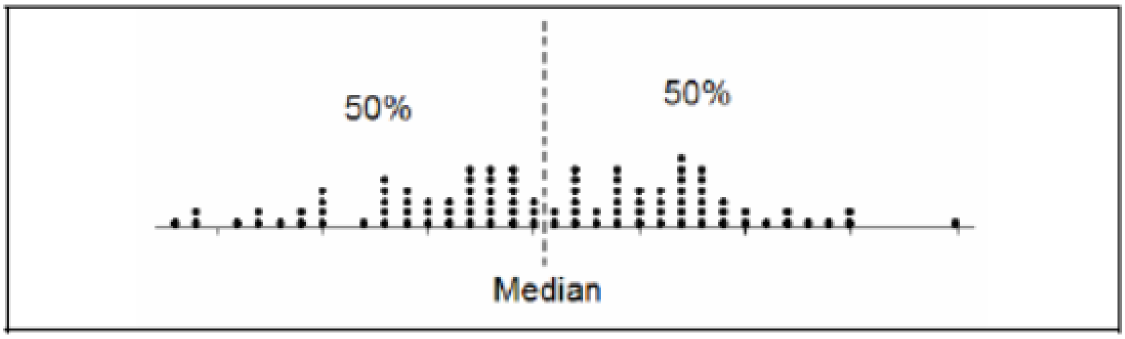
For example:

Sample	Data	mean
A	2 4 5 7 7 10	5.83
B	2 4 5 7 7 100	20.83

- The mean can only be found for quantitative variables.

Median:

The median of a finite set of numbers is that value which divides the **ordered array** into two equal parts. The numbers in the first part are less than or equal to the median and the numbers in the second part are greater than or equal to the median.



Notice that:

50% (or less) of the data is $\leq$ Median

50% (or less) of the data is $\geq$ Median

Calculating the Median:

Let $x_1, x_2, \ldots, x_n$ be the sample values. The sample size (n) can be odd or even.

- First we order the sample to obtain the ordered array.
- Suppose that the ordered array is:

$$y_1, y_2, \ldots, y_n$$

- We compute the rank of the middle value (s):

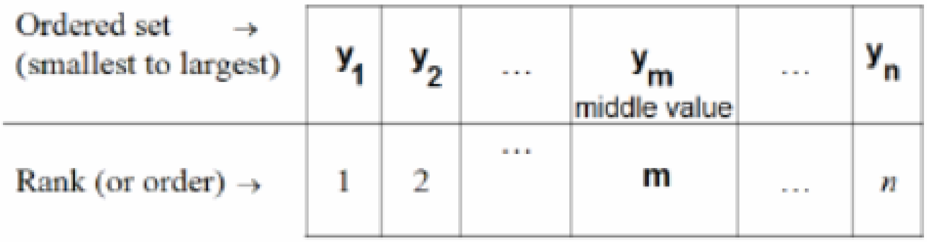
$$rank = \frac{n+1}{2}$$

- If the sample size (n) is an odd number, there is only one value in the middle, and the rank will be an integer:

$$rank = \frac{n+1}{2} = m \qquad \text{(m is integer)}$$

The median is the middle value of the **<u>ordered</u>** observations, which is:

$$\text{Median} = y_m.$$



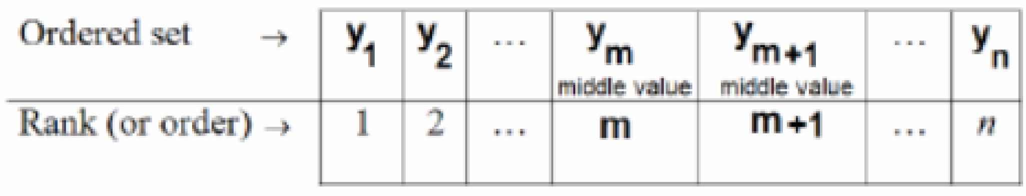
Ordered set → (smallest to largest)	y_1	y_2	…	y_m middle value	…	y_n
Rank (or order) →	1	2	…	m	…	n

- If the sample size (n) is an even number, there are two values in the middle, and the rank will be an integer plus 0.5:

$$rank = \frac{n+1}{2} = m + 0.5$$

Therefore, the ranks of the middle values are (m) and (m+1). The median is the mean (average) of the two middle values of the **ordered** observations:

$$\text{Median} = \frac{y_m + y_{m+1}}{2}.$$



Ordered set →	y_1	y_2	…	y_m middle value	y_{m+1} middle value	…	y_n
Rank (or order) →	1	2	…	m	m+1	…	n

Example (odd number):

Find the median for the sample values: 10, 54, 21, 38, 53.

Solution:

.$n = 5$ (odd number)

There is only one value in the middle.

The rank of the middle value is:

$$rank = \frac{n+1}{2} = \frac{5+1}{2} = 3. \quad (m=3)$$

Ordered set →	10	21	**38** (middle value)	53	54
Rank (or order) →	1	2	**3** **(m)**	4	5

The median =38 (unit)

Example (even number):

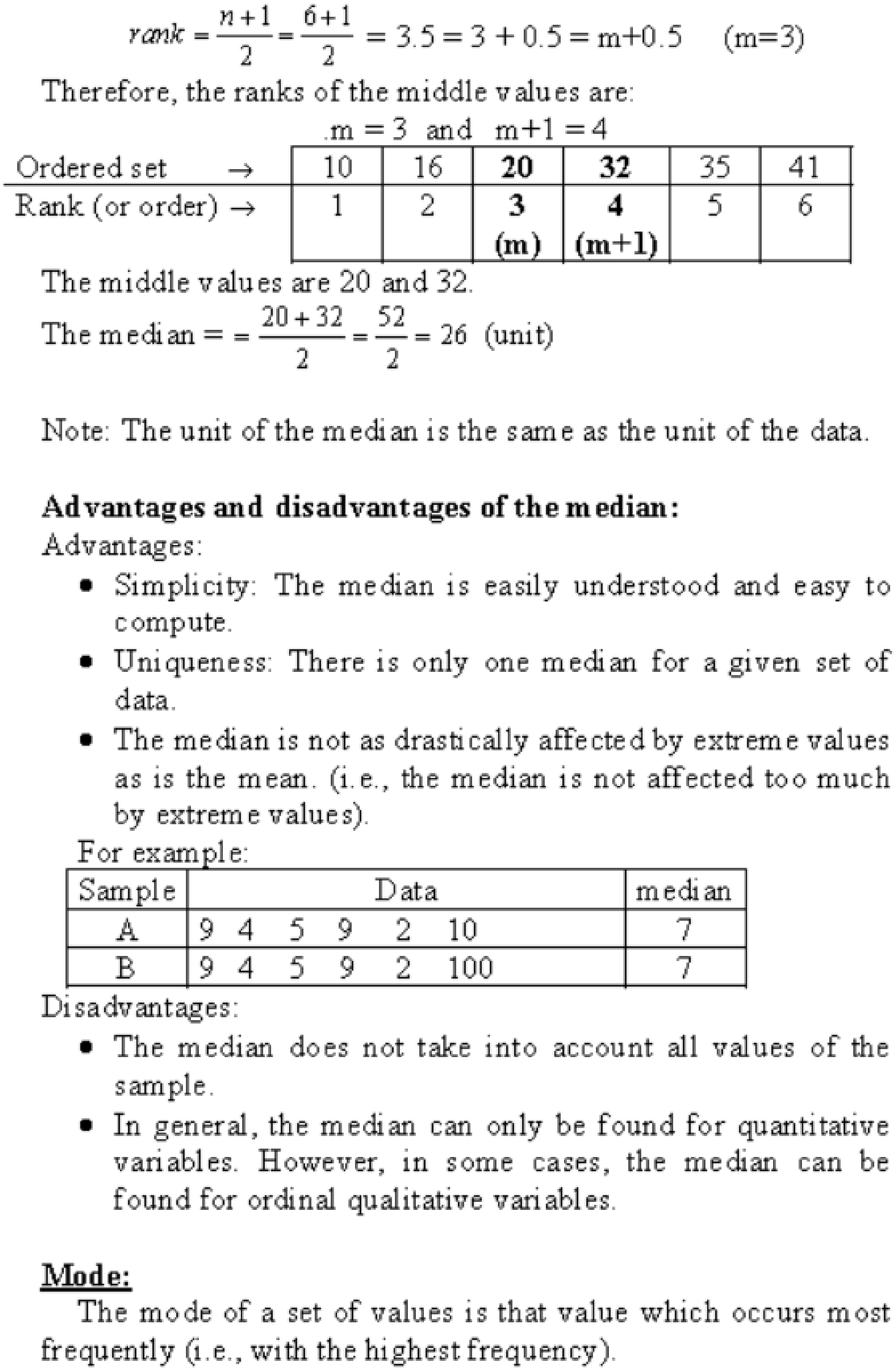
Find the median for the sample values: 10, 35, 41, 16, 20, 32

Solution:

.$n = 6$ (even number)

There are two values in the middle.

The rank is:



$$rank = \frac{n+1}{2} = \frac{6+1}{2} = 3.5 = 3 + 0.5 = m+0.5 \quad (m=3)$$

Therefore, the ranks of the middle values are:

.m = 3 and m+1 = 4

Ordered set →	10	16	**20**	**32**	35	41
Rank (or order) →	1	2	**3** **(m)**	**4** **(m+1)**	5	6

The middle values are 20 and 32.

The median = $= \frac{20+32}{2} = \frac{52}{2} = 26$ (unit)

Note: The unit of the median is the same as the unit of the data.

Advantages and disadvantages of the median:

Advantages:

- Simplicity: The median is easily understood and easy to compute.
- Uniqueness: There is only one median for a given set of data.
- The median is not as drastically affected by extreme values as is the mean. (i.e., the median is not affected too much by extreme values).

For example:

Sample	Data	median
A	9 4 5 9 2 10	7
B	9 4 5 9 2 100	7

Disadvantages:

- The median does not take into account all values of the sample.
- In general, the median can only be found for quantitative variables. However, in some cases, the median can be found for ordinal qualitative variables.

Mode:

The mode of a set of values is that value which occurs most frequently (i.e., with the highest frequency).

- If all values are different or have the same frequencies, there will be no mode.
- A set of data may have more than one mode.

Example:

Data set	Type	Mode(s)
26, 25, 25, 34	Quantitative	25
3, 7, 12, 6, 19	Quantitative	No mode
3, 3, 7, 7, 12, 12, 6, 6, 19, 19	Quantitative	No mode
3, 3, 12, 6, 8, 8	Quantitative	3 and 8
B C A B B B C B B	Qualitative	B
B C A B A B C A C	Qualitative	No mode
B C A B B C B C C	Qualitative	B and C

Note: The unit of the mode is the same as the unit of the data.

Advantages and disadvantages of the mode:

Advantages:

- Simplicity: the mode is easily understood and easy to compute..
- The mode is not as drastically affected by extreme values as is the mean. (i.e., the mode is not affected too much by extreme values).

For example:

Sample	Data	Mode
A	7 4 5 7 2 10	7
B	7 4 5 7 2 100	7

- The mode may be found for both quantitative and qualitative variables.

Disadvantages:

- The mode is not a “good” measure of location, because it depends on a few values of the data.
- The mode does not take into account all values of the sample.
- There might be no mode for a data set.
- There might be more than one mode for a data set.