CHAPTER FOUR - TREATMENT OF GROUPED SAMPLE DATA

Frequency Distribution

The probability distribution of a random variable is often very useful in studying the behaviour of the distribution if presented in a suitable form. Considerable information can be obtained by grouping our data into classes and determining the number of observations in each of the classes. Such an arrangement is called a ***frequency distribution***.

Frequency distribution is one of the ways in which we can organize a large number of data. Data that are presented in the form a frequency distribution are called ***grouped data.*** The number of observations falling in a particular class is called the ***class frequency*** and is denoted f.

The basic way to build frequency distribution is to divide the range of data values into classes or (class center) and limit the number of data within each class (or class center).

Example 1.

The following data represent the grades of 32 students in physics in the ministerial exam for the preparatory stage
22, 47, 88, 71, 34, 54, 62, 41, 36, 87, 76, 69, 48, 29, 33, 66, 42, 52, 58, 99, 53, 57, 59, 74, 39, 45, 42, 58, 63, 84, 55, 58.

The following table represents the frequency distribution of these grades.

classes	frequency
20-30	2
30-40	4
40-50	6
50-60	9
60-70	4
70-80	3
80-90	3
90-100	1
Total	32

From the above example, it is shown that the frequency distribution is a table consisting of classes, namely the values of observations or measurements, and the frequencies corresponding to these classes.

When building the frequency distribution, we have to take in view the following points:

1. Classes must be separated.
2. The classes should be of equal length.
3. The classes are sufficient to hold the data. This means that if we look at any value in the data we can put it in one class. This allows us to enter all the data in the frequency distribution classes and the sum of these frequencies equal to the number of data, i.e. if the number of data is n we have $\sum_{i=1}^{k} f_{i} = \mathrm{n}$ (where k represents the number of classes).

To construct the above frequency distribution table, we follow these steps:
-Calculate the range that equals the difference between the largest and smallest numerical values for that class. 99-22 = 77.
-One integer is added to the range to include the smallest and largest decimal (77 + 1 = 78).
-Choose a suitable class length so that we can get a number of classes between 6 and 15 (class length 10).
-The number of classes is calculated by dividing (range +1) by the length of the class and rounding the result to an integer (78/10 = 8).
-The classes are fixed by setting the minimum value of the class plus the length of the class.
-Data is entered into the classes and the number of frequencies for each class is recorded in the frequency column.
-Note that the sum of the frequencies is equal to the total number of the data.

Example.2

To illustrate the construction of a frequency distribution, consider the following data, which represent the lives of 40 similar car batteries recorded to the nearest tenth of year.

The batteries were guaranteed to last 3 years.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Let us choose 7 class intervals, to determine approximate class width, we divide the range by the number of intervals. Therefore the range is 4.7 - 1.6 = 3.1 and the class width can be no less than

$\frac{3.1}{7} = 0.443$ we choose 0.5.

The frequency distribution for the above data is given in the following table

Class interval	Class boundaries	class mark	Frequency
1.5 – 1.9	1.45 – 1.95	1.7	2
2.0 – 2.4	1.95 – 2.45	2.2	1
2.5 – 2.9	2.45 – 2.95	2.7	4
3.0 – 3.4	2.95 – 3.45	3.2	15
3.5 – 3.9	3.45 – 3.95	3.7	10
4.0 – 4.4	3.95 – 4.45	4.2	5
4.5 – 4.9	4.45 – 4.95	4.7	3
total			40

Graphic Representation

The information provided by a frequency distribution in tabular form is easier to grasp if presented graphically.

The most widely used form of graphic presentation of numerical data are bar charts, histograms and polygons.

In this chapter, we will learn

- how to define a histograms
- how to make and interpret histograms
- the differences between histograms and bar graphs

The following diagram shows the differences between a histogram and a bar chart.

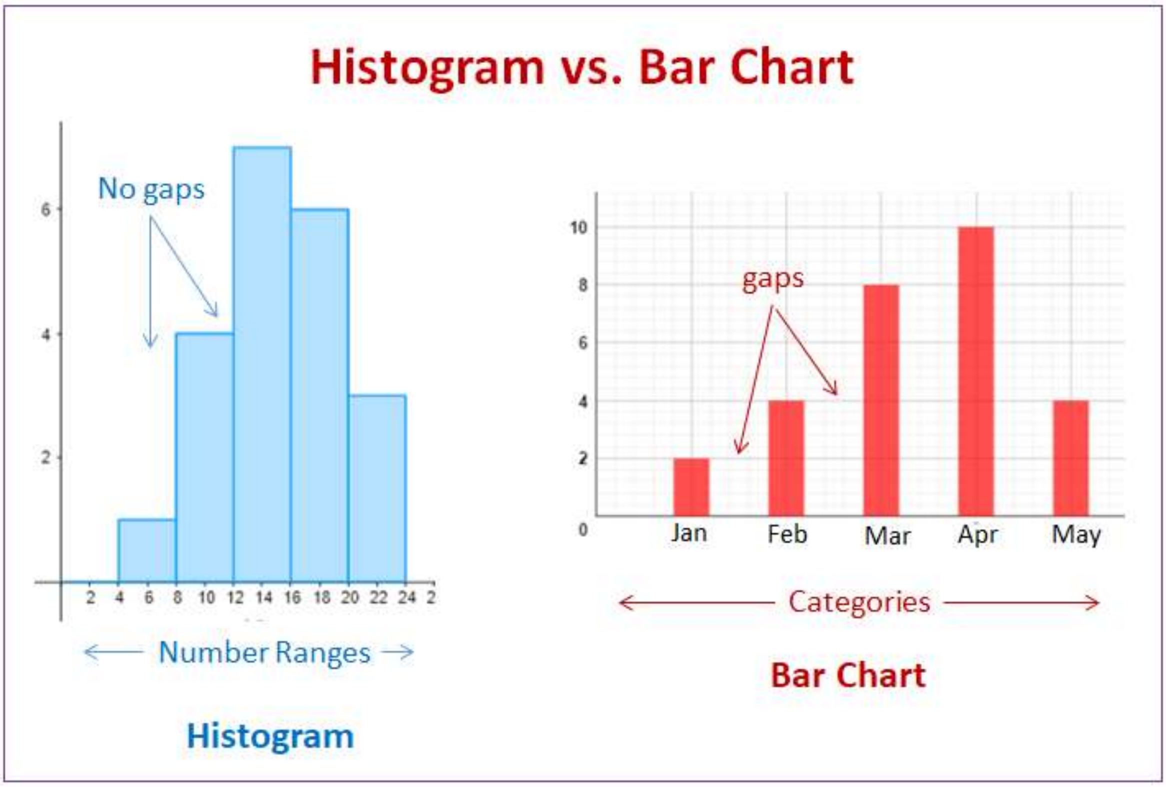

Figure 1: Histogram and Bar Chart

Compare Bar Graphs and Histograms

Histograms are used to show distributions of variables whereas bar charts are used to compare variables. Histograms plot quantitative data with ranges of the data grouped into intervals while bar charts plot categorical data.

Note that there are no spaces between the bars of a histogram since there are no gaps between the intervals. On the other hand, there are spaces between the variables of a bar chart.

Bar Charts

A **bar chart** represents the data as horizontal or vertical bars. The length of each bar is proportional to the amount that it represents.

The Bar Chart of the table for the frequency distribution in example.2 is shown in the following figure.

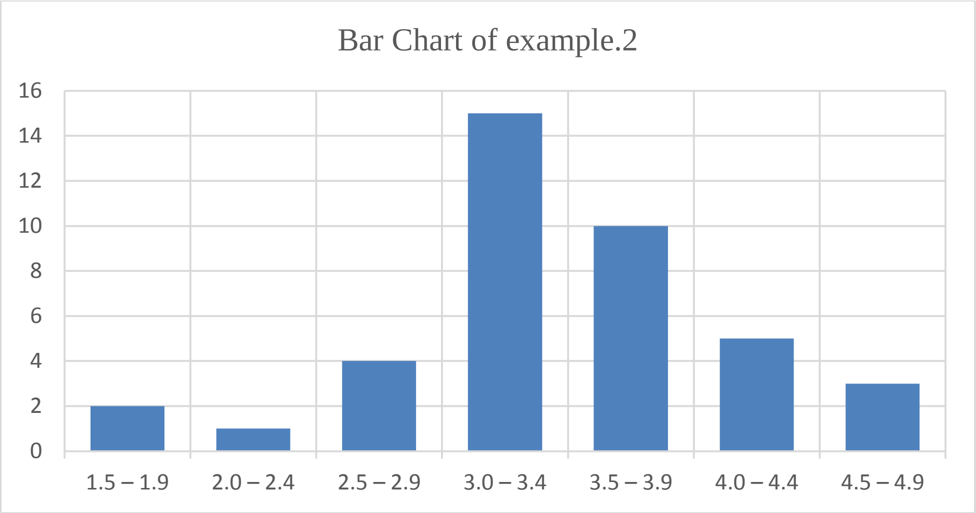

Figure 2: Bar Chart graph of example.2

Histograms

How to define a histogram, interpret a histogram and create a histogram from data?
A histogram is a bar graph that represents a frequency distribution. The width represents the interval and the height represents the corresponding frequency. There are no spaces between the bars.

Polygons

The frequency polygon is obtained by fixing the position of each mid class against the frequency of that class and then connecting these points by straight lines. We reached the two end points of the polygon by the previous mid class point from the left and the next mid class from the right. The polygon is joined by these two points, as in the following figure:

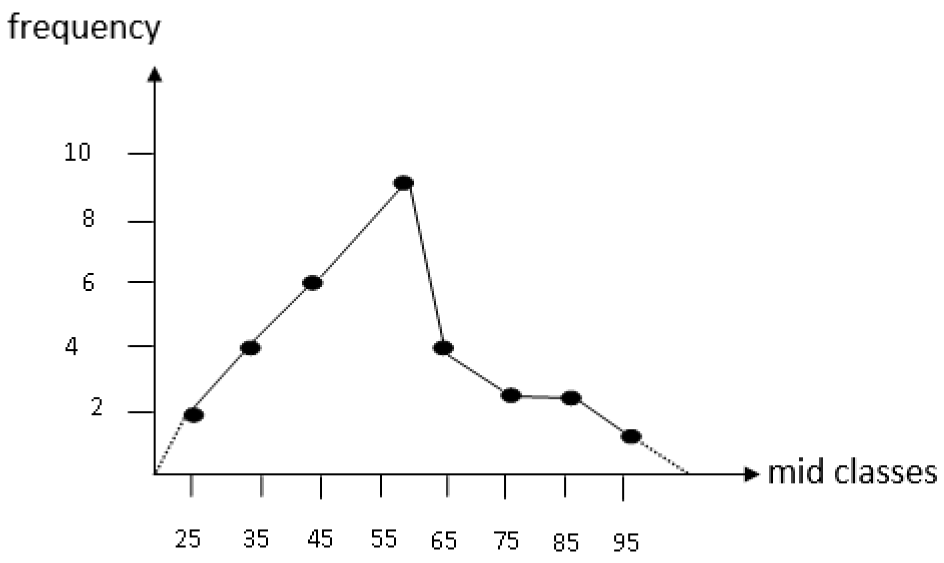

Figure 3: Polygon graph of example.1

The frequency polygon can also be obtained from the histogram by pointing the upper sides of the rectangles in the histogram and then connecting these points together with each other as in the following figure for example.1:

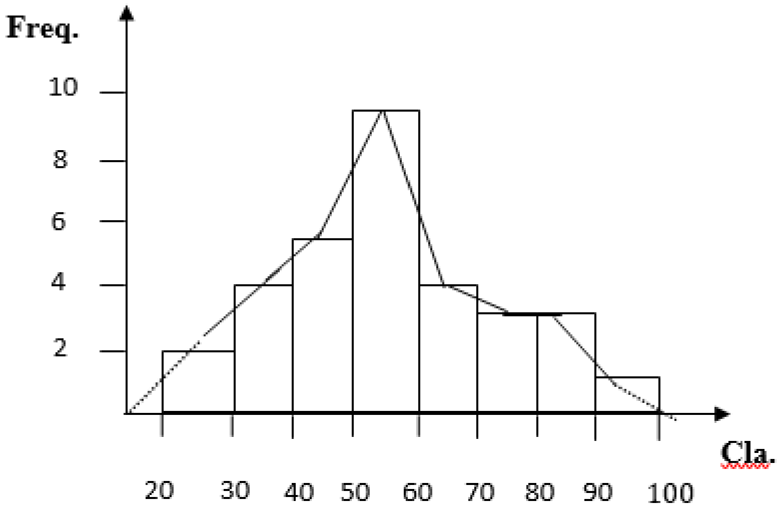

Figure 4: Polygon graph of example.1

The following graph represent the frequency polygon of Example.2 (Battery Lives).

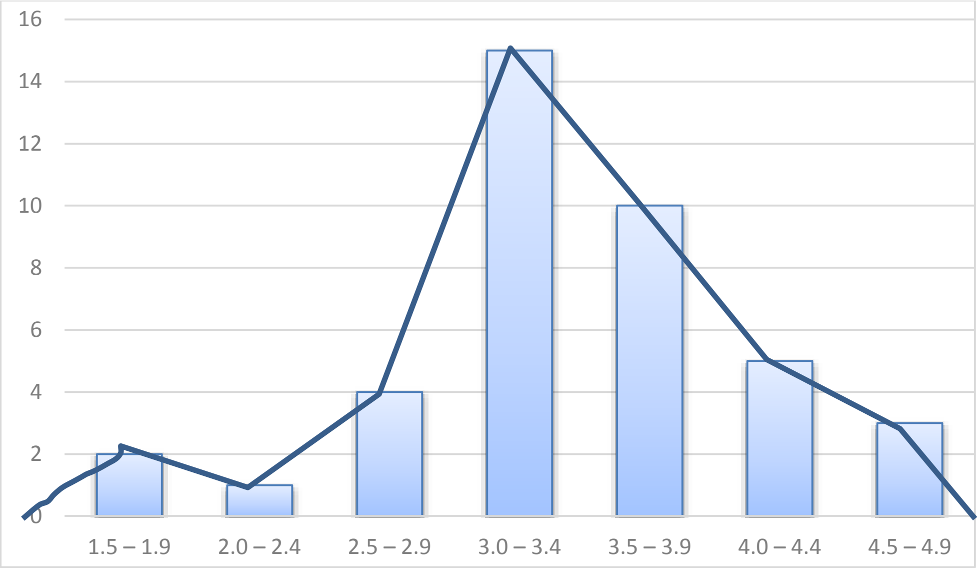

Figure 5: Polygon graph of example.2

Relative Frequencies:

The relative frequency of each class is the ratio of the frequency of that class to the total frequency. If the sum of the total frequencies is n and the frequency of class i is f_i , then its relative frequency is $p_i = \frac{f_i}{n}$. If we multiply the relative frequency p_i by 100 (p_i x100), then we get the percentage frequency as in the following examples:

Example.3

The following table shows the relative frequency and the percentage frequency of the frequency distribution table for example (1):

Classes	**freq.** f_i	**Relative freq.**	**Percentage freq. %**
20-30	2	1/16	6.25
30-40	4	1/8	12.5
40-50	6	6/32	18.075
50-60	9	9/32	28.125
60-70	4	1/8	12.5
70-80	3	3/32	9.375
80-90	3	3/32	9.375
90-100	1	1/32	3.125
Total	32		

Cumulative frequency distribution:

Often our interest is in the number of observation that are equal to or smaller than a given value.

The sum of the frequencies of all values that are equal to or smaller than a value is the cumulated frequency of that value.

Example.4

Below is the cumulated frequency distribution table for Example.1:

classes	freq. f_i	Cumulated freq.
20-30	2	2
30-40	4	6
40-50	6	12
50-60	9	21
60-70	4	25
70-80	3	28
80-90	3	31
90-100	1	32
Total	32	

The number of grades that fall in the class 50-60 or less is 21.

Mean, Median, and Mode

In chapter 3 we defined the mean of a set of observations to be their arithmetic average. If the data have been grouped we have lost the identity of the observations. To evaluate the mean we shall assume that all the observations within a given class interval fall at the class midpoint or class mark.

Definition.1 If $x_1, x_2, \ . \ . \ ., x_k$ are the class marks (class midpoints) of a set of grouped data with corresponding class frequencies $f_1, f_2, \ . \ ., f_k$ then **the mean** of our sample is

$$\bar{\mu} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

The computation of the mean for the data of (Battery Lives) is illustrated by the following table

Class interval	class mark (x_i)	Frequency (f_i)	$x_i . f_i$
1.5 – 1.9	1.7	2	3.4
2.0 – 2.4	2.2	1	2.2
2.5 – 2.9	2.7	4	10.8
3.0 – 3.4	3.2	15	48.0
3.5 – 3.9	3.7	10	37.0
4.0 – 4.4	4.2	5	21.0
4.5 – 4.9	4.7	3	14.1
Total		40	136.5

Hence, the mean $\mu = \frac{136.5}{40} = 3.4125$ years.

Definition.2 For grouped data (frequency distribution data), to find the **median** we first specified the median class. The median class is defined as that class who has cumulated frequency greater than or equal to N / 2 directly. After determining the median class, the median is calculated from the following formula:

$$M_e = X_l + \frac{\frac{N}{2} - f^c}{f} * c$$

where:

l is the number of classes.

$\sum_{\mathrm{i}=1}^{\mathrm{l}} f_i = N$ is the total number of frequencies.

X_{L} is the lower bound of the median class.

f^c is the cumulated frequency before the median class.

f is the frequency of the median class.

c is length the class interval.

Notice that the calculated median does not depend on all the values and does not affected by the extreme values.

Class interval	class mark (x_i)	Frequency (f_i)	$cum\ f$
1.5 – 1.9	1.7	2	2
2.0 – 2.4	2.2	1	3
2.5 – 2.9	2.7	4	7

3.0 – 3.4	3.2	15	22
3.5 – 3.9	3.7	10	32
4.0 – 4.4	4.2	5	37
4.5 – 4.9	4.7	3	40
Total		40	

MedianClass and ModeClass

For the above data our estimate of the median is 3.3

Where $X_{\mathrm{L}} = 3,\ f^{c} = 7,\ f = 15,\ \ c = 0.4,\ N = 40 \Longrightarrow M_e = 3 + \frac{20-7}{15} * 0.4 = 3.346$

Also our estimate of the mode is 3.2.

Definition.3 The **mode** X_m for usual data is defined as the value with the highest frequency. When the data is given in frequency distribution, the corresponding class must first be fixed. **The mode class is defined as the class with the highest frequency**. After finding the mode class we find the mode from the following formula:

$$M_o = \mathrm{X_L} + \frac{\Delta_1}{\Delta_1 + \Delta_2} * c$$

X_{L} is the lower limit of the mode class .

Δ_1 is the difference between the frequency of the mode class and the frequency of the previous class.

Δ_2 is the difference between the frequency of the mode class and the frequency of the subsequent class. c is the length of the mode class.

$(\mathrm{X_L} = 3,\ \Delta_1 = 15 - 4 = 11,\ \Delta_2 = 15 - 10 = 5, c = 0.4)\ for\ example.2$

Measurements of dispersion

Range

The range is defined as the difference between the highest value and the smallest value of data. If the range is small, it means that the data is confined to a close range and if the range is large, the data is within a long distance.

The range in the frequency distribution is also defined as the difference between the upper limit of the upper class and the lower limit of the lower class.

Mean Deviation (M.D)

In the case of the frequency distribution with mid classes $X_1, \mathrm{X}_2, \dots, X_l$ and their corresponding frequencies $f_1, f_2, \dots, f_l$, the mean deviation is:

$$M.D = \frac{\sum_{i=1}^{l} f_i \left| X_i - \overline{X} \right|}{\sum_{i=1}^{l} f_i}$$

Variance

In the case of the frequency distribution with mid classes $X_1, X_2, \dots, X_l$ and their corresponding frequencies $f_1, f_2, \dots, f_l$, the variance is:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{l} f_i\,(X_i-\bar{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{l} f_iX_i^2 - n\bar{X}^2\right]$$

Where $\sum_{i=1}^{l} f_i = n$.

Standard Deviation

The standard deviation is defined as the positive square root of the variance i.e.$S = \sqrt{S^2}$

Example.5

The table below shows the weights(kg) of members in a sport club. Calculate the mean, median, mode, mean deviation and standard deviation of the distribution.

Masses (kg)	40-49	50-59	60-69	70-79	80-89	90-99
Frequency (f_i)	6	8	12	14	7	3

Solution:

1. To find the **mean** we will use the formula

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

So we need to find the center of each class (x_i), then find $\sum_{i=1}^{k} f_i x_i$

Mass (kg)	**Frequency (f_i)**	Class mark (x_i)	$x_i f_i$	Cum f
40 – 49	6	44.5	267	6
50 – 59	8	54.5	436	14
60 – 69	12	64.5	774	26
70 – 79	14	74.5	1043	40
80 – 89	7	84.5	591.5	47
90 – 99	3	94.5	283.5	50
Total	50		3395	

Median class

Mode class

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{3395}{50} = 67.9$$

2. To find the **median** we first specify the median class which is represent the cumulative frequency greater than or equal to (50/ 2) directly.

$M_e = X_l + \frac{\frac{N}{2} - f^c}{f} * c$ where c=10 , $X_l = \frac{59+60}{2} = 59.5$, $f^c = 14$, $f = 12$

$$M_e = 59.5 + \frac{25-14}{12} * 10 = 68.66$$

3. We can find the mode class which contain highest frequency, then find the mode from the following formula

$$M_o = \mathrm{X_L} + \frac{\Delta_1}{\Delta_1 + \Delta_2} * c$$

the lower limit of the mode class is $X_L = \frac{69+70}{2} = 69.5$ and

$\Delta_1 = 14 - 12 = 2$, $\Delta_2 = 14 - 7 = 7$. Then:

$$M_o = 69.5 + \frac{2}{2+7} = 69.722$$

4. Mean deviation = $\frac{\sum_{i=1}^{l} f_i |x_i - \overline{X}|}{\sum_{i=1}^{l} f_i} = \frac{140.4+107.2+40.8+92.4+116.2+79.8}{50} = 11.536$
5. Variance $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{l} f_i X_i^2 - n\overline{X}^2\right]$

$$S^2 = \frac{1}{49}\{[6(44.5)^2 + 8(54.5)^2 + 12(64.5)^2 + 14(74.5)^2 + 7(84.5)^2 + 3(94.5)^2] - 50(67.9)^2\} = \frac{9522}{49}$$

Then the standard deviation $=\sqrt{S^2}$

S = 13.94

■

EXERCISES

1. The following scores represent the final examinations grade for an elementary statistics course:

23	60	79	32	57	74	52	70	82	36
80	77	81	95	41	65	92	85	55	76
52	10	64	75	78	25	80	98	81	67
41	71	83	54	64	72	88	62	74	43
60	78	89	76	84	48	84	90	15	79
34	67	17	82	69	74	63	80	85	61

a. Set up a frequency distribution using 10 intervals.
b. Find the mean, median, and the mod.
c. Construct a frequency histogram.
d. Construct a frequency polygon.
e. Find the relative frequency and the percentage frequency.

2. For the following data

class interval	frequency

7 – 13	5
14 – 20	4
21 – 27	3
28 – 34	2
35 – 41	2
42 – 48	1
49 – 55	2
56 – 62	1

a- Find the mean, median, and mode.

b- Find the mean deviation and standard deviation

c-. Construct a frequency histogram.

d- Construct a frequency polygon.

3. For the following data

class interval	frequency
10 – 14	1
15 – 19	3
20 – 24	4
25 – 29	4
30 – 34	5
35 – 39	3
40 – 44	1

1. find the mean, median, and mode.
2. Find the mean deviation and standard deviation

3. construct a frequency histogram.

4. construct a frequency polygon.

4. The following data represent the spending in dollars on extracurricular activities for a random sample of college students during the first week of the first semester

6 6 9 22 12 7 18 13 11 12 8 2 10 6

9	4	9	14	13	8	10	12	20	29	9	5	11	3
5	6	5	24	15	4	11	22	13	19	6	4	10	5

a - Set up a frequency distribution using 10 intervals.

b - Find the mean, median, and mod.

c- Find the relative frequency and the percentage frequency

d - Construct a frequency histogram.

e- Construct a frequency polygon.