Chapter 2 Descriptive Statistics

Raw data

⇨ data recorded in the sequence in which they are collected and before they are processed or ranked

Table 1: The weights of 20 students in kg (Quantitative raw data)

61	68	65	67	68	71	69	63	74	64
66	65	62	67	60	73	69	70	70	71

Table 2: The grades of AAMS2613 of 20 students (Qualitative raw data)

A	B	C	A	C	B	B	A	B	C
B	A	B	B	B	A	C	D	D	B

Arrays

⇨ an arrangement of numerical raw data in ascending order or descending order of magnitude

60 , 61 , 62 , 63 , 64 , 65 , 65 , 66 , 67 , 67 ,
68 , 68 , 69 , 69 , 70 , 70 , 71 , 71 , 73 , 74

Ungrouped data

⇨ Contains information on each member of a sample or population individually

⇨ Examples: Data presented in Table 1 and Table 2

Range

⇨ The difference between the largest and smallest numbers

⇨ Range = The largest value – the smallest value

✓ Range = 74 – 60 = 14 kg

Frequency distribution / table

⇨ A tabular arrangement of data by classes together with the corresponding class frequencies

Table 3: The weights of 20 students in kg

Weight (kg)	Tally	No of students
60 – 62		
63 – 65		
66 – 68		
69 – 71		
72 – 74		

The first class consists of weights from 60 to 62 kg. The class frequency is 3.

Grouped data
⇨ Data organised and summarised in the frequency distribution
⇨ Example: Data presented in Table 3

Note:
- Generally, the grouping process destroys some of the original information
- The classes are non-overlapping i.e. each value belongs to one and only one class

Class interval
⇨ an interval that includes all the values that fall within two numbers
⇨ Example: Third class interval =

Class limits
⇨ endpoints of each interval
  ✓ For third class interval =
  ✓ Lower class limit =
  ✓ Upper class limit =

Class Boundary
⇨ the dividing line between two classes
⇨ is given by the midpoint of the upper limit of one class and the lower limit of the next higher class
  ✓ Refer Table 3, upper second class limit =
  ✓ lower third class limit =
  ✓ Class boundary between the 2^{nd} and 3^{rd} class =

Class width / class size

⇨ is the difference between the upper and lower class boundary

⇨ Class width = Upper boundary – Lower boundary

- ✓ Lower boundary for second class = 62.5
- ✓ Upper boundary for second class = 65.5
- ⇨ Width of the second class =

Class mark / class midpoint

⇨ is the midpoint of the class interval

⇨ Class mark = (Lower class limit + Upper class limit) / 2

- ✓ The class mark for the interval 66 – 68 is
- ✓

Example 1:

Class intervals	Class boundaries	Class size	Class mark
11 – 20			
21 – 30			
31 – 40			
41 – 45			
46 – 60			

Example 2:

Class intervals	Class boundaries	Class size	Class mark
5 – < 10			
10 – < 15			
15 – < 20			
20 – < 30			

Histogram and frequency polygon

⇨ Two graphical representations of frequency distribution.

Three types of histogram

1. Frequency histogram

2. Relative frequency histogram
3. Percentage histogram

A frequency histogram consists of a set of rectangle that having

a) The bases on a horizontal axis with centres at the class marks and lengths equal to the class interval sizes
b) The areas proportional to the class frequencies

If the class intervals all have equal size

the height of the rectangles are proportional to the class frequencies

otherwise

the height of the rectangles must be adjusted

Polygon

⇨ A line graph formed by joining the midpoints of the tops of successive bars in a histogram

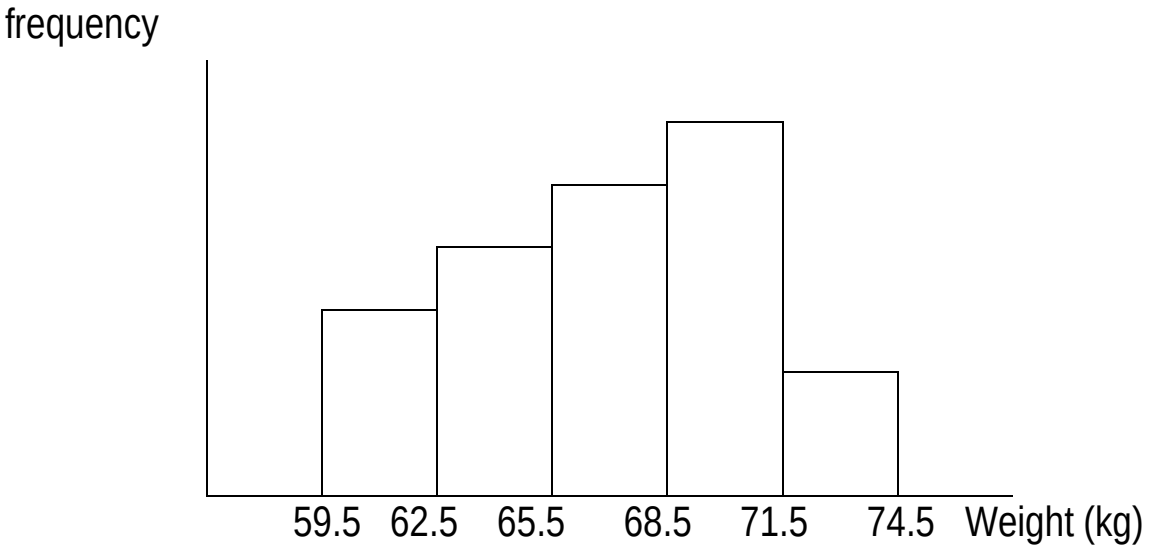
Three types of polygon

1. Frequency polygon
2. Relative frequency polygon
3. Percentage polygon

Table 4: The weights of 20 students in kg

Weight (kg)	Class mark	No of students
60 – 62	61	3
63 – 65	64	4
66 – 68	67	5
69 – 71	70	6
72 – 74	73	2

The frequency histogram and frequency polygon



For a given class,

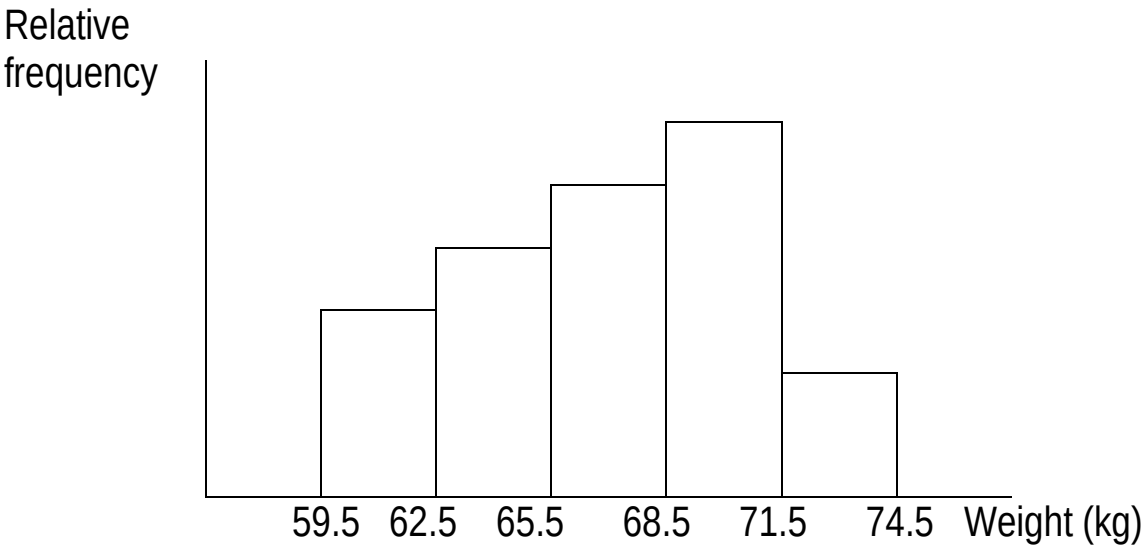
$$\text{Relative frequency} = \frac{\text{frequency of that class}}{\text{sum of all frequencies}} = \frac{f}{\Sigma f}$$

Percentage frequency = (Relative frequency) × 100

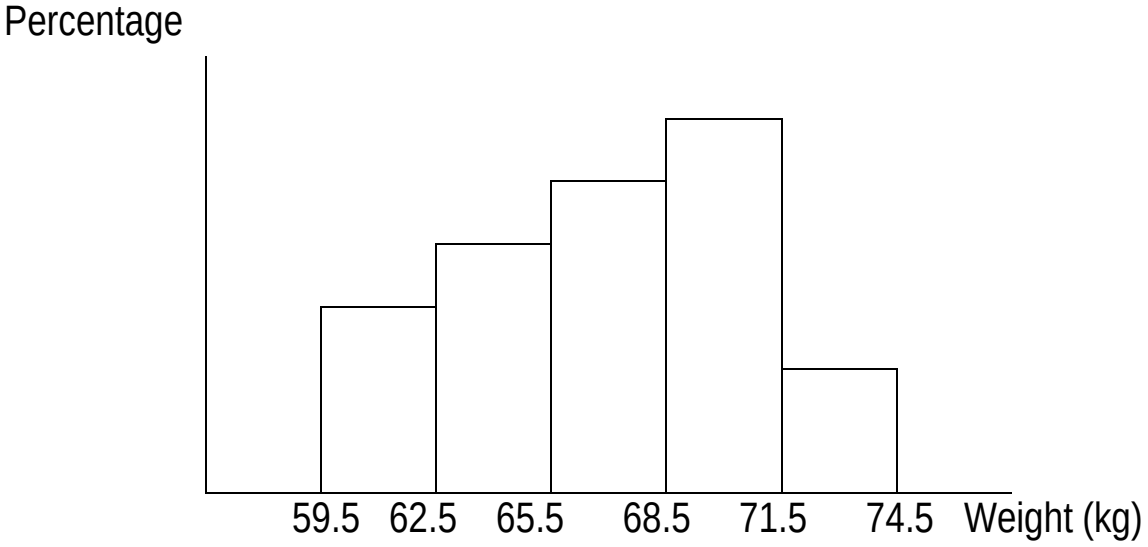
Table 5: The weights of 20 students in kg

Weight (kg)	Class mark	Frequency	Relative frequency	Percentage
60 – 62	61	3		
63 – 65	64	4		
66 – 68	67	5		
69 – 71	70	6		
72 – 74	73	2		
		$\Sigma f = 20$		

The relative frequency histogram



The percentage histogram



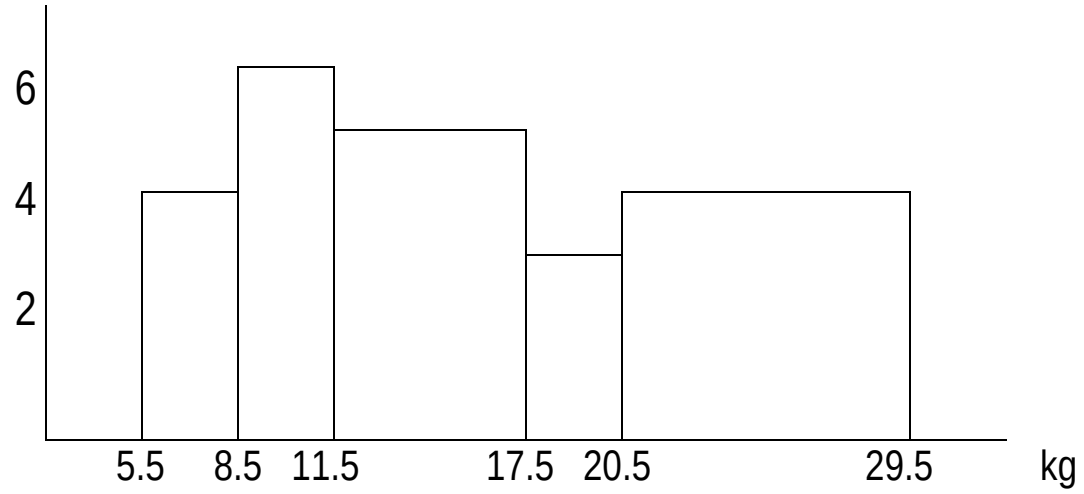
Example 3:
The frequency distribution gives the weight of 35 objects, measured to the nearest kg. Draw a histogram to illustrate the data.

Weight (kg)	6 – 8	9 – 11	12 – 17	18 – 20	21 – 29
Frequency	4	6	10	3	12

Solution:

Weight (kg)	Class width		Frequency	Height of rectangle (standard frequency)
6 – 8	3		4	
9 – 11	3		6	
12 – 17	6		10	
18 – 20	3		3	
21 – 29	9		12	

standard frequency



Cumulative frequency

⇨ The total frequency of all values that less than the upper class boundary of each class

Cumulative frequency distribution

⇨ A table that presenting the cumulative frequency

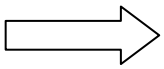
✓ Example: Table 6

$$\text{Cumulative relative frequency} = \frac{\text{cumulative frequency}}{\text{sum of all frequencies}}$$

Cumulative percentage = (Cumulative relative frequency) ✕ 100

Table 6

Weight (kg)	Frequency
60 – 62	3
63 – 65	4
66 – 68	5
69 – 71	6
72 – 74	2



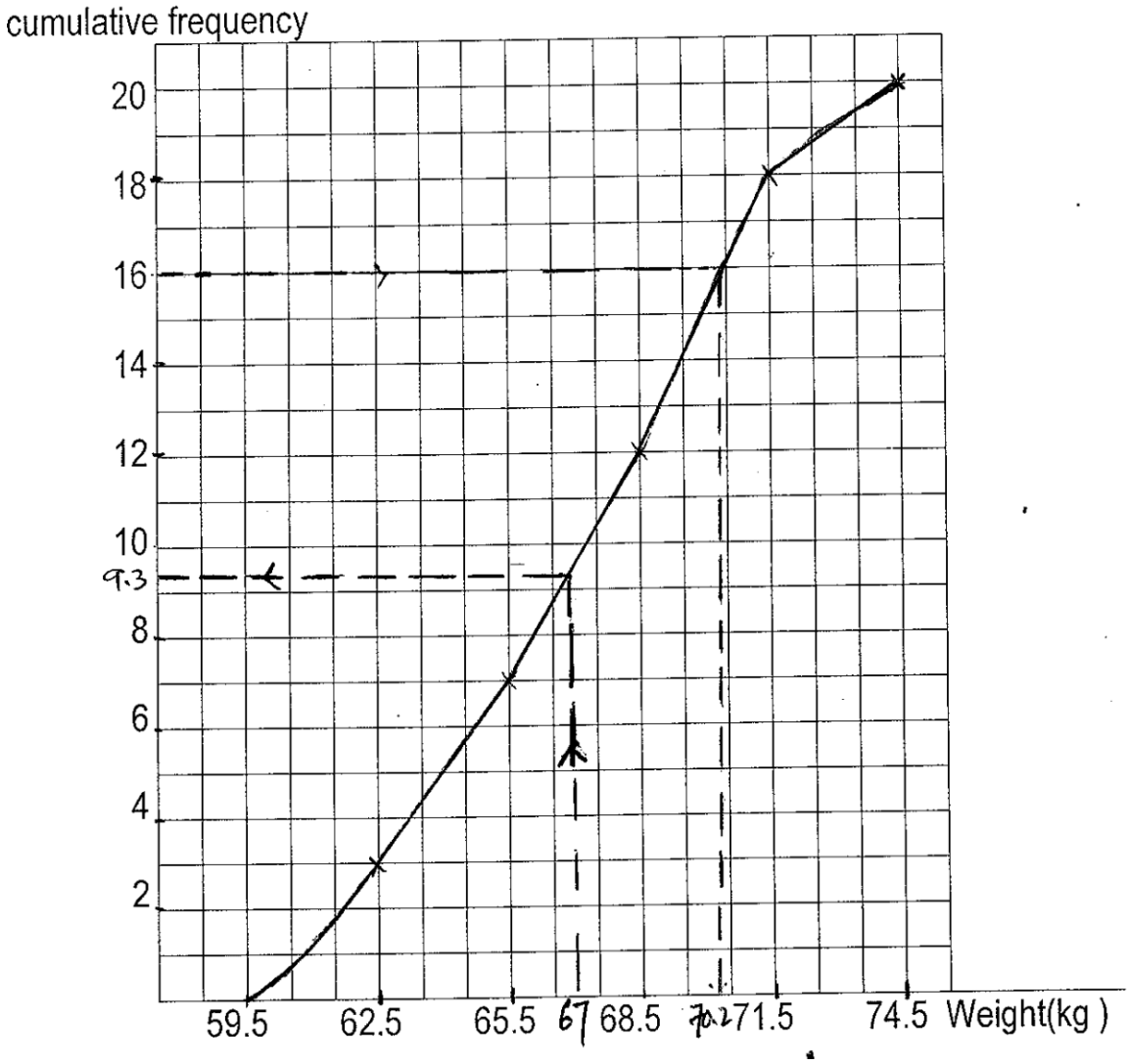
Weight (kg)	Cumulative frequency

Table 7: Cumulative relative frequency and cumulative percentage

Weight (kg)	Cumulative relative frequency	Cumulative percentage
< 59.5		
< 62.5		
< 65.5		
< 68.5		
< 71.5		
< 74.5		

Ogive, Cumulative frequency curve and Cumulative Frequency Polygon
⇨ The cumulative frequency are plotted against the upper class boundary

Note:
If relative cumulative frequency is used in placed of cumulative frequency, the graph is called relative cumulative frequency curve or percentage ogive



Example 4: Estimate from the above curve

a) the total number of students that their weight were less than 67 kg

b) the value of X ,if 20 % of the students were of weight X kg or more

Solution:

a) From the graph, 9.3 students were less than 67 kg

⇨ 9 students were less than 67kg

b) 20 % of the students were of weight X kg or more

⇨ 80 % of the students were of weight X kg or less

⇨ 16 students were less than X kg

⇨ From the graph, X = 70.2

Measures of central tendency

⇨ Represent a data set by some numerical measures (typical values)
⇨ Give the centre of a histogram or a frequency distribution curve

Consists of 3 measures
1. Median 2. Mode 3. Mean

The Median
⇨ the value of the middle term in a data set that has been ranked in increasing or decreasing order

Median = the value of the $\left(\frac{n+1}{2}\right)th$ term in a ranked data set; *n* = total number

Note:
1. If *n* is odd, then median ≡ the value of the middle term in the ranked data
2. If *n* is even, then median ≡ the average values of the two middle term

a) Raw data

Example 5:
Find the median of set A = { 10, 5, 19, 8, 3 } and set B = { 2, 7, 3, 6, 4, 5 }

Solution:

Note: 1. Median is not influence by the extreme value
2. Extreme values are values that are very small or very large relative to the majority of the values in a data set

b) Ungrouped frequency distribution

⇨ the median can be found either from ungrouped frequency distribution or from the cumulative frequency distribution

Example 6: Find the median of the following frequency distribution.

No. of children	0	1	2	3	4	5
Frequency	3	5	12	9	4	2

Solution:

c) Grouped frequency distribution

⇨ When calculating the median for grouped distributions, take the middle number as $(n + 1) / 2$ for the odd number but $n / 2$ for even number

Example 7: Estimate the median of the following frequency distribution.

Weight (kg)	60 – 62	63 – 65	66 – 68	69 – 71	72 – 74
Frequency	3	4	5	6	2

Solution:

Method 1: Linear Interpolation

$$Median = L_m + \frac{C_m}{f_m}\left(\frac{n}{2} - \sum f_{m-1}\right)$$

where

L_m = lower class boundary of the median class

C_m = the size of the median class

f_m = frequency of the median class

$\sum f_{m-1}$ = the cumulative frequency in the class preceding the median class

n = total frequency

Solution

Method 2: Cumulative frequency curve

From the cumulative frequency curve, the value corresponding to a cumulative frequency of 10 is 67.5

⇨ median = 67.5 kg

The Mode

⇨ the value that occurs with the highest frequency in a data set

a) Raw data

Example 8

Find the mode of each of the following data set.

i) 74, 9, 5, 8, 3, 8, 8

ii) 2, 6, 6, 6, 3, 8, 8, 8, 3

iii) 2, 2, 6, 6, 8, 8, 9, 9

iv) B, C, D, A, A, C, C, C, B, A

Solution:

i)

ii)

iii)

iv)

Note: 1. Mode is not influence by the extreme value
2. Mode may not exist, exist one mode or multi modes (not unique)
3. Mode can be used for both quantitative and qualitative data

b) Ungrouped frequency distribution

Example 9: Find the mode of the following frequency distribution.

No. of children	0	1	2	3	4	5
Frequency	3	5	12	9	4	2

Solution:

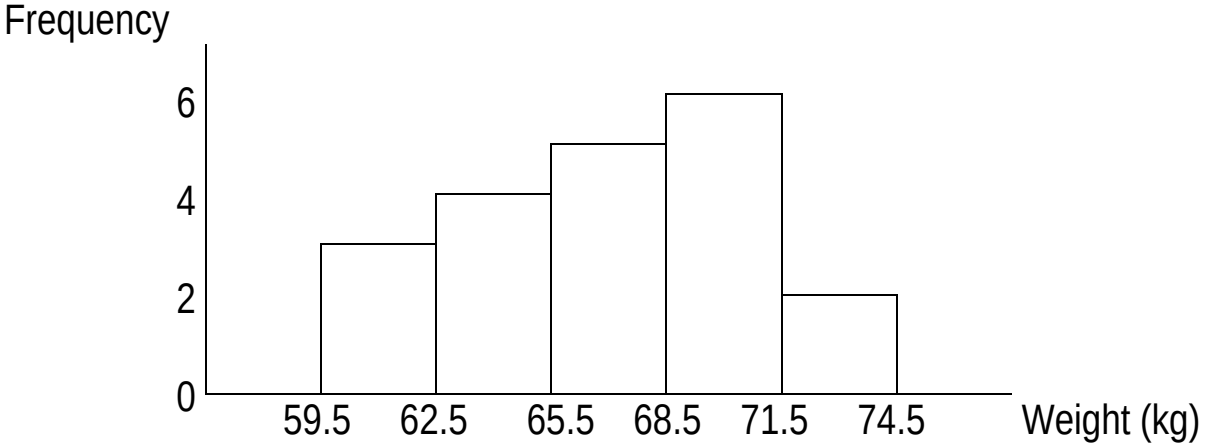
c) Grouped frequency distribution

Modal class
⇨ The class which has the largest standard frequency
- An estimate of the mode can be obtained from the modal class

Example 10: Estimate the mode of the following frequency distribution.

Weight (kg)	60 – 62	63 – 65	66 – 68	69 – 71	72 – 74
Frequency	3	4	5	6	2

Solution:
Method 1: Estimate the mode by graphically (histogram)



The modal class: 69 – 71
Estimated mode = 69.1 kg

Method 2: Estimate the mode by calculation

$$Mode = L_m + \frac{f_m - f_b}{2f_m - (f_b + f_a)} \times C$$

where

L_m = lower class boundary of the modal class

f_m = frequency of the modal class

f_b = frequency of the next lower class below the modal class

f_a = frequency of the next higher class above the modal class

C = the size of the modal class

The arithmetic mean

The arithmetic mean of the *n* numbers $x_1, x_2, \cdots, x_n$ is denoted by $\overline{x}$ and is defined as

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

a) Raw data

Example 11:
Find the arithmetic mean for the data set { 158, 189, 265, 127, 191 }

Solution:

Note: 1. Mean not necessary takes one of the values in the original data
2. Mean is influenced by extreme value
3. Mean is not suitable in the data set that contain extreme value

b) Ungrouped frequency distribution

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{n} = \frac{1}{n}\sum f_i x_i = \frac{\sum f_i x_i}{\sum f_i}$$

Example 12: Find the mean of the following frequency distribution.

x_i	2	5	6	8
f_i	1	3	4	2

Solution:

Note: It is particularly useful to use an assumed mean when dealing with large numbers or ones involving fractions or decimal.

c) Grouped frequency distribution

⇨ When data had been grouped into intervals, the midpoint x_i of the interval is taken to represent the interval

Example 13: Find the mean of the following frequency distribution.

AAMS1683 Scores	10 – 12	13 – 15	16 – 18	19 – 21
Number of students	4	12	20	14

Solution:

Some properties for the mean

For a data set $\{x_1, x_2, \cdots, x_n\}$ with mean $\bar{x}$. If each number in that data set

1. is added with a constant *b*, then the new mean is $\bar{x} + b$.
2. is multiplied by a constant *a*, then the new mean is $a\bar{x}$.
3. is multiplied by a constant *b* and then is added with a constant *a* ($y_i = a + bx_i;\ i = 1,2,\cdots,n$), then the new mean $\bar{y} = a + b\bar{x}$.

Measures of dispersion

⇨ Sometimes, the measures of central tendency only are not enough to reveal the whole picture of the distribution of a whole data set
⇨ The measure of central tendency does not describe how the data is distributed

Data set	Data	Mean	Median	Mode
A	1, 3, 6, 10, 10, 21, 26	11	10	10
B	7, 8, 10, 10, 10, 15, 17	11	10	10

✓ the mean, median and mode are the same for data set A and B but the distribution of the data is different

The Range

The range for a data set $\{x_1, x_2, \cdots, x_n\}$ is defined to be the difference between the largest value and smallest value.
⇨ Range = Largest value – Smallest value

Example 14: Find the range for data set A and data set B.

Solution:
For data set A, Range

For data set B, Range

Range based on quartiles

Quartiles are 3 summary measures that divide a ranked data set into 4 equal parts. These 3 measures are the first quartile, the second quartile, and the third quartile.

Consider *n* items arranged in ascending order. Then,

The first quartile = Lower quartile = $Q_1 = \frac{1}{4}(n+1)th$ value

The second quartile = Median = $Q_2 = \frac{1}{2}(n+1)th$ value

The third quartile = Upper quartile = $Q_3 = \frac{3}{4}(n+1)th$ value

The interquartile range (IQR) = $Q_3 - Q_1$

The semi-interquartile range = The quartile deviation = $(Q_3 - Q_1) / 2$

25 %	25 %	25 %	25 %

Q_1 Q_2 Q_3

a) Raw data

Example 15:
Find the lower quartile, upper quartile, and quartile deviation for the raw data

20, 28, 40, 12, 30, 15, 50

Solution:

Example 16:
Find the lower quartile and upper quartile using the above formula and using the definition of quartile for the raw data

75, 80, 68, 53, 99, 58, 76, 73, 85, 88, 91, 79

Solution:

Note: Quartiles are not necessary unique in a given data set

b) Ungrouped frequency distribution

Example 17: Find the Q_1 and Q_3 for the following distribution.

Marks	10	20	30	40	50	60
No of students	3	8	16	7	5	4
Cumulative frequency						

Solution:

c) Grouped frequency distribution

$$Q_1 = L_q + \frac{C_q}{f_q}\left(\frac{n}{4} - \sum f_{q-1}\right)$$

$$Q_3 = L_q + \frac{C_q}{f_q}\left(\frac{3n}{4} - \sum f_{q-1}\right)$$

where

L_q = lower class boundary of the quartile class

C_q = the size of the quartile class

f_q = frequency of the quartile class

$\sum f_{q-1}$ = the cumulative frequency in the class before the quartile class

n = total frequency

Example 18: Compute the Q_1 and Q_3 for the following distribution.

Weight (kg)	60 – 62	63 – 65	66 – 68	69 – 71	72 – 74
Frequency	3	4	5	6	2
Cumulative frequency					

Solution:

Variance

⇨ The variance is the average of the squared deviation of the data from the mean

Ungrouped data

Consider a population of *N* measurements $X_1, X_2, \cdots, X_N$

Population Mean = $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$

Population Variance = $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2$

Consider a sample of *n* measurements $X_1, X_2, \cdots, X_n$

Sample Mean = $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample Variance = $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2}{n-1}$

Grouped data

Population Variance = $\sigma^2 = \frac{\sum f_i (X_i - \mu)^2}{\sum f_i} = \frac{\sum f_i X_i^2}{\sum f_i} - \left(\frac{\sum f_i X_i}{\sum f_i}\right)^2$

Sample Variance = $S^2 = \frac{\sum f_i (X_i - \overline{X})^2}{n-1} = \frac{\sum f_i X_i^2 - \frac{1}{n}\left(\sum f_i X_i\right)^2}{n-1}$

Standard Deviation

⇨ The standard deviation is the positive square root of the variance

⇨ Sample standard deviation = S = $\sqrt{S^2}$

⇨ Population standard deviation = $\sigma = \sqrt{\sigma^2}$

- ⇨ A small standard deviation means that the data are distributed closely to their mean
- ⇨ A large standard deviation means that the data are widely scattered about their mean

Example 19:
Data shows the salary per day for *all* 6 employees of a small company.

29.50, 16.50, 35.40, 21.30, 49.70, 24.60

Calculate the variance and standard deviation for these data.

Solution:

Example 20:
A sample consists of 5 data values: 72, 49, 79, 55 and 57. Calculate the variance and standard deviation.

Solution:

Example 21:
Find the variance from the following frequency distribution if it represent
a) population b) sample

Height (m)	20 – 22	23 – 25	26 – 28	29 – 31	32 – 34
Frequency	3	6	12	9	2

Solution:

Height	Midpoint, x	Frequency, f	f x	f x ^{2}
20 – 22				
23 – 25				
26 – 28				
29 – 31				
32 – 34				

a)

$$\sigma^2 = \frac{\sum f_i X_i^{\,2}}{\sum f_i} - \left(\frac{\sum f_i X_i}{\sum f_i} \right)^2 =$$

b)

$$S^2 = \frac{\sum f_i X_i^{\,2} - \frac{1}{n}\left(\sum f_i X_i\right)^2}{n-1} =$$