# Unit 19 : Frequency distribution, graphical representations, central tendency and dispersion

## Learning Objectives

Students should be able to:

- organise raw data into a frequency distribution table.
- draw a histogram from a frequency distribution table.
- construct a cumulative frequency table.
- draw a cumulative frequency polygon.
- determine the mean, median and mode from ungrouped data
- determine the mean, median and mode from grouped data
- determine the range and inter quartile range.

## Activities

Use MS Excel to draw histogram and cumulative frequency polygon

## Reference

Suen, S.N. “Mathematics for Hong Kong 5A”; rev. ed.; Chapter 5; Canotta

# 1. Frequency distribution and their graphical representation

Data that have not been organised in any way are called raw data. They are collected by counting or measurement, or through other survey methods.

Table 1 shows data collected from 33 stocks in Hong Kong stock market as at 28 June 2002. Three different characteristics associated with a stock have been examined. Each such characteristic is an example of a variable associated with a stock. In Table 1, the variables observed are stock classification, lot size and yield percentage.

Table 1

| Stocks code | Stock Classification | Lot size | Yield % | Share price |
|---|---|---|---|---|
| 1 | Properties | 1000 | 2.48 | 64.5 |
| 2 | Utilities | 500 | 6.95 | 30.2 |
| 3 | Utilities | 1000 | 3.10 | 10.25 |
| 4 | Commerce | 1000 | 4.32 | 18.05 |
| 5 | Finance | 400 | 4.19 | 89.25 |
| 6 | Utilities | 500 | 5.57 | 29.25 |
| 8 | Info. Tech. | 1000 | 0 | 1.83 |
| 10 | Properties | 1000 | 6.07 | 7.25 |
| 11 | Finance | 100 | 5.90 | 83.00 |
| 12 | Properties | 1000 | 3.40 | 32.40 |
| 13 | Industries | 1000 | 3.02 | 57.25 |
| 14 | Properties | 1000 | 5.17 | 7.35 |
| 16 | Properties | 1000 | 2.62 | 59.25 |
| 17 | Properties | 1000 | 3.23 | 6.20 |
| 19 | Commerce | 500 | 2.91 | 38.50 |
| 20 | Commerce | 1000 | 1.21 | 6.20 |
| 23 | Finance | 200 | 3.50 | 14.45 |
| 66 | Transport | 500 | 4.14 | 10.15 |
| 83 | Properties | 2000 | 1.34 | 2.98 |
| 97 | Properties | 1000 | 3.59 | 6.40 |
| 101 | Properties | 500 | 4.52 | 8.85 |
| 179 | Industries | 500 | 1.16 | 8.90 |
| 267 | Commerce | 1000 | 4.76 | 16.80 |
| 291 | Properties | 2000 | 4.60 | 9.35 |
| 293 | Commerce | 1000 | 1.47 | 11.90 |
| 363 | Info. Tech. | 1000 | 3.28 | 14.65 |
| 494 | Info. Tech. | 2000 | 2.64 | 10.05 |
| 511 | Commerce | 1000 | 2.88 | 33.00 |
| 762 | Info. Tech. | 2000 | 0 | 5.95 |
| 883 | Industries | 500 | 2.48 | 10.10 |
| 941 | Info. Tech. | 500 | 0 | 22.65 |
| 992 | Commerce | 2000 | 1.79 | 2.85 |
| 1038 | Commerce | 1000 | 5.02 | 12.55 |

Variables can be divided into three different types:

- ***Categorical variable*** may be non-numeric or numeric. Its values describe the characteristics of the variable. For example, the colour of a mobile phone, the type of a car, the examination grade of a student, etc.
- ***Discrete variable*** is numeric. The values taken can only change in steps. For example: number of children in a family (which can take on values 0, 1, 2, etc. in steps of size 1), number of classrooms (which can only change in step of size 1, namely, 0, 1, 2, etc.), and the size of dresses (5, 6, 7, 8, 9…etc.).
- ***Continuous variable*** **is numeric**. The values taken can be any value in an interval. For example: weights of people, average exam marks of a student.

## 1.1 Tabular and graphical presentation of categorical variables

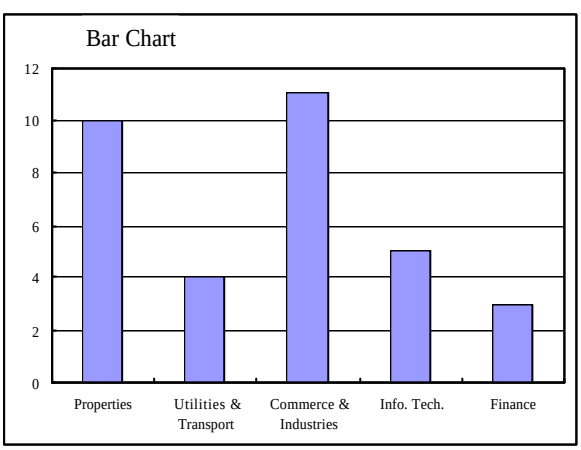
Consider the observation on stock classification in Table 1. There are six different stock classifications. We keep running a tally of the possible outcomes in a table. The presentation of data by listing them with the corresponding occurrence frequencies is called a 'frequency distribution'. A frequency distribution table can make data easier to interpret.

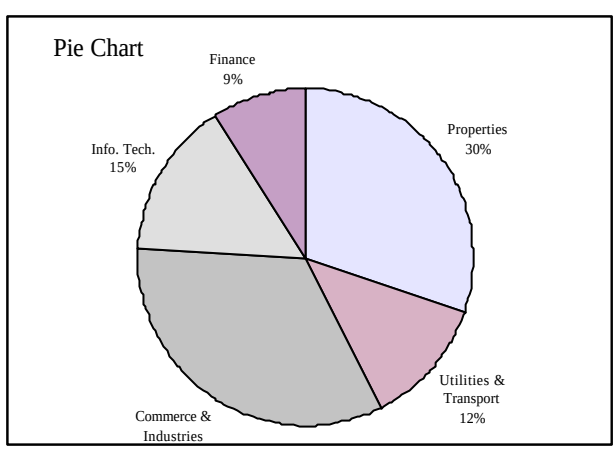
Table 2: Stock classifications

| Classification | Tally | Frequency | Relative Frequency |
|---|---|---|---|
| Finance | //// | 4 | 0.121 |
| Utilities | /// | | 0. |
| Properties | ~~////~~ ~~////~~ | 10 | 0.303 |
| Information Technology | ~~////~~ | | 0. |
| Commerce & Industries | ~~////~~ ~~////~~ / | 11 | 0.333 |
| Total | | | 1.000 |

Relative frequency of a class = frequency of the class / total frequency

Bar chart and pie chart are commonly used graphical devices for presenting categorical variables. In the bar chart the variable (classifications) is represented on the horizontal axis and the frequencies are represented by the height of vertical bars. In stead, in the pie chart a circle is drawn and it is divided into sectors having area proportional to the frequencies of the variable value.



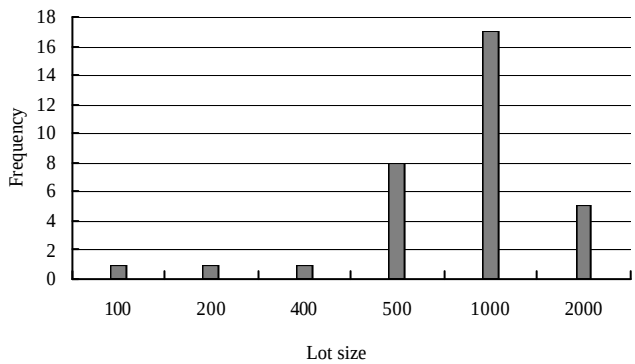
## 1.2 Tabular and graphical presentation of discrete variables

The lot size of 33 sample stocks in Table 1 is a discrete variable, because its possible values progress in steps, 100,200,... rather than any number in between 100 and 200. A bar chart may be used to present discrete variables.

Table 3 Lot size of 33 sample stocks

| Lot size | Tally of 33 stocks | Frequency |
|---|---|---|
| 100 | / | 1 |
| 200 | / | 1 |
| 400 | / | 1 |
| 500 | ~~///~~ /// | 8 |
| 1000 | ~~///~~ ~~///~~ ~~///~~ // | 17 |
| 2000 | ~~///~~ | 5 |
| | | Sum = _ |

Bar chart

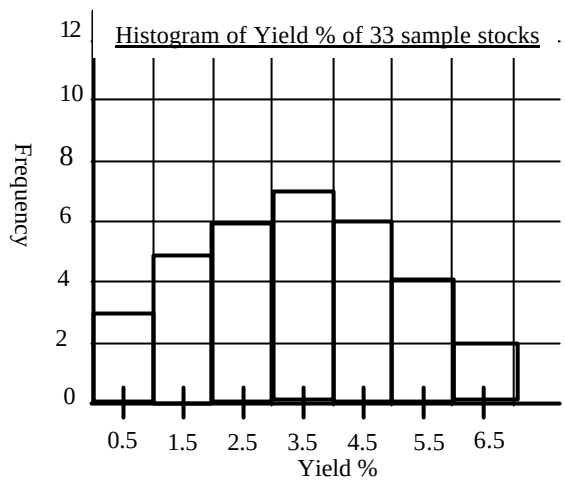



## 1.3 Tabular and graphical presentation of continuous variables

The yield % of 33 sample stocks in Table 1 is a continuous variable. To simplify the presentation of these data, we can group the data into classes. A histogram is used to present these data graphically.

Table 4 Yield % of 33 sample stocks

| Yield % interval | Tally of 33 stocks | Class mark | Frequency |
|---|---|---|---|
| $0 = x < 1$ | /// | 0.5 | 3 |
| $1 = x < 2$ | /// | 1.5 | 5 |
| $2 = x < 3$ | ~~///~~ //// | 2.5 | 6 |
| $3 = x < 4$ | ~~///~~ | 3.5 | 7 |
| $4 = x < 5$ | ~~///~~ / | 4.5 | 6 |
| $5 = x < 6$ | //// | 5.5 | 4 |
| $6 = x < 7$ | // | 6.5 | 2 |
| | | | Sum =33 |



### 1.3.1 Grouped frequency distribution

The steps of constructing a grouped frequency distribution are as follows:

Step 1: Construct the classes
- a. Pick out the highest value and the lowest value and find the range of the data.
- b. Determine the class intervals. Number of intervals should be between 5 and 12 and they usually have equal widths.
- c. Make sure that each item of the data set goes into one and only one class.

Step 2: Tally the data into these classes.

Step 3: Total the tallies in each class to give the class frequency.

Example 1

Suppose 40 students have taken an examination in Mathematics. The marks of the examination are

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 23 | 78 | 61 | 47 | 60 | 42 | 54 | 41 |
| 85 | 55 | 39 | 29 | 88 | 59 | 77 | 78 |
| 81 | 66 | 73 | 94 | 40 | 38 | 60 | 55 |
| 35 | 98 | 82 | 54 | 93 | 76 | 83 | 48 |
| 41 | 67 | 64 | 74 | 97 | 88 | 57 | 69 |

How would you present the results of the students in a frequency table?

Solution

Highest value = 98
Lowest value =:
The range = 98 – _ =

Judging from the range, it will be convenient to divide the data into 8 classes with a class width of 10. To make the scale simple, we start from 20 (which is convenient and is just smaller than the lowest value) and take the class intervals as 20 – 29, etc.

Tally and total the data into these classes.

| Class | Tally (No. of students) | Frequency |
|---|---|---|
| 20 – 29 | // | 2 |
| 30 – 39 | /// | 3 |
| 40 – 49 | ~~////~~ / | 6 |
| 50 – 59 | ~~////~~ / | 6 |
| 60 – 69 | ~~////~~ // | 7 |
| 70 – 79 | ~~////~~ / | 6 |
| 80 – 89 | ~~////~~ / | 6 |
| 90 – 99 | //// | 4 |
| | Total | |

### 1.3.2 Construction of a histogram from a frequency distribution table

- A histogram is a chart that can be used to present grouped data (usually given in a frequency distribution table) graphically.
- This is similar to the bar chart except that the bars are widened to form rectangles.
- Class intervals are shown on the *x*-axis.
- Frequencies are shown on the *y*-axis for equal intervals.
- The width of each rectangle is equal to the class interval. The boundaries of each rectangle correspond to the class boundaries.
- There is no gap between rectangles.
- The mid-point of the base of rectangle corresponds to the class mark. Usually the class marks are labelled along the *x*-axis
- The area of each rectangle is equal to the frequency of that class.

Steps for drawing a histogram from raw data:

1. Set up a frequency distribution table.
2. Determine intervals with class boundaries on the *x*-axis.
3. On each interval, draw a rectangle of height proportional to the number of observations in the interval.

Example 2

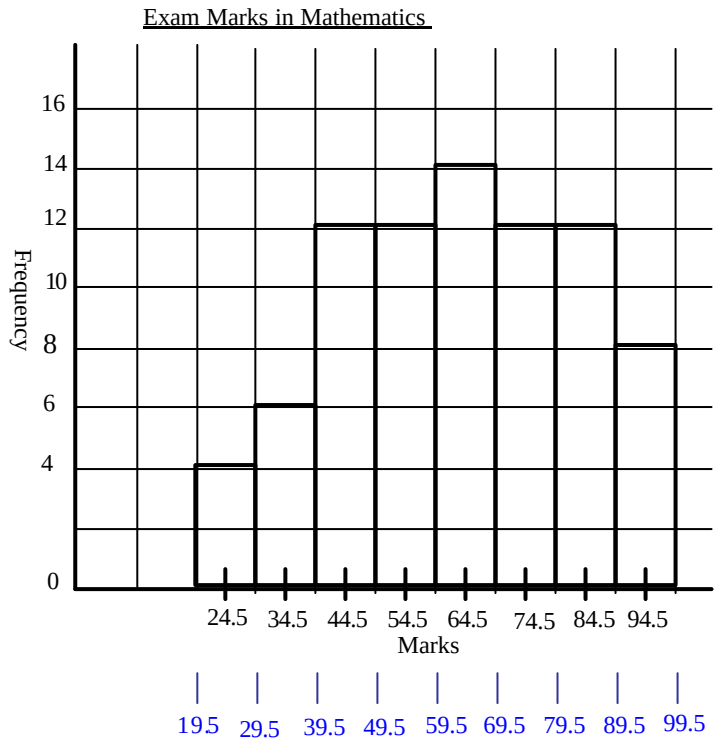
Draw a histogram for the following frequency distribution table:

| Class | 20 – 29 | 30 – 39 | 40 – 49 | 50 – 59 | 60 – 69 | 70 – 79 | 80 – 89 | 90 – 99 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 2 | 3 | 6 | 6 | 7 | 6 | 6 | 4 |

Solution

Consider a marking scheme of an examination, the exam marks are corrected to the nearest integer. An exam mark of 40 corresponds to an actual mark that may be anywhere in the interval from 39.5 up to but not including 40.5.

| Class | Frequency | Class boundary | Class mark |
|---|---|---|---|
| 20 – 29 | 2 | 19.5 – 29.5 | 24.5 |
| 30 – 39 | 3 | 29.5 – 39.5 | |
| 40 – 49 | 6 | 39.5 – 49.5 | 44.5 |
| 50 – 59 | 6 | 49.5 – 59.5 | |
| 60 – 69 | 7 | 59.5 – 69.5 | 64.5 |
| 70 – 79 | 6 | 69.5 – 79.5 | |
| 80 – 89 | 6 | 79.5 – 89.5 | 84.5 |
| 90 – 99 | 4 | 89.5 – 99.5 | 94.5 |


Exam Marks in Mathematics
Frequency
16
14
12
10
8
6
4
0
24.5
34.5
44.5
54.5
64.5
74.5
84.5
94.5
Marks
19.5
29.5
39.5
49.5
59.5
69.5
79.5
89.5
99.5

# 2. Cumulative Frequency and Graphical Representation

## 2.1 Construction of a cumulative frequency table

This table shows how many data are below or above a certain value.
Intervals are joined successively into cumulative intervals.
The cumulative frequencies are found by adding each frequency to the total of the previous ones.

Example 3

Construct a cumulative frequency table from the frequency table below:

| Class boundaries | 19.5 – 29.5 | 29.5 – 39.5 | 39.5 – 49.5 | 49.5 – 59.5 | 59.5 – 69.5 | 69.5 – 79.5 | 79.5 – 89.5 | 89.5 – 99.5 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 2 | 3 | 6 | 6 | 7 | 6 | 6 | 4 |

Solution

| Marks in Mathematics | Cumulative Frequency |
|---|---|
| Less than 19.5 | 0 |
| Less than 29.5 | 2 = 0 + 2 |
| Less than 39.5 | = 2 + |
| Less than 49.5 | 11 = 5 + 6 |
| Less than 59.5 | = 11 + |
| Less than 69.5 | 24 = 17 + 7 |
| Less than 79.5 | 30 = 24 + 6 |
| Less than 89.5 | = 30 + |
| Less than 99.5 | 40 = 36 + 4 |

## 2.2 Construction of a cumulative frequency polygon

A cumulative frequency polygon is a graphical presentation of the cumulative frequency table.

Steps to construct a cumulative frequency polygon:
1. On the *x*-axis, mark the class boundaries.
2. For each *x*, plot a point of *y* ordinate equal to the cumulative frequency.
3. Join the points with line segments.

Example 4

Draw a cumulative frequency polygon from the cumulative frequency table below:

| Marks less than | 19.5 | 29.5 | 39.5 | 49.5 | 59.5 | 69.5 | 79.5 | 89.5 | 99.5 |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | 0 | 2 | 5 | 11 | 17 | 24 | 30 | 36 | 40 |

Hence
a) find number of students
 i) who passed the examination if the passing mark is 40;
 ii) who got distinction if the distinction mark is 85; and
b) the passing marks if the passing rate of the class is 40%.

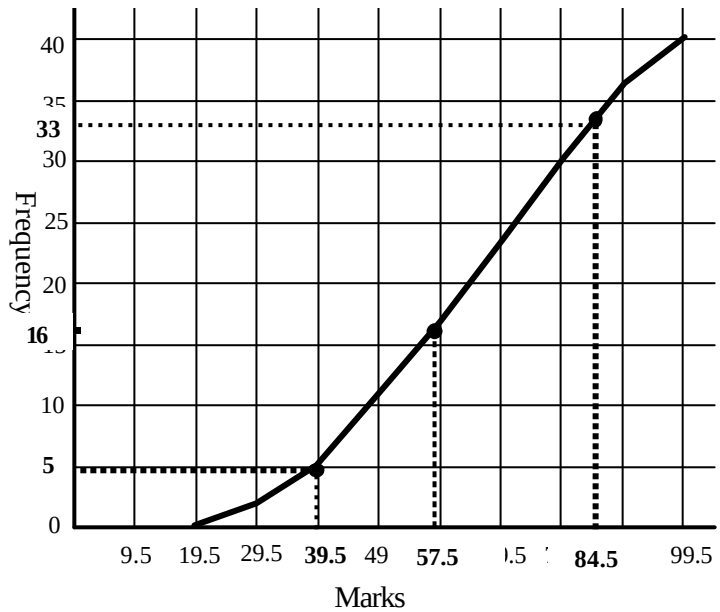
Solution

5 students got marks less than 39.5.
The no. of student passed in the examination is 40 – 5 = 35

33 students got marks less than 84.5, the no. of students obtained distinction award is
40 – =

The no. of student failed
= 40 (1 – %) =
The passing mark is

Less than Cumulative frequency polygon for marks of maths




## 3. Measure of central tendency: mean, median and mode

For a set of data, we determine a quantity used to summarise the whole set of data. This quantity is termed a measure of central tendency. The most commonly used measures are mean, medium and mode.

| **3.1** | For ungrouped data | For grouped data, |
|---|---|---|
| **Mean** | $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$ | $\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{f_1 + f_2 + f_3 + \ldots + f_n}$ |

Example 5

a) Find the mean of the set of data: 25, 36, 42, 38, 36
b) Find the mean from the set of grouped data

| Class mark | 10.5 | 30.5 | 50.5 | 70.5 | 90.5 | 110.5 |
|---|---|---|---|---|---|---|
| Frequency | 19 | 6 | 3 | 2 | 1 | 2 |

Solution

a) mean $\bar{x} = \frac{25+36+42+38+36}{5} = 35.4$

b)

| $x$ | $f$ | $xf$ |
|---|---|---|
| 10.5 | 19 | 199.5 |
| 30.5 | 6 | 183.0 |
| 50.5 | 3 | 151.5 |
| 70.5 | 2 | 141.0 |
| 90.5 | 1 | 90.5 |
| 110.5 | 2 | 221.0 |
| sum | 33 | 765.5 |

mean = ____ ÷33 =

**Use calculator to check your answer.

Example 7

The HK Consumer Price Index B from 1996 to 2001 was as following:

| Year | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|---|---|
| Index | 99.7 | 105.5 | 108.5 | 103.4 | 99.4 | 98.1 | 95 |

Calculate the average consumer price index B:

a) For the first 4 years, (1996 – 1999).
b) For the next 3 years, (2000 – 2002)
c) For all 7 years
d) Suppose the original data was lost, and only the 4- and 2-year averages in a) and b) were available. Would it still be possible to calculate the overall 6-year average? How?

Solution

a) From 1996 - 1999, n = 4.

The average price index = (99.7 + 105.5 + 108.5 + ____) ÷ __ =

b) From 2000 - 2001, n = 2.

The average price index = (99.4 + 98.1 + 95.0) ÷ 3 = 97.5

c) From 1996 - 2001, n = .

The average price index
= (99.7 + 105.5 + 108.5 + 103.4 + 99.4 + 98.1 + 95.0) ÷
=

d) The average price index over 7 years = $(104.3 \times 4 + 97.5 \times 3) \div (4 + 3)$
= 101.4

| **3.2** | For ungrouped data | For grouped data, |
|---|---|---|
| **Median** | 1. the middle datum, when $n$ is odd<br>2. the mean of the two middle data, when $n$ is even. | Step 1: Draw the cumulative frequency polygon.<br>Step 2: The median is the datum corresponding to the middle value of the cumulative frequency. |

Example 6

a) Find the median of 2, 3, 10, 12, 999.
b) Find the median of 2, 3, 10, 12, 22, 123.
c) The provisional figures on the population by age group in Hong Kong as at 9/2001 are tabulated below. Draw a cumulative frequency polygon and determine the median age for the population.

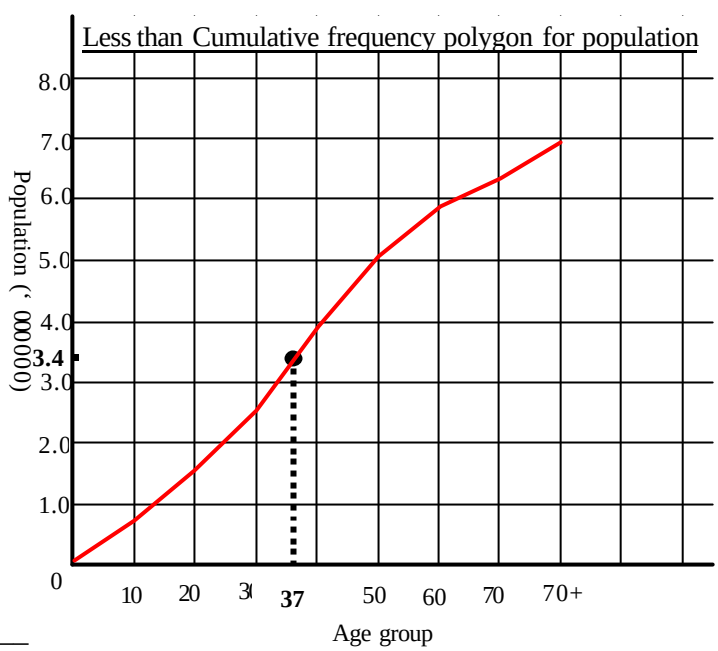
| Age group | 0 – 9 | 10 – 19 | 20 – 29 | 30 – 39 | 40 – 49 | 50 – 59 | 60 – 69 | ≥ 70 |
|---|---|---|---|---|---|---|---|---|
| Population ('000) | 676 | 885 | 1000 | 1267 | 1208 | 677 | 503 | 499 |

Solution

a) Median =

b) Median = (10+ )÷2 =

| Age group $x$ | Population ('000) | cumulative population ('000 000) |
|---|---|---|
| $x < 10$ | 676 | 0.676 |
| 10 ? $x < 20$ | 885 | |
| 20 ? $x < 30$ | 1000 | 2.561 |
| 30 ? $x < 40$ | 1267 | 3.828 |
| 40 ? $x < 50$ | 1208 | |
| 50 ? $x < 60$ | 677 | 5.713 |
| 60 ? $x < 70$ | 503 | 6.216 |
| $70 < x$ | 499 | |




c) The rank of median = ____/2 = 3.____
From the cumulative polygon, the median age = ____

| 3.3 | For ungrouped data | For grouped data, |
|---|---|---|
| **Mode** | the datum that has the highest frequency | modal class is the class that has the highest frequency |

Example 7

a) Find the mode of the data:
   1, 2, 2, 2, 3, 3, 9
b) Find the modal class

| Class | 10 - 14 | 15 - 19 | 20 - 24 | 25 - 29 |
|---|---|---|---|---|
| Frequency | 2 | 8 | 7 | 3 |

Solution

a) The mode is .
b) The modal class is 15 – .

Remark
Mean seems to be the most commonly used (and often misused) quantity for measuring central tendency. If the distribution of the data set shows a strong degree of skewness, then mean is not a reliable measure as it is strongly affected by the extreme values. In this case, medium may be a better choice. Mode is used when there is reason to choose the most commonly occurring data value as the representative for the whole data set.

## 4. Measure of dispersion: Range and Inter-quartile range Standard deviation

Apart from using a measure of central tendency to summarise a set of data, we need a quantity to measure the degree of dispersion of the set of data (so that we can determine the reliability of the set of data). Range is a measure that is very simple to use but it provides relatively little information on dispersion. Quartile deviation is used in association with the median whereas standard deviation goes with the mean.

| 4.1 | For ungrouped data | For grouped data, |
|---|---|---|
| **Range** | the difference between the largest datum and the smallest datum. | the difference between the highest class boundary and the lowest boundary. |

Example 8

a) Find the range of the data:
1, 2, 2, 2, 3, 3, 9
b) Find the range of the grouped data

| Class | 10 - 14 | 15 - 19 | 20 - 24 | 25 - 29 |
|---|---|---|---|---|
| Frequency | 2 | 8 | 7 | 3 |

Solution
a) The range = 9 – = .
b) The range = 29.5 – ___ =

## 4.2 Inter quartile range

Inter quartile range = $Q_3 - Q_1$

where $Q_1, Q_2, Q_3$ are called quartiles which divide the data (which have been ranked, i.e. arranged in order) into four equal parts.

Moreover,
$Q_2$ is the median of the whole set of data,
$Q_1$ is the median of the lower half,
$Q_3$ is the median of the upper half.
Quartile deviation, Q.D. = ½ $(Q_3 - Q_1)$

Example 9

The following frequency distribution gives the life hours of a sample of 50 light bulbs:

| Life hours ('000) | Frequency | Cumulative frequency |
|---|---|---|
| 0.6 to under 0.7 | 2 | 2 |
| 0.7 to under 0.8 | 4 | 6 |
| 0.8 to under 0.9 | 6 | 12 |
| 0.9 to under 1.0 | 14 | 26 |
| 1.0 to under 1.1 | 13 | 39 |
| 1.1 to under 1.2 | 7 | 46 |
| 1.2 to under 1.3 | 4 | 50 |

Find the median and the inter-quartile range of the data.

Solution

The rank of median
= ½ ×50 = 25
The median of life hours is 980 hrs.

The rank of upper quartile
= ¾ ×50 =
= , to the nearest integer
The upper quartile $Q_3$ is 1090 hrs

The rank of lower quartile
= ¼ ×50 =
= 13 , to the nearest integer
The lower quartile $Q_1$ is 900 hrs.

The inter-quartile range = $Q_3 - Q_1$
= 1090 –
= hrs.

Quartile deviation = ½ ($Q_3 - Q_1$)
= ½ (1090 – ____) =

Less than cumulative frequency polygon
for life hours of 50 sample light bulbs