Unit 10: Basic Statistical Measures And Their Interpretations

10.1 Measure of central tendency: Mean, Weighted Mean, Median and Mode.

This is a revision topic. Both the methods of displaying data and the measures of central tendency of a data set should be reviewed. They should know the usefulness, merits and dismerits of each of the above measures. In order to arouse students' interest, more daily life examples should be given.

Everyday examples: (These examples are extracted from the packages Living with Statistics produced from Census and Statistics Departments, H.K.)

(i) The mean height of the students in a class is 1.73 m.
(ii) The mean daily maximum air temperature for the year 1994 was 25.9 °C.
(iii) The median income of households in Hong Kong in the 3 rd quarter of 1994 was $15,000.
(iv) The median age of men who married in Hong Kong in 1993 was 30. (Only first marriages are considered in this example.)
(v) The modal household size in Hong Kong was 4 as revealed in the 1991 Population Census.
(vi) Manufacturers would need to consider modal sizes for various commodities. A shirt manufacturer for example would obviously need to make more shirts of the modal size than any others.

More examples are given below:

Example 1
Get the values of something which is published in the newspaper every day such as share prices, world temperatures, etc. and compare them on different days using measure of central tendency.

Example 2
Write down the amount of pocket money the students get each week (correct to HK $10). Determine the mean, median and mode of the amounts. Which of these would you regard as the best indicator of the average pocket money? Students may notice in their examples that there may be more than one mode.

Example 3
The annual salaries bill of a small business company is as follows:

Director	$ 2,400,000
Manager	$ 2,000,000
Salesman	$ 300,000
Chief operator	$ 150,000
3 Operators at	$ 100,000 each
2 Secretaries at	$ 90,000 each
1 Apprentice	$ 80,000

Find the mode, the median and the mean salary of these 10 persons. Which of these would you regard as the best indicator of the average salary? Explain your answer.

Weighted mean

The concept of weighted mean is commonly known in daily life. The mean of grouped data is in fact a mean of the class marks weighted by the class frequencies. The examination scores use the idea of weighted mean.

Weighted means have important application in the construction of **index numbers,** such as price index.

Index numbers

An index is a number used to measure the relative change in a set of observations over time compared to a base period. The index is taken as 100 in the base period. An index number is expressed as a percentage of the base period..

There are different price indexes in Hong Kong, for example, Consumer Price Index(A), Consumer Price Index(B), and the Hang Seng Price Index. They differ in weighting systems to reflect the different expenditure patterns of different households.

Example 1
The Consumer Price Index(A) covers about 50 % of the urban households in Hong Kong with a monthly expenditure of between HK $2,500 and HK $ 9,999 in the base period October 1989 to September 1990. Subsequent index numbers measure the relative change of prices of consumer goods and services as a percentage.

Example 2
The Hang Seng Index of Hong Kong Stocks was taken as 100 on 31 July 1964. It was 14,864 on 29 September 1997, showing an increase of over 100 times in market value.

Price Index

Commodity	Price in base period	Price period t	Weight
1	c_{01}	c_{t1}	w_1
2	c_{02}	c_{t2}	w_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	c_{0k}	c_{tk}	w_k

A price index for the group of commodities under consideration for the period t

$$= \frac{\text{weighted mean of the prices in the period } t}{\text{weighted mean of the prices in the base period}} \times 100$$

$$= \frac{\dfrac{w_1c_{t1} + w_2c_{t2} + w_3c_{t3} + \cdots + w_kc_{tk}}{w_1 + w_2 + w_3 + \cdots + w_k}}{\dfrac{w_1c_{01} + w_2c_{02} + w_3c_{01}3 + \cdots + w_kc_{0k}}{w_1 + w_2 + w_3 + \cdots + w_k}} \times 100$$

$$= \frac{w_1c_{t1} + w_2c_{t2} + w_3c_{t3} + \cdots + w_kc_{tk}}{w_1c_{01} + w_2c_{02} + w_3c_{03} + \cdots + w_kc_{0k}} \times 100$$

<u>Example 3</u>
Find the price index of the group of commodities shown in table below for 1995 using 1992 as the base year.

Commodity	Price in 1992	Price in 1995	Weight
Rice	28	30	5
Pork	21	26	4
Fish	30	36	4
Vegetables	6	8	3
Others	4	6	2

The required price index

$$= \frac{5\times30+4\times26+4\times36+3\times8+2\times6}{5\times28+4\times21+4\times30+3\times6+2\times4} \times 100$$

= 117.3

Suggested Index number project

(i) An index of personal expenditure on leisure items.
(ii) An index of traffic expenditure of the students.

Applications of Central Measures

One way of looking at this is by an example:

Mr. Chan is employed to replace broken windows. From the records of the last few months, his daughter has done some calculations. She finds that the average time to replace a window-pane can be given as follows.

Mean = 2.0 hours
Mode = 1. 8 hours
Median = 1. 5 hours

How can Mr. Chan use these figures?

(i) If 100 windows in a school had been broken by thunderstorm, it would take him a total of $100 \times 2.0 = 200$ hours to replace all of them. The <u>mean</u> is the measure to use for this sort of calculations.

(ii) If Mr. Chan is asked to state how long he takes to replace a window, then the most likely time he will take is the mode.

(iii) Replacing windows on a one-off basis is Mr. Chan's normal job. He is asked by a company to replace a window within 2.0 hours. Will Mr. Chan accept such requirement to do the job? Why?

Through other examples, advantages and disadvantages of the mean, median and mode can be discussed.

10.2 Measures of Dispersion

At the beginning of the lesson, teachers are suggested to introduce two sets of data with same mean and same median but different dispersion. They should point out that it is not sufficient to represent data with a measure of central tendency only. It is also important to have a measure of the way in which the data are distributed or dispersed about the mean. This measure is known as the measure of dispersion.

Types of measure of dispersion

The following measures of dispersion are in common use:

(i) range

(ii) interquartile range or semi-interquartile range

(iii) variance and standard deviation

The definition of each term can then be illustrated with some concrete examples. If possible, graphs are suggested to be used in illustrating different measures. Like many mathematical concepts, visual representations of these statistical concepts are often easier to understand than: the symbolic or numerical representations.

(i) Range

Daily life examples: Weather reporting; Temperature difference; Fluctuation of stock price; Fluctuation of water wave.

Two different distributions which have the same range may be illustrated graphically.

(ii) Interquartile range and Semi-interquartile range

Daily life examples: Salary; Academic results of students in an entrance examination.

The method of interpolation can be used for estimation of quartiles, however it depends on the nature of the set of data

Note: Instead of using quartiles, percentiles may be used.

(iii) Variance and Standard deviation

The range, interquartile range and semi-interquartile range are usually ineffective to measure the dispersion of a set of data. When some data change, these measures of dispersion may remain unchanged. An useful measure that describes the dispersion of **all** the values is the **variance or standard deviation.**

Example 1

Find the standard deviation of the height of boys and girls of different ages and draw appropriate conclusions.

Example 2

The following are two sets of data of an experiment obtained by two different students:

	Volume of acid measured (cm^3)									
Student A	8	12	7	9	3	10	12	11	12	14
Student B	7	6	7	15	12	11	9	9	13	11

(i) What is the mean volume of acid measured by each student?

(ii) What is the standard deviation?

(iii) Which set of results is more reliable?

Compare and Contrast various measures

The advantages and disadvantages of respective measures of dispersion should be stressed.

10.3 Frequency distribution and Cumulative frequency distribution, their graphical representations and interpretation

Teachers are suggested to revise this topic through some concrete daily life examples.

The significance of and difference between grouped and ungrouped data should be stressed. Special attention should be paid to the suitability of a graph in providing information and conclusions of a certain type. Emphasis should be laid on interpretation and also the use of graphs for prediction.

New advances in technology have resulted in software packages that construct graphs for given data sets. It is recommended that once students have developed an understanding of various types of graphs, they can use the computer as an useful tool to focus on analyzing similarities and differences in their different forms. Teachers may use computer as an teaching aid to facilitate their students' understanding.

The following topics are suggested as a class project for investigation:

1. Topic: Body Temperatures

   Problem: Does everyone have the same body temperature? Do males and females have similar temperatures? Are females warmer than males? Are their temperatures more variable?

   Collection of data: Using clinical thermometers take the temperatures of 50-100 people. Most clinical thermometers now measure in degrees Celsius and will require readings to be taken to the nearest 0.1°C.

2. Topic: Traffic distributions

   Aim: To investigate the distribution of the number of vehicles passing a particular point in one minute intervals.

   Collection of data: Choose a fairly busy road on which you might expect a uniform flow of traffic. Decide what type of vehicles are to be included in the count and where the passing points is to be. Count the number of vehicles passing in one minute intervals, in one direction, over a period of one hour. Repeat, if time allows, at a point where traffic flow is interrupted, e.g. near traffic lights, or junctions. (In a class project, a variety of sites could be chosen.)

3. Topic: Pocket money of Sixth-form students

   Problem: Do male and female sixth-form students have similar amount of pocket money each week?

Collection of Data: Ask about 50 sixth-form students the amount of pocket money they have (which should be a week in school term time). If they do not have any pocket money, record the answer as zero. Record males and females separately.

4. Devise and carry out opinion polls on a number of different topics. (e.g. pop music, political parties, educational standards etc.)

5. How many French fries do you get in an order of "medium fries" at a snackbar? Students try to do so, then realize that they must agree on how to handle the broken pieces of French fry. Soon the class has several different suggestions. How shall we describe the results? What do you see? Experienced students may discuss the distribution about their length or shape. How small a count of fries will lead you to make a complain? Which is the best buy: large, medium, or small fries? How does this snackbar compare with another one?

10.4 Box-and-whisker diagrams

The box-and-whisker diagram (also known as box plot) is used for the purpose of displaying important features of a set of data, including the minimum value, lower quartile, median, upper quartile, and maximum value of the data.

A box-and-whisker diagram is constructed on a horizontal scale as follows:

Step 1. Draw a narrow box that reaches from the lower quartile to the upper quartile on the scale;
Step 2. Draw a vertical bar across the box at the median;
Step 3. Draw a line segment from the left end of the box to the minimum value; and
Step 4. Draw a line segment from the right end of the box to the maximum value.

The two line segments drawn in Steps 3 and 4 are called whiskers.

The general configuration of a box-and-whisker diagram is shown in Fig. 10.4(i).

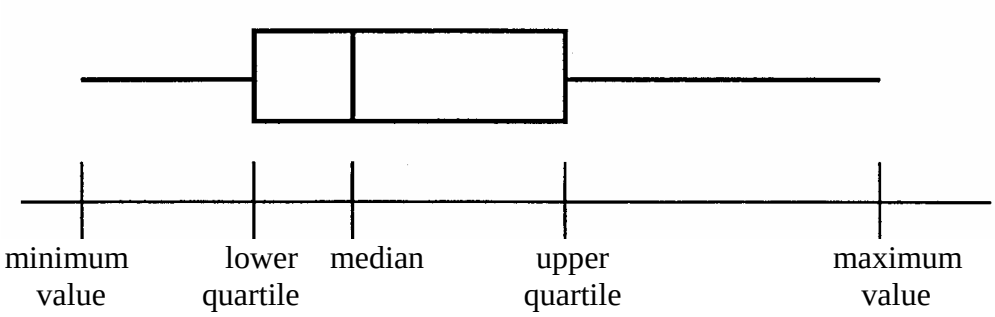


Fig. 10.4 (i)

The width of the box of a single box-and-whisker diagram can be as narrow or as wide as a user wishes without affecting the meaning of the display.

The construction of a box-and-whisker diagram is illustrated by the following example.

Example 1
The weekly expenditure on soft drinks of 20 football players are given in the following table:

Expenditure on soft drinks
of 20 football players (in dollars)

26	34	31	36	16
27	32	21	43	41
30	6	15	38	27
58	35	28	21	20

From these figures,

Maximum value	= 58
Upper quartile	= 35.5
Median	= 29
Lower quartile	= 21
Minimum value	= 6

The diagram is drawn as in Fig. 10.4(ii) below.

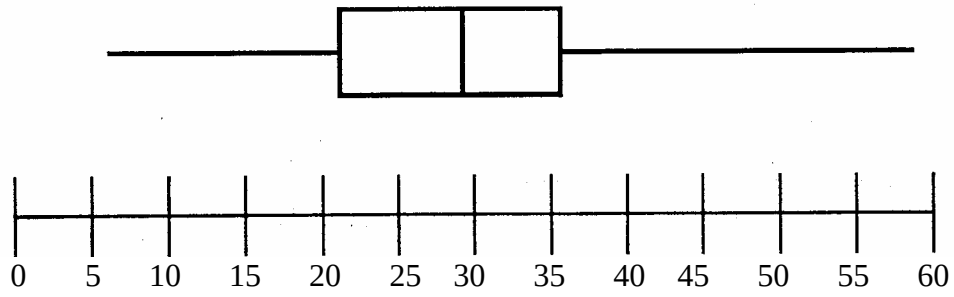


Fig. 10.4(ii) Box-and-whisker diagram for the expenditure on soft drinks data

A box-and-whisker diagram can also be constructed on a vertical scale by making appropriate adjustments in the foregoing steps (refer to Fig. 10.4(i))

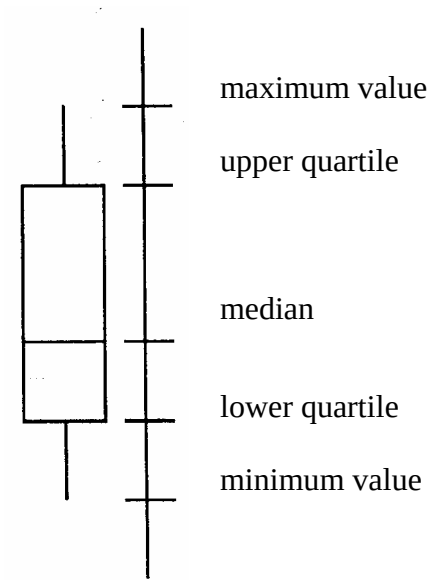


Fig. 10.4(iii)

A box-and-whisker diagram depicts the location of the centre, spread and symmetry of a set of data.

1. The bar at the median shows the location of the centre of the data.
2. The length of the box shows the spread of the middle half of the data.
3. The lengths of the whiskers show the spread of the lower and upper quarters of the data.

4. The shape of the diagram gives a quick impression of the degree of symmetry of the distribution of the data about the median.
   If the diagram is symmetrical, then the set of data is likely to be symmetric with respect to the median. Variations from symmetry in the diagram reflect similar variations in the set of data (see Fig. 10.4(iv) and 10.4(v)).

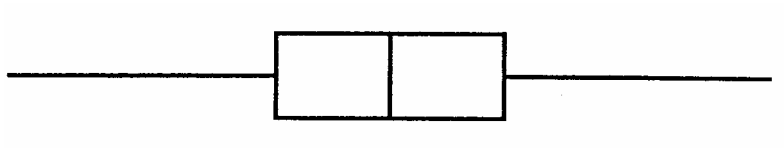


Fig. 10.4(iv) A distribution symmetrical about the median.

Fig. 10.4(v) Distributions not symmetrical about the median.

Box-and-whisker diagrams are also effective for comparing different sets of data. Features such as location of centres, spread and symmetry .for these sets of data could be compared.
Fig. 10.4(vi) shows 3 box-and-whisker diagrams of the salary distributions of 3 cities.

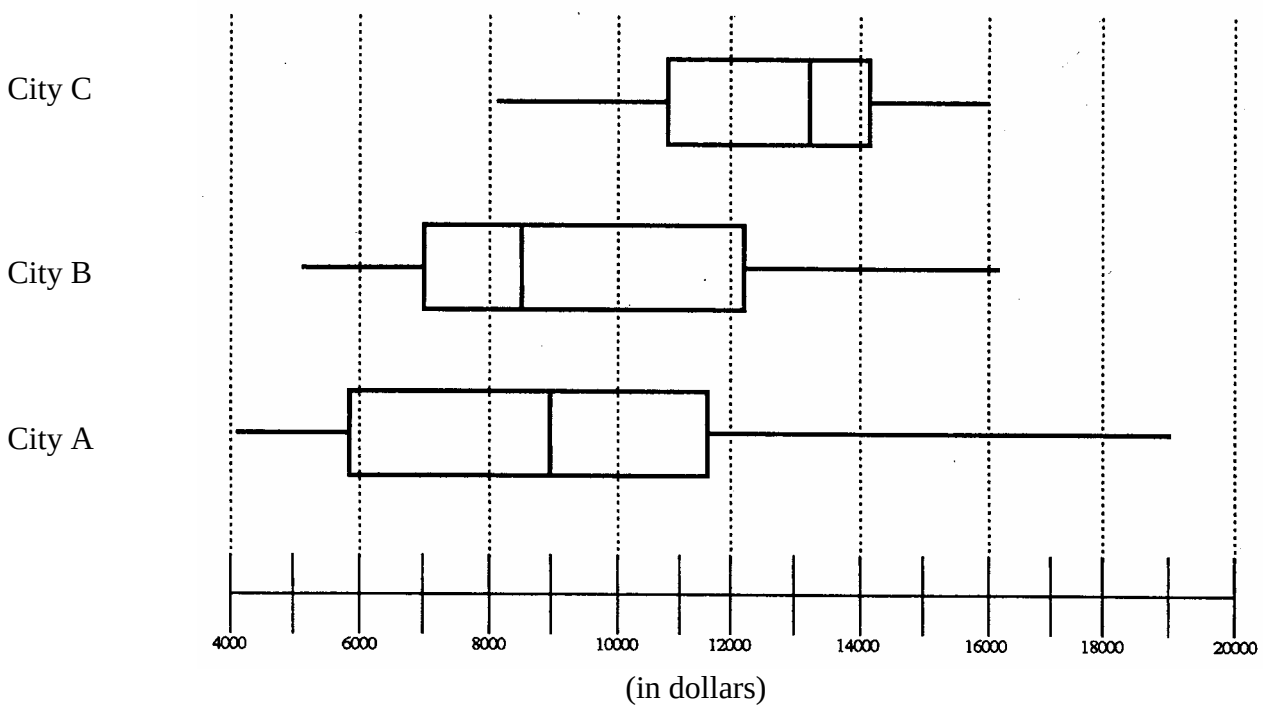


Fig. 10.4(vi) The salary distributions of 3 cities

It can be seen that City C has the highest median salary, City A has the largest salary spread, and nearly half of the people in City B have salary lower than the minimum salary of City C.

10.5 Stem-and-Ieaf diagrams

A stem-and-leaf diagram is a combination of a table and a graph. It visually resembles the numerical histogram and is basically a refined grouped frequency distribution. Instead of tally marks, usually the last digit of the data values, which are called '**leaves**', are used to indicate the observation. These observations are grouped according to the one or more digits preceding the leaves, which are called '**stems**' and arranged in an ascending order and listed to the left of the leaves.

Let's illustrate the construction of the stem-and-leaf diagram of the marks of 40 students in a mathematics test:

49	73	58	59	40	61	45	61
51	43	80	55	81	52	54	45
58	53	55	38	59	72	46	68
42	65	65	64	29	48	39	67
47	56	66	48	42	56	51	53

Step 1. Draw a vertical straight line

Step 2. Select the leading digits (**stems**) from the data.

Step 3. Arrange the stems in an ascending order and list them on the left-hand side of the vertical line.

Step 4. Go through the data and write the last digits of the numbers (**leaves**) to the right of the appropriate stems one after the other in ascending order.

Stem (in 10)	Leaves (in 1)
2	9
3	89
4	02235567889
5	11233455668899
6	11455678
7	23
8	01

Fig. 10.5 (i)

Example 1

Stem (in 100)	Leaves (in 10)			
1	11	23	47	48
2	23	27		
3	55	63	80	
4	21	23		
5	00			

Fig. 10.5(ii)

In the above diagram, **stems** are hundreds and **leaves** are units.
The set of data in the diagram represents:
111, 123, 147, 148, 223, 227, 355, 363, 380, 421, 423, 500

Example 2
A householder's weekly consumption of electricity in kilowatt-hours during a period of nine weeks in a winter were as follows:

338, 354, 341, 353, 351, 341, 353, 346, 341.

For the electricity of consumption data the first figure in each observed value is 3, so there is no variation. Then, looking at the second figure there is variation, since it may be either 3, 4 or 5. So to the left of the vertical line we put in the figures 33, 34 and 35. To the right of the vertical line we record the remaining figure in each observed value (the 'leaves'). Finally we order the figures in each row to the right of the vertical line by size to give the completed stem and leaf diagram shown in Fig. 10.5 (iii).

Stem (in 10)	Leaves (in 1)			
33	8			
34	1	1	1	6
35	1	3	3	4

Fig. 10.5 (iii)

Example 3
Examination results of 11 students:

English:	23,	39,	40,	45,	51,	55,	61,	64,	65,	72,	78
Chinese:	37,	41,	44,	48,	58,	61,	63,	69,	75,	83,	89

One way to compare their performances in the two subjects is by means of side by side stem-and-leaf diagrams.

English

Stem (in 10)	Leaf (in 1)
2	3
3	9
4	0 5
5	1 5
6	1 4 5
7	2 8
8	

Chinese

Stem (in 10)	Leaf (in 1)
2	
3	7
4	1 4 8
5	8
6	1 3 9
7	5
8	3 9

Fig. 10.5(iv)

The comparison can be made more dramatic by back-to-back stem-and-leaf diagram.

English Leaves (in 1)	Stem (in 10)	Chinese Leaves (in 1)
3	2	
9	3	7
5 0	4	1 4 8
5 1	5	8
5 4 1	6	1 3 9
8 2	7	5
	8	3 9

Fig. 10.5(v)

For some data, in order to make the diagram more compact, outlying values may be listed separately, low values at the low end and high values at the high end. For example, the following set of data may be displayed in the form as in Fig. 10.5(vi) or Fig. 10.5(vii).

12.0	21.2	23.1	24.8
22.0	25.9	29.8	31.9
21.2	21.6	22.8	22.5
26.3	25.1	21.7	21.6
23.5	27.0	38.5	28.5

Stem (in 1)	Leaves (in 0.1)
LO = 12.0	
21	2 2 6 6 7
22	0 5 8
23	1 5
24	8
25	1 9
26	3
27	0
28	5
29	8
HI = 31.9, 38.5	

Fig. 10.5(vi)

Stem (in 1)	Leaves (in 0.1)
LO = 12.0	
21	2 2 6 6 7
22	0 5 8
23	1 5
24	8
25	1 9
26	3
27	0
28	5
29	8
30	
31	9
HI = 38.5	

Fig. 10.5(vii)

10.6 First Notion of population parameters and sample statistics

We introduce the concept of a population and its parameters as well as the concept of a sample drawn from the population and statistics calculated from the sample. One of the purposes of drawing a sample is to obtain estimates for the population parameters from which we can make statistical inference. The following terms should be introduced through examples: **a population, a sample, sample variance and sample mean.** Examples to clarify the ideas of such terms can then be given.

Population and Samples

Example 1

Ages distribution:	The distribution of the ages of all students in a secondary school.
The population:	All students in a secondary school.
Sample:	To draw 100 students randomly from the school.

Example 2

TV viewers:	A survey is conducted to investigate the most popular TV program during 6 pm to 10 pm.
The population:	All the viewers during 6 pm to 10 pm.
Sample:	To draw 800 persons randomly from the population.

Example 3

Quality control:	Check the life-time of electric light bulbs produced by a factory.
The population:	All electric light bulbs produced.
Sample:	To select 30 light bulbs randomly for inspection.

Example 4

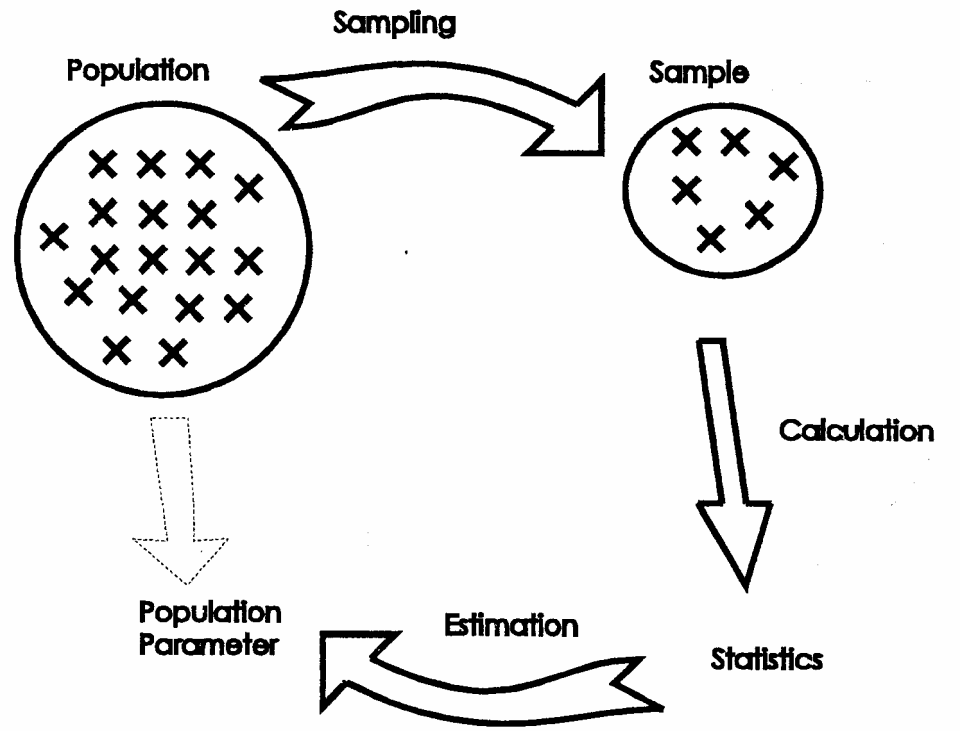
Suppose that the manager of the Customer Service Division of Hong Kong was interested primarily in determining whether customers who had purchased a video cassette recorder over the past twelve months were satisfied with their products. Using the warranty cards submitted after the purchase the manager was planning to survey 1,425 of these customers.

i) Describe both the population and sample of interest to the manager.

ii) Describe the type of data that the manager primarily wishes to collect.

A population parameter is a summary measure describing a characteristic of the entire population, while a **statistic** is a summary measure that is computed from a sample to describe a characteristic of the population.

Teachers may illustrate the terms with the following diagram:



Sample Mean and Sample Variance

Students are expected to know the following:

(i) sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$, where $n = \sum_{i=1}^{k} f_i$;

(ii) sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right)^2$, where $n = \sum_{i=1}^{k} f_i$;

(iii) Sample mean $\bar{x}$ and the sample variance s^2 tend to population mean μ and population variance σ^2 respectively as the sample size is sufficiently large.

(iv) for finite population, population variation

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} f_i\left(x_i - \mu\right)^2$, where N is the population size.

Students are not required to understand the meaning of the **mean of the sample mean** and the **variance of the sample** mean.

<u>Example 1</u>

A metal coin is marked with the numbers 3 and 4 on opposite sides. It is biased so that on the average for every five tosses, 3 appears four times and 4 appears one time. We shall take these numbers as our population: 3, 3, 3, 3, 4.

The population mean $\mu = \frac{3+3+3+3+4}{5} = 3.2$

and the population variance $\sigma^2 = \frac{4\times(3-3.2)^2 + (4-3.2)^2}{5} = 0.16$

Suppose the students do not know how it is biased. They throw this coin three times experimentally, all the possible samples from the population are (3, 3, 3), (3, 3, 4), (3, 4, 3), (4, 3, 3), (3, 4, 4), (4, 3, 4), (4, 4, 3), (4, 4, 4).

Take this sample (3, 3, 3) as an example:

The sample mean, $\overline{x} = \frac{3+3+3}{3} = 3$

The sample variance $s^2 = \frac{(3-3)^2 + (3-3)^2 + (3-3)^2}{3-1} = 0$

Take another sample (3,4,4) as an example:

The sample mean, $\overline{x} = \frac{3+4+4}{3} = 3\frac{2}{3}$

The sample variance $s^2 = \frac{(3-3\frac{2}{3})^2 + (4-3\frac{2}{3})^2 + (4-3\frac{2}{3})^2}{3-1} = \frac{1}{3}$

Teachers can then point out that the sample mean and the sample variance are not identical to the population mean and population variance.