

CHAPTER ONE : INTRODUCTION TO STATISTICS

1. What is statistics?

Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting numerical or quantitative data for the purpose of assisting in making a more effective decision (**From data to information**).

2. What is the application of statistics?

The application of statistics to our life is definitely wide. If you look through your university catalog, you will find that statistics is required for many college programs such as School of Engineering, School of Business and Psychology or Sociology Department. However, the course content is the same. In business and economics, we are interested in profit, cost, demand and supply quantity. Psychology is interested in IQ and EQ test scores. In engineering, they may be interested in production quantity on a particular machine.

3. Why we study statistics?

We must study and learn statistics because we are living in the world of numeric information. Everyday, when you read the newspaper, watch television, or walk down the street, you may hear people say:

- The Ministry of Education has reported that the total number of students enrolled in local universities has increased from 30,000 last year to 37,000 this year.
- The economists predict that the prices of national cars will drop by at least 50% to 60% after the imposition of AFTA.
- One major research conducted by a group of university students in 2000 revealed that the life expectancy of a male in Malaysia is 72 years and of a female is 75 years.

All the examples above are the language of statistics. People need statistics to plan their budgets, companies need statistics to control their stocks and cash flows, banks need statistics to reduce their risks, investors need statistics to make a decision on the area of their investments. Now, you know why we need statistics, the more informative you are, the more powerful you are! Statistics can tell you much information just simply from a set of number – **data**.

4. Types of Statistics

In general, statistics study is done in two ways and we can classify them into descriptive statistics and inferential statistics.

- **Descriptive Statistics:**

A method of organising, summarising and presenting data in an informative way. Examples are histogram, pie chart, bar chart, box plot, etc.

- **Inferential Statistics:**

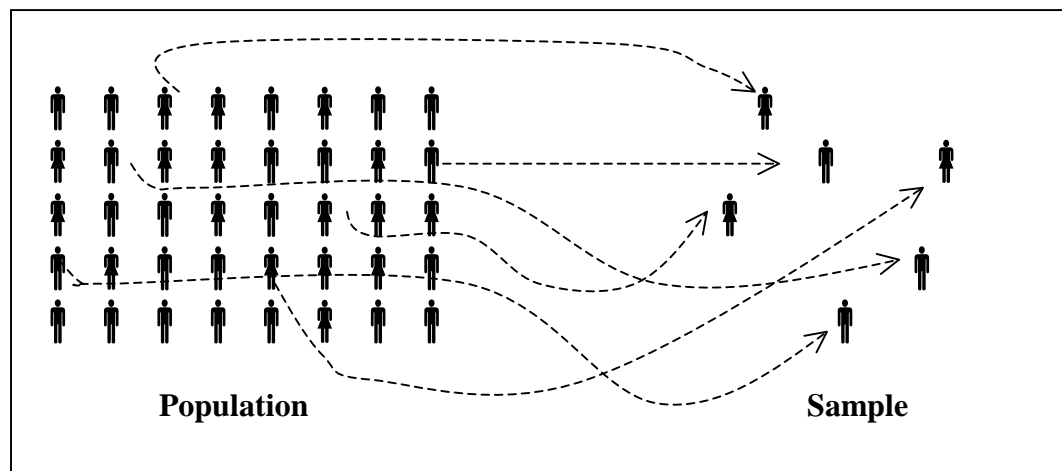
A decision, estimate, prediction, or generalization about a population based on a sample.

- **Population:** A population is a collection of all possible individuals, objects or measurements of interest.
- **Sample:** A sample is a portion or part of the population of interest.

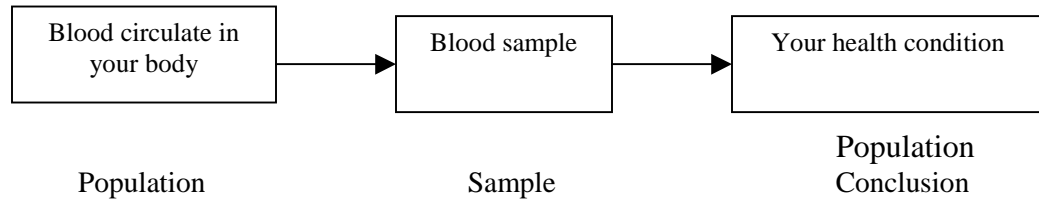
Example: An investor may be interested in studying the influence of rock music among the students of Malaysia.

Population: All the students of Malaysia.

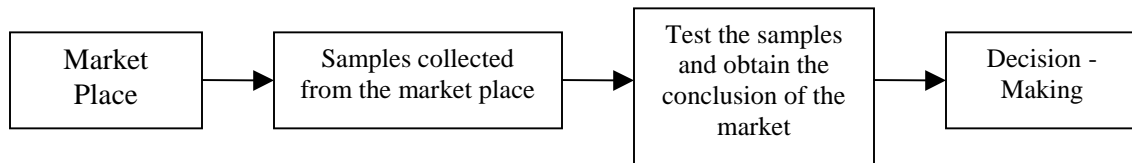
Sample: Since Stamford College gets students from all parts of Malaysia and is represented by all communities, students in Stamford College may be taken to be the sample. The sample information can be collected and inferences can be made about the population.



To think on how inferential statistics works; you may just think on the process of your medical checkup. The doctor only takes a small portion of blood sample from your body for testing and then can conclude the overall condition of your health.



The same sampling principle applies in business decision-making:



For example, if one new computer manufacturer wants to launch its CPU in the market for the first time, the marketing manager may need to collect some sample prices of other brands of CPU in the market such as Dell, Acer, IBM, Toshiba and so on to make a comparison. After the comparison and investigation in the price and other areas, the marketing manager will decide to set a reasonable price for its CPU in order to compete with its competitors in the market.

5. Why we want to study sample?

Studying sample is easy, less costs, save time, sometimes it may not be possible to study the entire population – endurance testing of electric bulbs.

6. Types of Variables

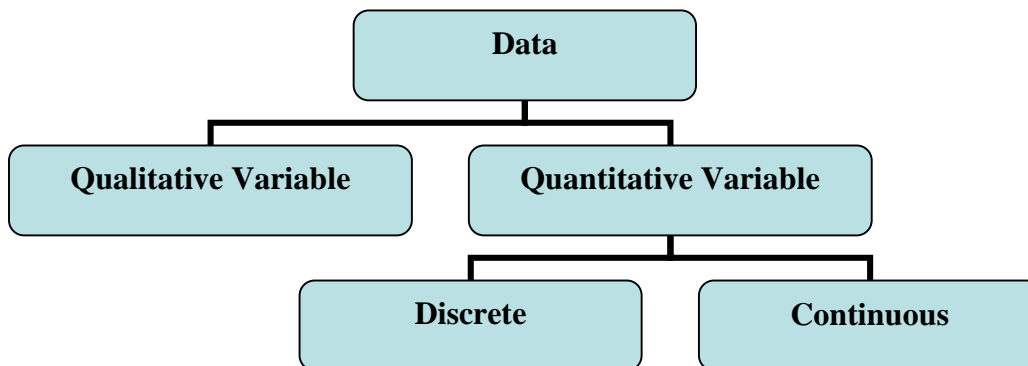
- **Variable:** The characteristics of each individual, object or measurement of interest in a population that we want to study for statistics purpose. In general, there are two types of variables – Qualitative variable and quantitative variable.

Example: To study the living standard of residents in one area, we are interested to study the variables – sex, age and monthly income of the residents in that area.

- **Qualitative Variable:** The characteristic of the variable being studied is non-numeric. Example: Color, gender, occupation, quality and grade.
- **Quantitative Variable:** The variable can be reported numerically and can be classified as :

- a. **Discrete Variable:** The value of the variable can be measured precisely or the value is any integer of 1, 2, 3, 4, ..., 100, 101, 102, ..., K. Example: The number of children, the number of claims made in one year.
- b. **Continuous Variable:** The value of the variable can assume any value within a specific range. Example: The depth of the river, the height of the students, and time taken to finish a task. The possible values can be 12.3, 0.88, 177.7, 14.6, 5.57, etc.

Summary of the components of the data



7. Level of Measurement

Data can be collected in four different **levels of measurements**: Nominal level, ordinal level, interval level and ratio level.

- **Nominal Level:** Data can only be classified into categories and counted. It cannot be arranged in an ordering scheme. The ranking is not important in this level. Example: Counting the number of M&M candies by **colors**, counting the number of students entering the stadium by **gender**. The table below is another example of long distance telephone usage by carrier:

<u>Carrier</u>	<u>Number of Calls</u>	<u>Percent</u>
AT&T	108,115,800	75
MCI	20,577,310	14
Sprint	8,238,740	6
Other	7,130,620	5
Total	144,062,470	100

The table shows only the total number of calls made by different carriers. There is no natural order; which category placed first is not important.

- **Ordinal Level:** This involves data that may be arranged in some order, but differences between data values cannot be determined or are meaningless. Example: Rating the quality of a product as “good” or “poor”.

Rating of a finance professor

<u>Rating</u>	<u>Frequency</u>
Superior	6
Good	28
Average	25
Poor	12

In the rating of a finance professor above, we know that ‘Superior’ is better than ‘Good’ and ‘Good’ is better than ‘Poor’. However, we cannot tell the magnitude of how ‘Superior’ is better than ‘Good’ and how ‘Poor’ is worse than ‘Good’.

- **Interval Level:** Similar to the ordinal level, with the additional property that meaningful amount of differences between the data can be determined, meaning that we know the magnitude of how better between two categories is. There is no natural zero point, meaning that zero is not an absent of something or empty. Example, temperature. We can tell that 32°C is warmer than 24°C, and 0°C does not mean empty, it just telling us that the temperature is very cool at that time!
- **Ratio Level:** The ratio level includes all the characteristics of interval level but in addition, the 0 point is meaningful and the ratio or difference between two numbers is meaningful. Example is **Money**. If John earns \$0 meaning that John earns nothing (empty), if Peter earns \$1000 and Mary earns \$500, then Peter earns more than Mary does. The table given below shows another example of ratio level measurement.

Father-Son Salary Combinations

<u>Name</u>	<u>Income</u>	
	<u>Father</u>	<u>Son</u>
Jones	80,000.00	40,000.00
White	90,000.00	30,000.00
Rho	60,000.00	120,000.00
San	75,000.00	130,000.00

8. Collecting Raw Data

“Raw data” means unorganized or unprocessed data. It is very important for us to collect an appropriate set of data for the purpose of our study or research. A data set collected inappropriately may cause a **biased** result in your research. The sample results may not be representative of the population and may cause you to make a wrong decision or judgment.

a. Reasons for collecting data

- To provide the necessary input to a survey.
- To provide the necessary input to a study.
- To measure performance of an ongoing service or production process.
- To evaluate the standard of a service or process.
- Decision making process.
- To satisfy our curiosity.

b. Identify source of data

Data can be collected in many ways, namely:

- **Data which is already distributed by an organization or an individual**
Example: Newspaper, magazine, Internet, government agencies, etc.
A source is **primary** if the data collector is the one who uses the data for analysis.
A source is **secondary** if one organisation or individual has compiled the data and it is used by another organisation or individual.
- **Experiment**
In an experiment, strict control is exercised over the treatments.
Example: In a study of testing the effectiveness of new laundry detergent, the researcher determines that brand in the study is more effective in cleaning soiled clothes by actually washing dirty laundry instead of asking the customers which brand they believe is more effective.
- **Survey**
Here, no control is exercised over the behavior of the people being surveyed. This means that the respondents are free to answer the questionnaires based on their beliefs, attitudes, behaviour and their characteristics. Responses are then edited and compiled for analysis.
- **Observation**
Here, the behaviour of an object is observed directly in its **natural setting**. Most data on animal behaviour are collected in this way. Most of our data in the area of social sciences like psychology experiment are obtained by observation.

c. Design of survey research

One of the most popular sources of data is through conducting a survey. Survey design is an art that improves with experience. The general procedure for designing a survey involves five basic steps given below:

Step 1: Choose an appropriate mode of response:

The response from the survey is affected by the mode of response.

Example: E-mail, mail, internet, magazine, newspaper, personal interview or telephone.

Step 2: Identify broad categories

List the broad categories or options that reflect the theme of the survey. These categories should not overlap and must be completed.

Step 3: Formulate accurate questions

The questions must be short, simple and clear.

Example: Do you wish to further your master degree?
Do you wish to further your master degree in business? (Clearer)

Step 4: Conduct a pilot test

After completing your survey questionnaires, you must test your questionnaire on a small group of participants before implementing your survey to all the selected respondents. The pilot test can help you to detect the clarity and time taken to answer your survey.

Step 5: Write a cover letter

If the survey is conducted by mail, a cover letter must be included to explain the purpose and importance of the survey, and why they are the selected individuals for the survey. In some cases, offer an incentive in the form of a gift for the respondent's participation.

9. Types of survey: Sampling Method

A sample is the portion of the population that has been selected for analysis. Instead of taking a complete census of the whole population, statistical sampling focuses on collecting a small representative group of a larger population and making inferences on the population.

Why sampling rather than census?

- Save time.
- Save cost.
- Simple and more practical.

- It may be impossible for us to conduct the survey on the entire population.

10. Types of sampling:

The sample collected from sampling methods can be classified into probability sample and non-probability sample.

- **Probability sample: Used to make inferences on the population**

A sample is selected in such a way that each item or person in the population has a known likelihood of being included or selected in the sample. It can be conducted in four different ways as shown below:

- a) Simple random sample
- b) Systematic random sample
- c) Stratified random sample
- d) Cluster random sample

Throughout the course, the random sample we are using is a probability sample.

- **Non-probability sample: Cannot be used to make inferences on the population**

Not all items or people have a chance of being selected in the sample. Thus, there may be **bias**. It depends on personal judgment or **self-selected**. Thus, non-probability sampling method is not used throughout our statistics course.

Types of probability sample

- a) **Simple random sample**

This is most popular and is widely used in statistics. A sample is selected so that each person or item in the population has the same chance of being included. There are two methods by which samples are selected.

Sampling with replacement:

After a person or an item is selected, it is returned to the frame, where it has the same probability of being selected again. In Malaysia, the lottery result is drawn from a machine containing balls numbered from 0 to 9, and the ball is replaced each time it is drawn. Thus, the digits of any number selected can be the same.

Sampling without replacement:

Once a person or item is selected, it is not returned to the frame and therefore cannot be selected again. Usually when you conduct a survey on a population in one area, you do not administer the questionnaire on the same person again.

b) Systematic random sampling:

The items or individuals of the population are arranged in some way such as in an alphabetical order, in a file drawer by date received or by some other methods. A random starting point is selected, and then every, say, 10th member of the population is selected for the sample.

For example, we would like to interview 40 students from a population of 800 students in one local university. To use systematic random sampling method, first you may arrange the population of 800 students into 40 groups with 20 students for each group. You randomly select a student from the first group, and then every 20th student starting from the first student will be selected for your sample. See the figure below.

	<u>Group</u>							
	1	2	3	4	5	40
1	X_{11}	X_{21}	X_{31}	X_{41}	X_{51}	$X_{40\ 1}$
2	X_{12}	X_{22}	X_{32}	X_{42}	X_{52}	$X_{40\ 2}$
3	X_{13}	X_{23}	X_{33}	X_{43}	X_{53}	$X_{40\ 3}$
4	X_{14}	X_{24}	X_{34}	X_{44}	X_{54}	$X_{40\ 4}$
5	X_{15}	X_{25}	X_{35}	X_{45}	X_{55}	$X_{40\ 5}$
6	X_{16}	X_{26}	X_{36}	X_{46}	X_{56}	$X_{40\ 6}$
.....
.....
}								Selected samples
20	$X_{1\ 20}$	$X_{2\ 20}$	$X_{3\ 20}$	$X_{4\ 20}$	$X_{5\ 20}$	$X_{40\ 20}$
.....
.....
}								Selected samples
40	$X_{1\ 40}$	$X_{2\ 40}$	$X_{3\ 40}$	$X_{4\ 40}$	$X_{5\ 40}$	$X_{40\ 40}$

Advantage: Requires less time and sometimes results in lower cost.

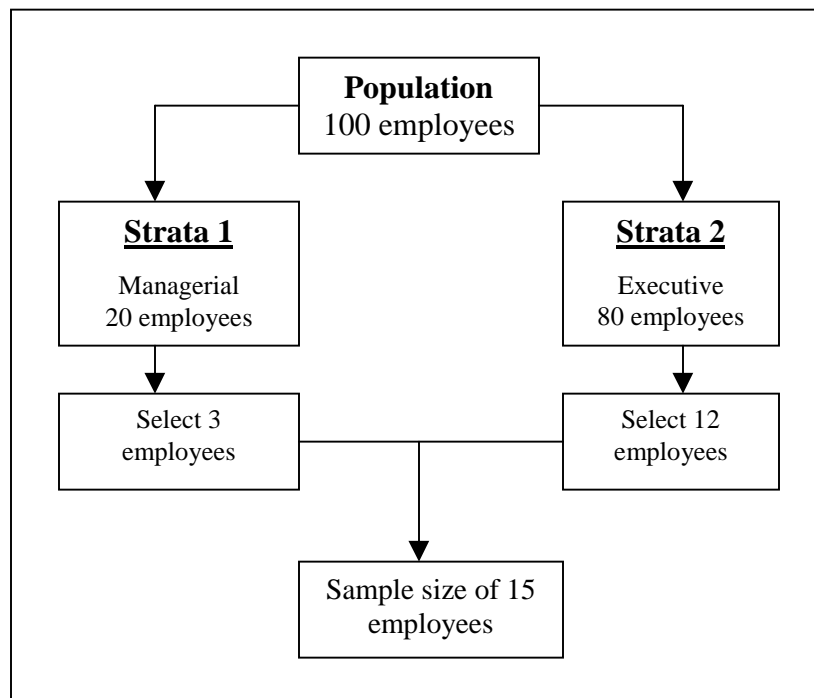
Disadvantages: Each element has an equal chance being selected but each sample does not have an equal chance of being selected.

c) **Stratified random sampling:**

A population is divided into subgroups by characteristics, called strata and a sample is selected from each stratum by proportion. For example, to randomly select 15 employees from one company with 20 managerial levels and 80 executive levels by using stratified random sampling method.

	Number	Percent
Managerial	20	20%
Executive	80	80%
Total	100	100%

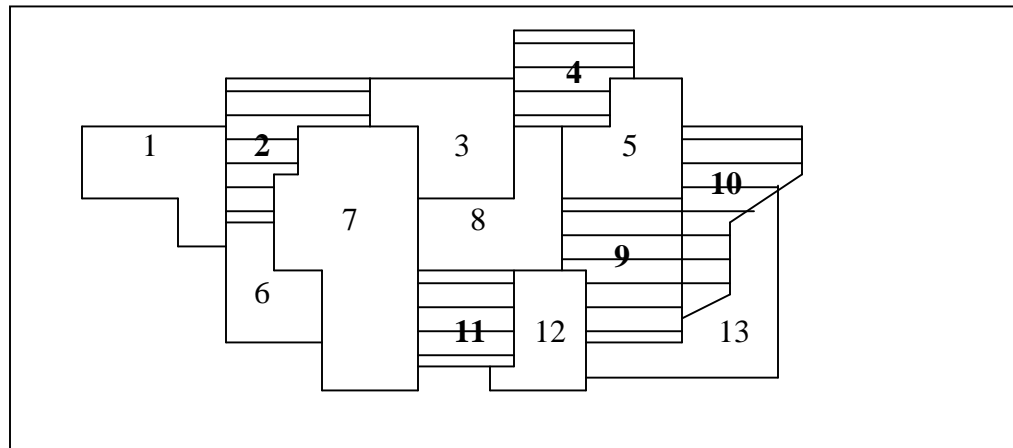
To select a sample of 15 employees by proportion, select 3 employees from managerial levels (15×0.2) and 12 employees from executive level (15×0.8).



Stratified sampling has the **advantage**, in some cases, of more accuracy reflecting the characteristics of the population than does simple random or systematic random sampling.

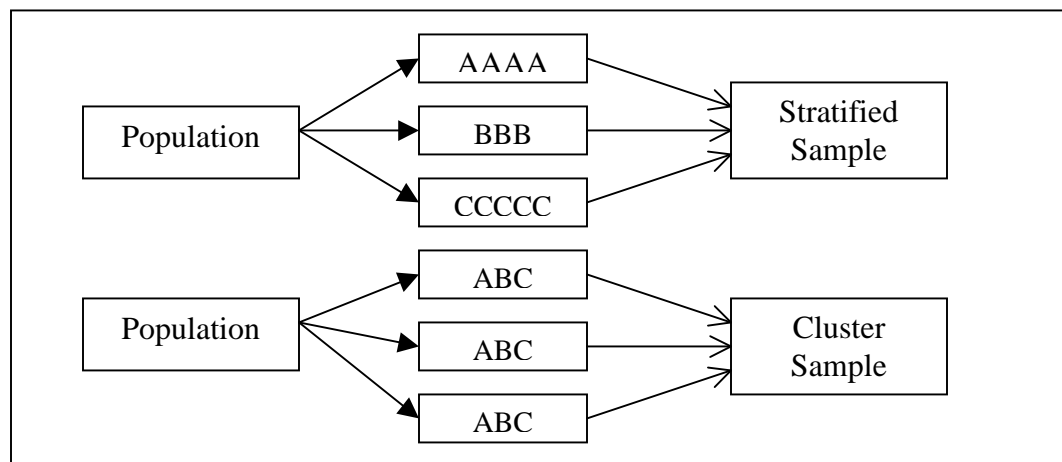
c) Cluster sampling:

Cluster sampling often employed to reduce the cost of sampling a population scattered over a large geographic area. In this method, divide the population into groups or clusters, afterward you select the clusters randomly then survey all the elements in selected clusters. For example, to study the views of federal environmental protection policies in Malaysia, first we divided Malaysia into 13 states (cluster), afterward we randomly selected says 5 states (cluster) in Malaysia and study the sample views from those selected states.



 **The different between stratified and cluster sampling.**

1. For both stratified and cluster sampling, the population is divided into subgroups.
2. Stratified sampling is used when each group is homogeneous, but groups vary from each other.
3. Cluster sampling is used when there is much variation within each group, but the groups are similar to each other.



6. Survey Errors:

Nothing is perfect in this world. Even when surveys use random probability sampling methods, they are subject to potential errors. A good survey is designed to attempt to reduce or minimize the errors. In general, there are four types of survey errors:

- a. Coverage error or selection bias
- b. Non-response error or non-response bias
- c. Sampling error
- d. Measurement error

a. Coverage Error

Coverage error occurs if certain groups of subjects are excluded from this frame listing so that they have no chance of being selected in the sample. Coverage error results in a selection bias. Non-probability sampling method may cause this error.

b. Non-response Error

Not everyone is willing to respond to a survey. Non-response error arises from the failure to collect data on all subjects in the sample and result in **non-response bias**.

c. Sampling Error

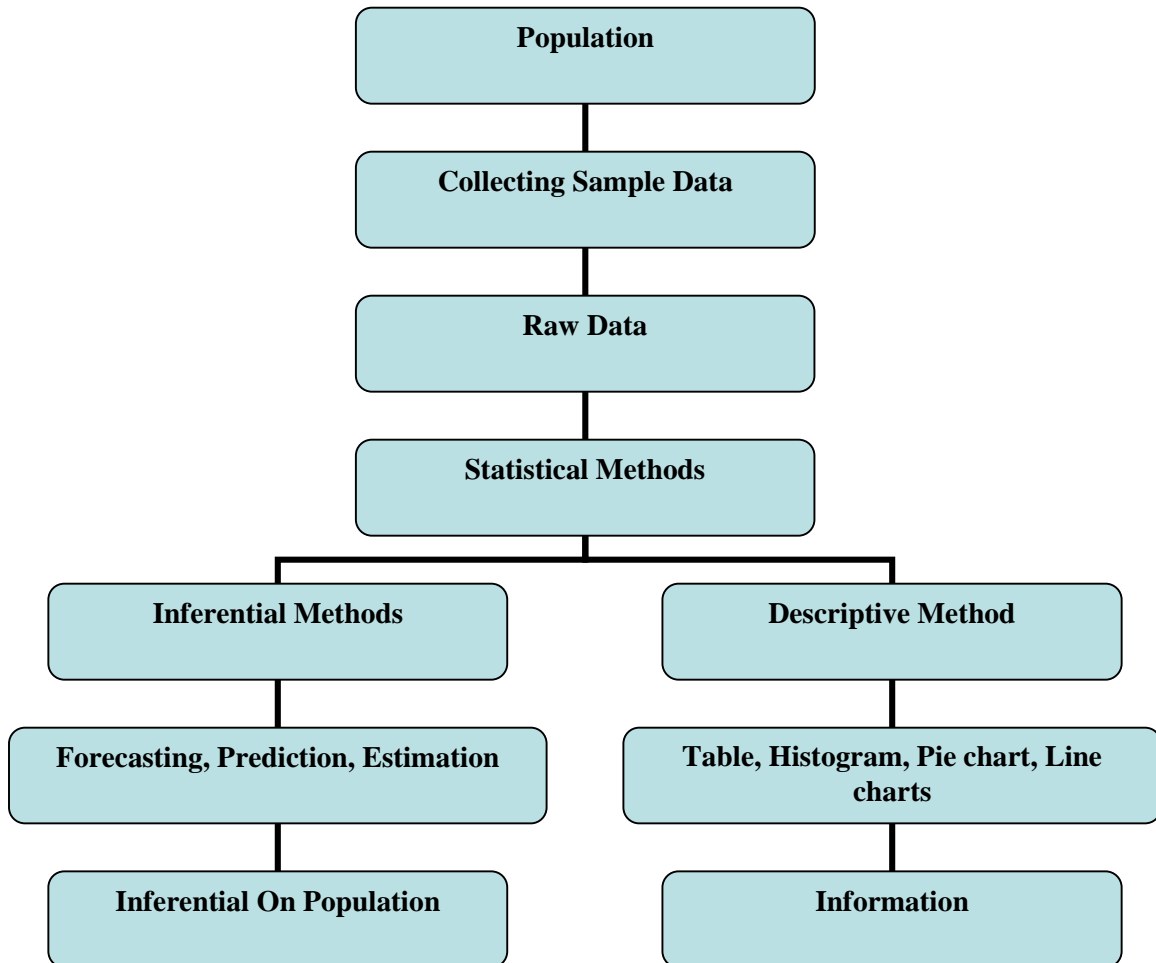
Sampling error reflects the heterogeneity of “chance differences” from sample to sample based on the probability of particular individuals or items being selected in the particular samples. When you read about the results of surveys or polls in newspapers or magazines, there is often a statement of error margin or precision. For example, “The result of this poll is expected to be within + or - 4 standard error of the actual value.” The sampling error can be reduced by increasing the sample size. Sampling error cannot be avoided but we can minimise it.

d. Measurement Error

Measurement error refers to inaccuracies in the recorded responses that occur because of a weakness in question wording, an interview effect on the respondent, or the effort made by the respondent. There are three sources of measurement error:

- **Ambiguous wording of questions**
- **Halo effect:** Occurs when the respondent feels obligated to please the interviewer. This error can be reduced by proper interview training.
- **Respondent error:** Occurs as a result of over-zealous or under-zealous effort by the respondent.

Data Generating Processing Flow



Learning Outcomes

- Students should be able to define what is statistics.
- Students should be able to explain what is descriptive statistics and inferential statistics
- Students should be able to distinguish between a qualitative variable and a quantitative variable.
- Students should be able to distinguish between a discrete variable and a continuous variable.
- Students should be able to distinguish among the nominal, ordinal, interval and ratio levels of measurement.
- Students should be able to use a sample to describe a population.

Basic Reading

1. Saravanan Kullandavelli (1994) LCCI Business Statistics; 5th ed. Malaysia; Stamford College Group Publishing.
2. A Francis (1995) Business Mathematics and Statistics; 4th ed. London DP Publications Ltd.

Revision Questions

1. Distinguish between the following terms:
 - a) Population and Sample
 - b) Random and Non-Random Sampling
 - c) Quantitative and Qualitative data
 - d) Chronological and Geographical data
 - e) Primary and Secondary data
 - f) Discrete and Inferential Statistics
2. For each of the following, decide whether the data is qualitative or quantitative; if quantitative, decide also whether it is discrete or continuous:
 - a) the temperature in a room
 - b) the number of boxes
 - c) the makes of microcomputer used on a company
 - d) the time taken to complete a process
 - e) the number of staff employed in the Human Resource Department