Jannatul Ferdous
Assistant Professor
Department of Computer Science & Engineering
Metropolitan University, Sylhet
Cell : +880 1990 07 55 26 (Whatsapp, Imo)
E-mail: jannat@metrouni.edu.bd

M.S. in Statistics
University of Chittagong, Chittagong, Bangladesh

B.Sc. (Hons.) in Statistics
University of Chittagong, Chittagong, Bangladesh

Content engagement

STA 215 (3.0 CREDITS):

Basic Statistics & Probability

CSE 50 & 51

Book Reference
Methods of Statics by K.C. Bhuyan.
Business Statistics by S.P. Gupta and M.P. Gupta.

STA 215 (3.0 CREDITS):

Basic Statistics & Probability

Chapter 1: Introduction to Statistics: Definition, Importance, Scopes & limitations. Population & sample, Parameter & statistic, Variables.

Chapter 2: Collection of data, Presentation of statistical data.

Chapter 3: Measures of Central Tendency: Arithmetic Mean, Geometric Mean, Harmonic mean, Median, Mode.

Chapter 4: Measures of Dispersion: Range, Standard deviation, Coefficient of Variation.

Chapter 5: Shape characteristics: Skewness & Kurtosis.

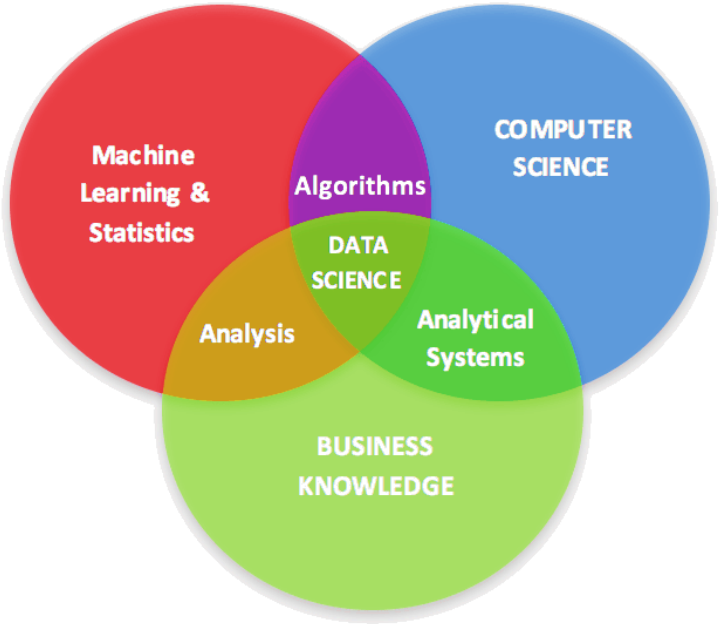
Chapter 6: Correlation.

Chapter 7: Regression.

Chapter 8: Probability: Introductory concept of Probability, Various mathematical problems related to probability.

Statistics ???

Scientific method for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.



Importance: Statistics is important for all branches of science, because it is incredibly helpful in decision making and testing the accuracy of the decisions at the end.

In computer science and machine learning, the efficiency of algorithms can be significantly increased by using statistics. Without understanding statistics, it is not possible to understand the logic behind algorithms and to develop them.

Human brain makes decisions based on data that it gets from its surroundings via sensors and it uses statistics and data to come a conclusion for an action. Your brain uses statistics even if you are not aware of this process at every instant of your life.

Functions of Statistics

1. Presents facts in simple form:

Statistics presents facts and figures in a definite form. That makes the statement logical and convincing than mere description. It condenses the whole mass of figures into a single figure. This makes the problem intelligible.

2. Reduces the complexity of data:

Statistics simplifies the complexity of data. The raw data are unintelligible. We make them simple and intelligible by using different statistical measures. Some such commonly used measures are graphs, averages, dispersions, skewness, kurtosis, correlation and regression etc. These measures help in interpretation and drawing inferences. Therefore, statistics enables to enlarge the horizon of one's knowledge.

3. Facilitates comparison:

Comparison between different sets of observation is an important function of statistics. Comparison is necessary to draw conclusions. The object of statistics is to enable comparison between past and present results to ascertain the reasons for changes, which have taken place and the effect of such changes in future. So to determine the efficiency of any measure comparison is necessary. Statistical devices like averages, ratios, coefficients etc. are used for the purpose of comparison.

4. Testing hypothesis:

Formulating and testing of hypothesis is an important function of statistics. This helps in developing new theories. So statistics examines the truth and helps in innovating new ideas.

Functions of Statistics .....continue

5. Formulation of Policies :

Statistics helps in formulating plans and policies in different fields. Statistical analysis of data forms the beginning of policy formulations. Hence, statistics is essential for planners, economists, scientists and administrators to prepare different plans and programmes.

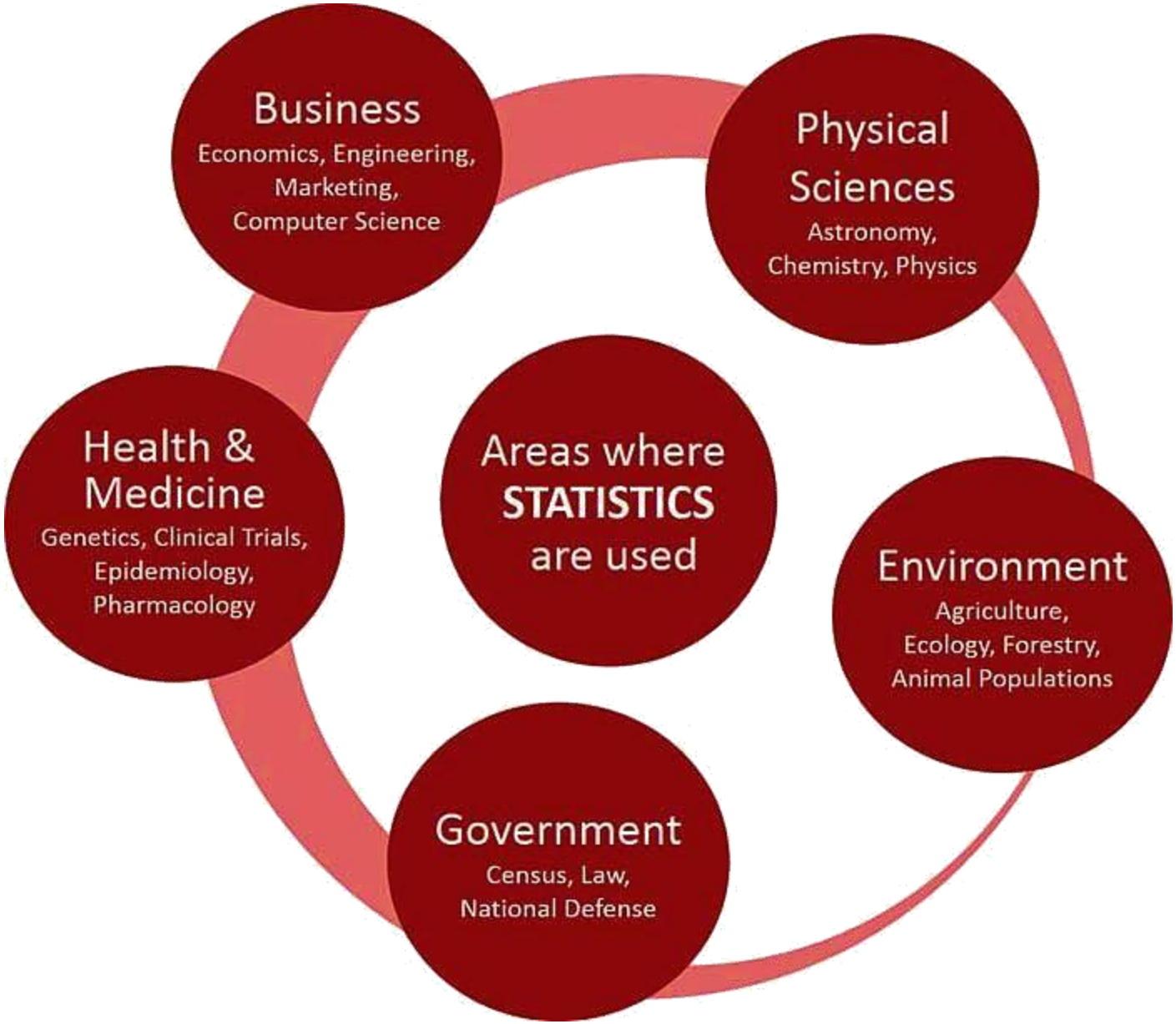
6. Forecasting :

The future is uncertain. Statistics helps in forecasting the trend and tendencies. Statistical techniques are used for predicting the future values of a variable. For example a producer forecasts his future production on the basis of the present demand conditions and his past experiences. Similarly, the planners can forecast the future population etc. considering the present population trends.

7. Derives valid inferences :

Statistical methods mainly aim at deriving inferences from an enquiry. Statistical techniques are often used by scholars planners and scientists to evaluate different projects. These techniques are also used to draw inferences regarding population parameters on the basis of sample information.

Uses of Statistics



	MACHINE LEARNERS	STATISTICIANS
Network/Graphs vs. Models	Network/Graphs to train and test data	Models to create predictive power
Weights vs. Parameters	Weights used to maximize accuracy scoring and hand tuning	Parameters used to interpret real-world phenomena - stress on magnitude
Confidence Interval	There is no notion of uncertainty	Capturing the variability and uncertainty of parameters
Assumptions	No prior assumption (we learn from the data)	Explicit a-priori assumptions
Distribution	Unknown a priori	A-priori well-defined distribution
Fit	Best fit to learning models (generalization)	Fit to the distribution

Limitations of Statistics

- Statistics is not suited to the study of qualitative phenomena. Even qualitative information is converted into numerical data by the method of ranking, scoring or scaling
- It is not a study of individual, it is a study of group
- Statistics laws are not exact (unlike physical and natural sciences statistical laws are only approximations not exact)
- Homogeneity of data is an essential requirement
- It results truly based on averages
- It can be misused (the use of statistical tools by inexperienced and untrained persons might lead to very fallacious conclusion):-
    i. If the people volunteer are paid some amount of money they can skew the results in favor of experiment
    ii. How were the questions worked to the participants in the study?(how the questions is framed will change the result, hence one should be very careful how clearly the experiments has defined the problem)
    iii. Who paid for study (e.g. few years ago a study reported balding men were more likely to have heart attack than non-balding men this was financed by a drug company that manufacturer a popular hair-growing product)

Basic concept of statistics

Population & Sample

A **population** is a group of phenomena that have something in common. The term often refers to a group of people, as in the following examples: All registered voters in Bangladesh

Often, researchers want to know things about populations but do not have data for every person or thing in the population. If a company's customer service division wanted to learn whether its customers were satisfied, it would not be practical (or perhaps even possible) to contact every individual who purchased a product. Instead, the company might select a **sample** of the population.

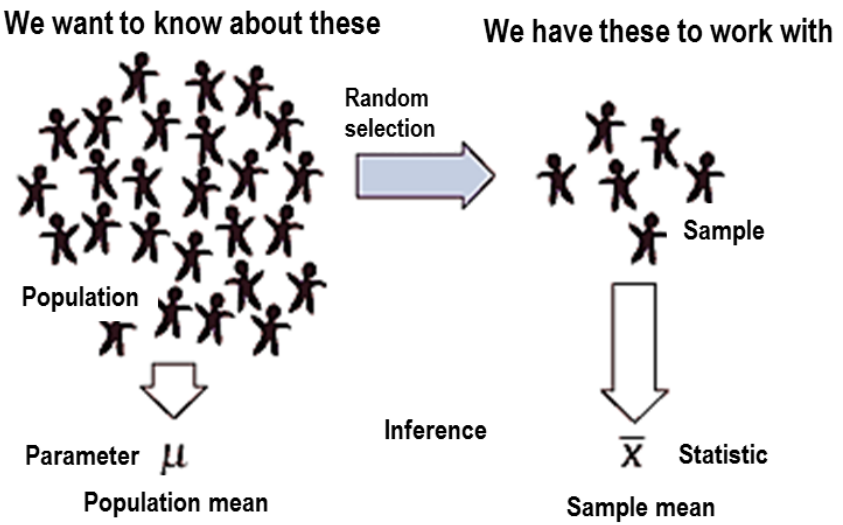
A **sample** is a smaller group of members of a population selected to represent the population. In order to use statistics to learn things about the population, the sample must be **random**. A random sample is one in which every member of a population has an equal chance of being selected.

- A **population** includes all of the elements from a set of data
- A **sample** consists one or more observations drawn from the population

Real World Example of Population

For example, let's say a denim apparel manufacturer wants to check the quality of the stitching on its blue jeans before shipping them off to retail stores. It is not cost effective to examine every single pair of blue jeans the manufacturer produces (the population). Instead, the manufacturer looks at just 50 pairs (a sample) to draw a conclusion about whether the entire population is likely to have been stitched correctly.

A **parameter** is a characteristic of a population. A **statistic** is a characteristic of a sample. Inferential statistics enables you to make an educated guess about a population parameter based on a statistic computed from a sample randomly drawn from that population.



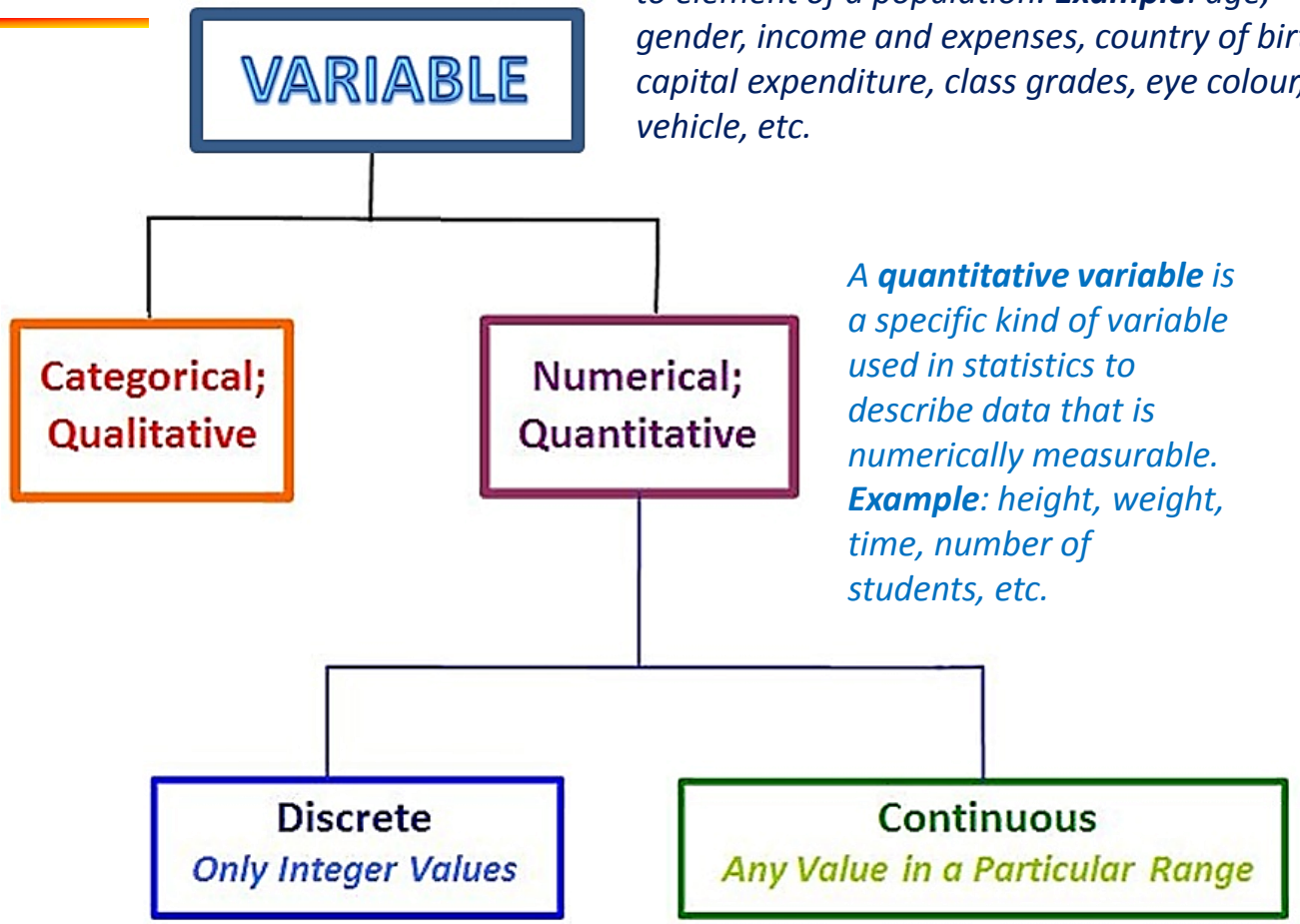
Population	Sample
The measurable characteristic of the population like the mean or standard deviation is known as the parameter.	The measurable characteristic of the sample is called a statistic.
Population data is a whole and complete set.	The sample is a subset of the population that is derived using sampling.
A survey done of an entire population is accurate and more precise with no margin of error except human inaccuracy in responses. However, this may not be possible always.	A survey done using a sample of the population bears accurate results, only after further factoring the margin of error and confidence interval.
The parameter of the population is a numerical or measurable element that defines the system of the set.	The statistic is the descriptive component of the sample found by using sample mean or sample proportion.

Parameter		Statistic
A value, usually a numeric value that describes a population		A value, usually a numeric value that describes a sample
Derived from measurements of the individuals in the population		Derived from measurements of the individuals in the sample
For the same population the value of the parameter is same		For the same population the value of the statistic varies from sample to sample
Example: Population mean (μ), Standard deviation (σ), Correlation co-efficient (ρ), Regression co-efficient (β), etc.		Example: Sample mean ($\overline{X}$), Standard deviation (S), Correlation co-efficient (r), Regression co-efficient (b), etc.

Basic concept of statistics

Variables

*A **variable** is a measurable characteristics, number, or quantity that can vary from element to element of a population. **Example**: age, gender, income and expenses, country of birth, capital expenditure, class grades, eye colour, vehicle, etc.*



*A **qualitative variable** is a specific kind of variable which* can't be counted (i.e. has no numerical value). It is measured by categories or groups. **Example**: *gender, class grades, eye colour, occupation, etc.*

*A **quantitative variable** is a specific kind of variable used in statistics to describe data that is numerically measurable. **Example**: height, weight, time, number of students, etc.*

A **discrete variable** *is a specific kind of quantitative variable in which data always take integer values. **Example**: number of students, number of eggs, etc.*

*A **continuous variable** is a specific kind of quantitative variable that describe data within a certain limit. **Example**: height, weight, time, distance, age, etc.*