IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

Important of Statistics

Prof. Ashwini Naresh Kudtarkar and Prof. Tarannum Ansari
Shri G. P. M. Degree College, Vile Parle (E), Mumbai, Maharashtra, India

Abstract: *This paper is focused on the issues related to optimizing statistical approaches in the emerging fields of Computer Science and Information Technology. More emphasis has been given on the role of statistical techniques in modern data mining. Statistics is the science of learning from data and of measuring, controlling, and communicating uncertainty. Statistical approaches can play a vital role for providing significance contribution in the field of software engineering, neural network, data mining, bioinformatics and other allied fields. Statistical techniques not only helps make scientific models but it quantifies the reliability, reproducibility and general uncertainty associated with these models. In the current scenario, large amount of data is automatically recorded with computers and managed with the data base management systems (DBMS) for storage and fast retrieval purpose. The practice of examining large preexisting databases in order to generate new information is known as data mining. Presently, data mining has attracted substantial attention in the research and commercial arena which involves applications of a variety of statistical techniques. Twentyyears ago mostly data was collected manually and the data set was in simple form but in present time, there have been considerable changes in the nature of data. Statistical techniques and computer applications can be utilized to obtain maximum information with the fewest possible measurements to reduce the cost of data collection..*

Keywords: Statistics, Data Mining, SoftwareEngineering, DBMS, Neural Networks, etc

I. INTRODUCTION

Statistics is the part of mathematical science which pertains to the collection, classification, tabulation, analysis, interpretation and presentation of data. Some of the eminent researchers consider Statistics to be a separate mathematical science rather than a branch of mathematics. On the other hand, many scientific investigations make use of data, Statistics is concernedwith the use of data in the background of uncertainty and judgment making inthe face of uncertainty.

In personal life, we have used statistics for general calculation of householdbudget. Generally, there are two type of information i.e., quantitative and qualitative information. Thus, this subject is used by the people to take appropriate decision about the problems/ budget on the basis of the both types of information's.

DEFINITION OF STATISTICS

The definition of statistics has been given by the different statistician indifferent ways. Some important definitions of statistics are given below;

L. Bowley defined that "Statistics may be called the science of counting".He also said that "Statistics may rightly be called the science of average".

According to **Boddington** "Statistics is the science of estimates andprobabilities".

According to **Selligman** "Statistics is the science which deals with the methods of collecting, classifying, tabulation, comparing and interpretingnumerical data collected to throw some light on any sphere of enquiry".

Croxton and Cowden defined that "statistics as the collection, tabulation,presentation, analysis and interpretation of numerical data".

TYPES OF STATISTICS

There are the two broad ways of classifying statistics, one is the on thebasis of function and another is the on the basis of distribution of data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

Statistics Based on Function:

There are three types of statistics on the basis of subject matter/ function. The types of statistics based on function have been given infollowing chart;

Types of Statistics Based on Function		
Descriptive Statistics Descriptive Statistics is the branch, which deals with descriptions of obtained data. It is a summary statistic which summarizes features/ characteristics from a collection of information. Moreover, it include classification, tabulation, and measurement of central tendency as well as variability. The researcher's use of these measures to understand about the tendency of data/ scores. Which further enhance the ease in description of thephenomena	**Correlational statistics** In correlational statistics, the obtained data are disclosed for their inter correlations and It includes various types of methods to compute the relationship (correlations) among data. It also provide description about sample or population for their further analyses purpose to explore the significance of sampling and population averages.	**Inferential statistics** Statistical inference (SI) is the process of data analysis to deduce properties of probability distribution. Inferential statistical analysis infers properties of a population or census through the testing hypotheses and deriving estimates which is based on the primary assumption i.e., the observed data set is sampled from a larger population. It is also deals with the drawing of conclusions about population/ census. Moreover, It provide technique to compute the probabilities of future behaviour of the subjects/areas

CHARACTERISTICS OF STATISTICS/ STATISTICAL DATA

The characteristics of statistics can be divided in two groups on thebasis of its meaning;

Characteristics of Statistical Data

Statistics is the aggregates of facts
Statistics/ data can be represented in numbers.
Statistics/ data are affected in sufficient quantity for avariety of reasons.
Calculation of data or estimation accuracy is basedon some level of significance.
The compilation of data is on pre-determined objects.
The data/ figure must be individually independent.
The data are presented in a mutually related form.

Characteristics of Science of Statistics

Statistics is a group of methods or techniques.
Use of statistical science is almost universal.
Statistics deals with the aggregate of numerical facts.
Statistics is both a science and an art.

SCOPE AND DIVISION OF STATISTICS

It can be divided into two parts;

Statistical Methods: Johnson and Jackson said that "Statistics is the process of the methods which are related to collection, classification, tabulation, analysis, interpretation and

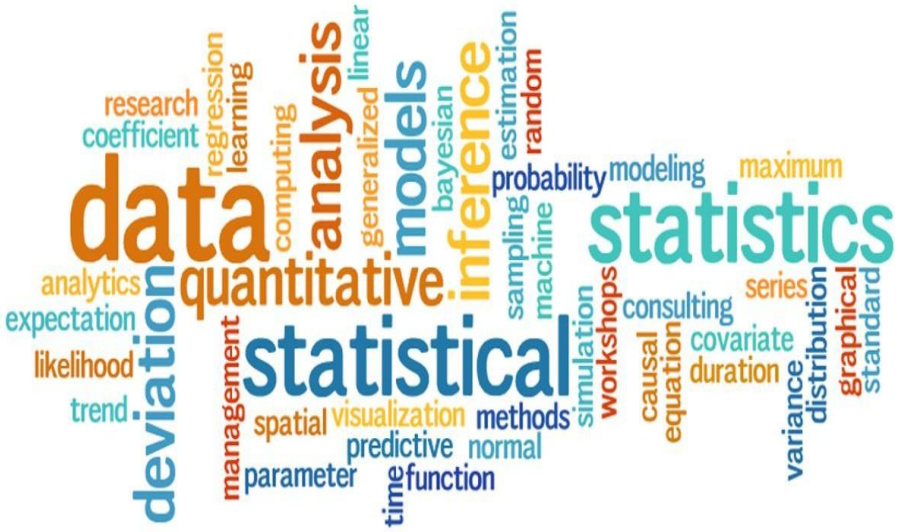
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

presentation of data."
Collection of data.
Organisation of data.
Presentation of data.
Analysis and interpretation of data.
Forecasting of data.

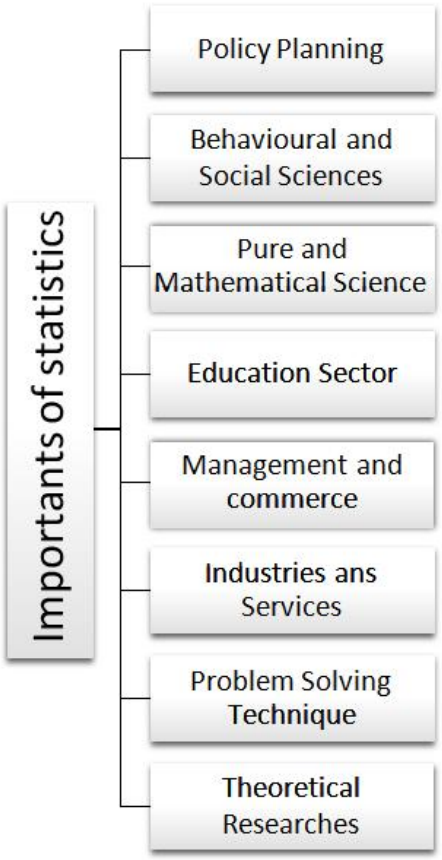
Applied Statistics
Descriptive Statistics: In this, the data compiled in thepast or present of region is studied.
Scientific Applied Statistics: Its help to construct scientific laws of behavioural science and its confirmation

IMPORTANCE OF STATISTICS

When analysts use statistical procedures correctly, they tend to produce accurate results. In fact, statistical analyses account for uncertainty and error in the results. Statisticians ensure that all aspects of a study follow the appropriate methods to produce trustworthy results. These methods include:
Producing reliable data.
Analysing the data appropriately.
Drawing reasonable conclusions
There is the wide importance of Statistics in several areas/ subject. Statistical applications have a wide scope and uses. Some of the majorimportance's are given below:
It forms the bases of scientific approach to problem solving. The students cannot carry out their researches without knowing the scientific method of approach to problem solving in their disciplines.
It helps in understanding events or characteristics of human behaviouroccurred in past, in present on variety of tasks in human life. It also helps us in predicting the future performance of the students in a course or success in a job.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023



Policy Planning: To finalise a government or individual policy, it requires some relevant data from previous documents or expected environment thatthe policy can be effectively utilised with maximum favourable benefits/ results.

Behavioural and Social Sciences: In social sciences especially in Economics, the both types of information i.e., quantitative and qualitative are used to analysis and draw policy recommendations. Moreover, statistics helps the academician/ researchers to alter the information in a comprehensive way to analysis and predict the patterns of behaviour or trends.

Pure and Mathematical science: The tools of statistical are also used to have precise measures in pure and mathematical sciences and to see differences on different occasions in various conditions.

Education Sector: The statistical tools or instruments is also used in the area of education. Statistics used to create patterns and trends of variables on the basis of past and present conditions and hence showing the direction of development in education sector. Further, these trends helps to crates the policies and planning of the education.

Management and Commerce: Statistics is very useful tool in management and commerce. It organisation the various aspects of work and well-being ofthe employees. It is also a very useful instrument for account, which is the branch of commerce. Moreover, it also keep an eye on the progress trend ofthe organisation.

Industries and Service: Statistics is a basic tool to analysis the progress ofindustry as well as service sector and it also helps to make further strategiesfor the development of these sectors.

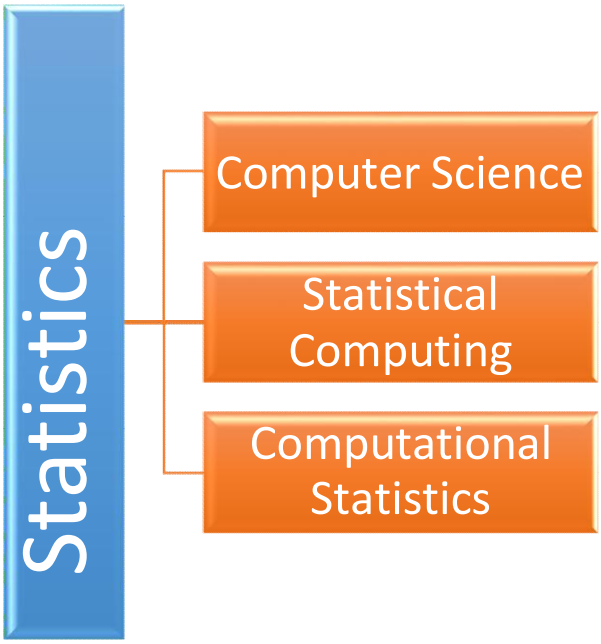
Problem Solving Technique: Statistics is provide the problem solving tool between two or more variables. To find out the best applicable solution to a problem situation, we can use statistical technique and it is possible because of statistics.

Theoretical Researches: On the basis of statistical analysis, we can establish the significance relationship of those facts for a particular paradigmor phenomena, which theories evolve the facts obtained from the field.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

It forces us to be definite and exact in our thinking and procedure to be adopted inmaking the concepts clear.
However, even if one will never be a researcher himself/herself, he would like to have his acquaintance with certain topics, because his profession desires him to apply research literature. He will be required to comprehend and evaluate detailedresearch literature, much of which is couched in statistical terms.

Statistical Computing: The term "statistical computing" to refer to the computational methods that enable statistical methods. Statistical computing includes numerical analysis, database methodology, computer graphics, software engineering and the computer-human interface.
Computational Statistics: The term "computational statistics" somewhat more broadly to include not only the methods of statistical computing but also modern statistical methods that are computationally intensive. Thus, to some extent, "computational statistics" refers to a large class of modern statistical methods. Computational statistics is grounded in mathematical statistics, statistical computingand applied statistics. Computational statistics is related to the advance of statistical theory and methods through the use of computational methods. Computation in statistics is based on algorithms which originate in numerical mathematics or in computer science. The group of algorithms highly relevant for computational statistics from computer science is machine learning, artificial intelligence (AI), and knowledge discovery in data bases or data mining. These developments have given rise to a new research area on the borderline between statistics and computer science.
Computer Science vs. Statistics: Statistics and Computer Science are both about data. Massive amounts of data is present around today's World. Statistics lets us summarize and understand it with the use of Computer Science. Statistics also lets data do our work for us.



STATISTICAL APPROACHES IN COMPUTATIONAL SCIENCES
Statistics is essential to the field of computer science in ensuring effectiveness, efficiency, reliability, and high-quality products for the public. Statistical thinking not only helps make scientific discoveries, but it quantifies the reliability, reproducibility and general uncertainty associated with these discoveries. The following terms are abrief listing of areas in computer science that use statistics to varying degrees at various times.
Data Mining: Data mining is the analysis of information in a database, usingtools that look for trends or irregularities in large data sets. In other words "finding useful information from the available data sets using statistical techniques"

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

• Speech Recognition: Speech recognition is the identification of spoken words by a machine. The spoken words are turned into a sequence of numbers and matched against coded dictionaries.
• Vision and Image Analyses: Vision and image analyses use statistics to solve contemporary and practical problems in computer vision, image processing, and artificial intelligence.
• Human/Computer Interaction: Human/Computer interaction uses statistics to design, implement, and evaluate new technologies that are useable, useful, and appealing to a broad cross-section of people.
• Network/Traffic Modelling: Network/Traffic modelling uses statistics to avoid network congestion while fully exploiting the available bandwidth.
• Artificial Intelligence: Artificial intelligence is concerned with modelling aspects of human thought on computers
• Machine Learning: Machine learning is the ability of a machine or system to improve its performance based on previous results.
• Capacity Planning: Capacity planning determines what equipment and software will be sufficient while providing the most power for the least cost.
• Storage and Retrieval: Storage and retrieval techniques rely on statistics to ensure computerized data is kept and recovered efficiently and reliably.
• Quality Management: Quality management uses statistics to analyse the condition of manufactured parts (hardware, software, etc.) using tools and sampling to ensure a minimum level of defects.
• Software Engineering: Software engineering is a systematic approach to the analysis, design, implementation, and maintenance of computer programs.

STATISTICS IN SOFTWARE ENGINEERING

Software engineering aims to develop methodologies and procedures to control the whole software development process. Nowadays researchers attempt to bridge the islands of knowledge and experience between statistics and software engineering by enunciating a new interdisciplinary field: statistical software engineering. Design of Experiments (DOE) uses statistical techniques to test and construct models of engineering components and systems. Quality control and process control use statistics as a tool to manage conformance to specifications of manufacturing processes and their products. Time and methods engineering uses statistics to study repetitive operations in manufacturing in order to set standards and find optimum (in some sense) manufacturing procedures. Reliability engineering uses statistics to measures the ability of a system to perform for its intended function (and time) and has tools for improving performance. Probabilistic design uses statistics in the use of probability in product and system design. Essential to statistical software engineering, is the role of data: wherever data are used or can be generated in the software life cycle, statistical methods can be brought to bear for description, estimation, and prediction. The department of software engineering and statistics trains multi skilled engineers in the processing of information, both in its statistical and computational forms, for use in various business professions.

STATISTICS IN HARDWARE MANUFACTURING

The hardware manufacturing companies are applying statistical approaches to create a plan of action that will work more efficiently for forecasting the future productivity of the hardware enterprise. Adopted statistical approaches for:
• Forecasting production, when there is a stable demand and uncertain demand.
• Pinpoint when and which inputs of a specific model will be the cause of uncertainty
• Calculate summary statistics in order to set sample data.
• To make market analysis and process optimizations.
• Statistical tracking and predicting for quality improvement.

STATISTICS IN DATABASE MANAGEMENT

Databases are packages designed to create, edit, manipulate and analyse data. To be suitable for a database, the data must consist of records which provide information on individual cases, people, places, features, etc. Optimizer statistics are a collection of data that describe more details about the database and the objects in the database. The optimizer statistics are stored in the data dictionary. They can be viewed using data dictionary views. Because the objects in a

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

database can be constantly changing; statistics must be regularly updated so that they accurately describe these database objects. These statistics are used by the query optimizer to choose the best execution plan for each SQL statement. Optimizer statistics include the following:

• Table Statistics
o Number of rows
o Number of blocks
o Average row length
• Column Statistics
o Number of distinct values (NDV) in column
o Number of nulls in column
o Data distribution (histogram)
• Index Statistics
o Number of leaf blocks
o Levels
o Clustering factor
• System Statistics
o I/O performance and utilization
o CPU performance and utilization

Statistical packages for databases are SAS, SPSS, R, etc. and these are available over a wide range of operating systems. Numerous other packages have been developed specifically for the PC DOS environment. S is a commonly available statistical package for UNIX.

STATISTICS IN DATA MINING

Data Mining is a process of discovering previously unknown and potentially useful hidden pattern in the data. Advances in information technology have resulted in a much more data based society. Data touch almost every aspect of our lives like commerce on the web, measuring our fitness and safety, doctors treat our illnesses, economic decisions that affect entire nations, etc. Alone, data are not useful for knowledge discovery. Data mining are transitioning from data-poor to data-rich by using the methods like data exploration, statistical inference and understanding of variability and uncertainty.

Statistical Elements Present in Data Mining

• Contrived serendipity, creating the conditions for fortuitous discovery.
• Exploratory data analysis with large data sets, in which the data are as far as possible allowed to speak for themselves, independently of subject area assumptions and of models which might explain their pattern. There is a particular focus on the search for unusual or interesting features.
• Specialised problems: fraud detection.
• The search for specific known patterns.
• Standard statistical analysis problems with large data sets.

Data Mining from Statistical Perspective

• Data sets which are relatively large and homogeneous might be reasonable to us mainstream statistical techniques on the whole or a very large subset of the data.
• All analyses done by mainstream statistics have intended outcome like set of data to a small amount of readily assimilated information.
• The outcome may include graphs, or summary statistics, or equations that can be used for prediction or a decision tree.
• Large volume of data without loss of information be reduced to a much smaller summary form, this can enormously aid the subsequent analysis task.
• It becomes much easier to make graphical and other checks that give the analyst assurance that predictive models or other analysis outcomes are meaningful and valid.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

Feature	Statistics	Data Mining
Type of Problem	Well structured	Unstructured / Semi-structured
Inference Role	Explicit inference plays great role in any analysis	No explicit inference
Objective of the Analysis and Data Collection	First – objective formulation, and then - data collection	Data rarely collected for objective of the analysis/modelling
Size of data set	Data set is small and hopefully homogeneous	Data set is large and data set is heterogeneous
Paradigm/Approach	Theory-based (deductive)	Synergy of theory based and heuristic-based approaches (inductive)
Type of Analysis	Confirmative	Explorative
Number of variables	Small	Large
Methods/Techniques	Dependence Methods: Discriminant analysis, Logistic regression Interdependence Methods: Correlation analysis, Correspondence analysis, Cluster analysis	Predictive Data Mining: Classification, Regression Discovery Data Mining: Association Analysis, Sequence Analysis, Clustering

PROPERTIES OF STATISTICAL PACKAGES

Statistical packages offer a range of types of statistical analysis. Statistical packages includes: Database functions, such as editing, printing reports.

- Capabilities for graphic output, particularly graphs but many also produce maps.
- Common packages are SAS, SPSS, R, etc.
- Available over a wide range of operating systems.
- Some have been "ported" to (rewritten for) the IBM PC.
- Numerous other packages have been developed specifically for the PC DOS environment.
- S is a commonly available statistical package for UNIX.

II. CONCLUSION

In this paper, many areas of computer science have been described in which statistics plays a very vital role for data and information management. Statistical thinking fuels the cross-fertilization of ideas between scientific fields (biological, physical, and social sciences), industry, and government. The statistical and algorithmic issues are both important in the context of data mining. Statistics is an essential and valuable component for any data mining exercise. The future success of data mining will depend critically on our ability to integrate techniques for modelling and inference from statistics into the mainstream of data mining practice.

REFERENCES

[1] Dr. Sanjeev Kumar, Introduction to statistics, Economics Department
[2] Government Women College Gandhinagar, Importance of statistics
[3] Murtada Iihasme, Importance of statistics, Wasit University
[4] Neeraj Tiwari, Professor & Head Department of Statistics, Kumaun University SSJ Campus, Almora Uttarakhand, India

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, January 2023

[5] Elder, J. F. and Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. Advances in Knowledge Discovery and Data Mining, MIT Press, pp.83-115.
[6] Gentle, J.E. (2004). Courses in statistical computing and computational statistics. The American Statistician, Vol. 58, pp.2–5.
[7] Grier, D.A. (1991). Statistics and the introduction of digital computers. Chance, Vol. 4(3), pp.30–36.
[8] Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. Statistics and Computing, Vol. 9, pp.123-143.
[9] Jay L. Devore: Probability and Statistics for Engineering and the Sciences, Engage learning.
[10] Murray R. Speigel: Theory & Problems of Statistics, Schaum's publishing Series.