Data analysis using Microsoft Excel

Introduction to Statistics

Statistics may be defined as the science of collection, organization presentation analysis and interpretation of numerical data from the logical analysis.

1.Collection of Data

It is the first step, and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of data such as census, sampling, primary, secondary, etc., and the investigator should make use of correct method.

2.Organization of data

The mass data collected cannot conclude anything. So, the collected data should be condensed into suitable form and format. The arrangement and categorization will be done in this process.

3.Presentation of data

The collected data should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular or diagrammatic or graphic form.

4.Analysis of data

The data presented should be carefully analyzed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, regression etc.,

5.Interpretation of data

The final step is drawing conclusion from the data collected. A valid conclusion must be drawn based on analysis. A high degree of skill and experience is necessary for the interpretation.

Data type

Measurement consists of counting the number of units or parts of units displayed by objects and phenomena. Measurement is a process of assigning numbers or symbols to any facts or objects or products or items according to some rule. It is a tool by which individuals are distinguished on the variables of area under study.

There are three types of data measurement

Nominal data

It is the simplest type of data, also known as categorical data. It is lowest level of measurement. It is simply a system of assigning number or the symbols to objects or events to distinguish one from another or in order to label them. The symbols or the numbers have no numerical meaning. The arithmetic operations cannot be used for these numerals. The orders of the symbol have no mathematical meaning.

For example, gender, occupation, religion are measured in nominal data. If we use data 1 for male programmer and 2 for female programmer for measuring the gender of

programmer, then 1 and 2 have no numeric meaning. It is used to distinguish male and female. In this case the number of a set of objects is not comparable to the other set. Mathematical operations such as addition, subtractions, multiplication, and division cannot be performed for the variable having nominal data. Frequency count is possible in this case so that percentage, mode can be obtained, and chi square test can be performed for test of significance for variable having nominal data.

Ordinal data

The second and the lowest level of ordered data is the ordinal data. It is the quantification of items by *ranking.* In this data, the numerals are arranged in some order but the gaps between the positions of the numerals are not made equal. It is used to rate preference of respondents. It indicates relative extent to which object possess certain characteristic. It represents qualitative values in ascending or descending order. The rank orders represent ordinal data and mostly useful in scaling the qualitative phenomena.

For example, qualification levels, preference to different statistical software, preference to different made of Laptops are measured in ordinal data. If we are going to study the computer literacy and we took the qualification level of people as primary, lower secondary, secondary, higher secondary, bachelor, master and Ph. D then we can use 1, 2, 3, 4, 5, 6 , 7 to represent primary , lower secondary , secondary , higher secondary , bachelor , master and Ph.D. respectively in ascending order. Mathematical operations such as addition, subtraction, multiplication, division cannot be performed for variable having ordinal data. Frequency count is possible and due to ranking partition values can be determined for variable having ordinal data. In this case median, mode, rank correlation also can be obtained.

Scale data

In addition to ordering the data, this data uses equidistant units to measure the difference between scores It assumes data have equal intervals. It is most powerful data of measurement. It possesses the characteristics of nominal, ordinal data. It can be expressed as relative level of measurement.

For example, Numbers on the data indicates the actual amount of property being measured. The ratio involved in the scale data possesses measure property and it facilitates comparison. Mathematical operations like addition, subtraction, multiplication and division all can be performed.

It represents actual number of variables and it is used to measure the physical dimensions. Some examples of scale data variables are disk space, amount expenses in IT department in different year, age etc. Arithmetic mean, Geometric mean, Harmonic mean, coefficient of variation along with all other measures can be obtained for variable having scale data. All statistical techniques can be applied to the variable present in scale data.

The following is the properties of categories, rank, equal interval of three data.

Level of measurement	Property		
	Categories	Ranks	Equal intervals
Nominal	Yes	No	No
Ordinal	Yes	Yes	No
Scale	Yes	Yes	Yes

Data analysis using Microsoft Excel

Variable

Variable is characteristic which can take more than one value. It may be characteristics of persons, things, events, groups, objects, feelings or any other category to be measured. It can take many values. Examples of variable are age of computer, sex of programmer, income of computer operator, job satisfaction, attitude of employee etc. If anyone is interested to find average age of CSIT first year second semester students, then the collection of information on variable age should be done. In this case value of variable age is expressed in years. To gather information on age one has to ask only one question "how old you are?" to the students of CSIT second semester students.

Types of Data

A Data provides the information about the individual. When the data taken by a variable is numeric then it is called numerical data. It is also called quantitative data. It can be divided into discrete and continuous data.

Discrete data is the data which can take countable values or whole numbers. For example, discrete variable, number of students in classes, number of rooms in houses, Number of vehicles in offices etc., then it takes only the whole numbers. So, the data taken by these variables are discrete and the respective variable is said to be discrete variable.

Continuous variable is the variable which can take all possible values i.e. whole numbers and fractions (real numbers). Examples of continuous variable are temperature recorded, height of students, wages of programmer, sales of Laptop etc. Here the data taken by these variables is within certain range. For example, wage of programmer is measured in Rupees as 5000 – 10000, 10000 – 20000, 20000 – 30000, etc then the data here collected are continuous in nature.

When the data taken by a variable is non-numeric then the variable is called qualitative variable. It is also called categorical variable. Examples of qualitative variable are religion of students, gender of sales man, education level of business executives etc.

Statistics and Technology

Statistics respond appropriately to the new demands of research and development in various areas of experimental sciences. With the advances in computer technology, many innovative techniques for use in applied statistics have recently been developed. In medicine, agriculture, business and government, these applications have solved important problems. Performing calculations almost at the speed of light, the computer has become one of the most useful research tools in modern times. Computers are ideally suited for data analysis concerning large research projects. Researchers are essentially concerned with huge storage of data, their faster retrieval when required and processing of data with the aid of various techniques. In all these operations, computers are of great help. Their use, apart expediting the research work, has reduced human drudgery and added to the quality of research activity.

Data analysis using Microsoft Excel

The computers can perform many statistical calculations easily and quickly. Computation of means, standard deviations, correlation coefficients, 't' tests, analysis of variance, analysis of covariance, multiple regression, and various nonparametric analyses are just a few of the programs and subprograms that can be solve very easily using computer. Similarly using computer, linear programming, multivariate analysis, Monte Carlo simulation etc. are also can be done very easily. Software packages are readily available for the various simple and complicated analytical and quantitative techniques of which researchers generally make use of. The only work a researcher has to do is to feed in the data he/she gathered after loading the operating system and particular software package on the computer.

The output, or to say the result, will be ready within seconds or minutes depending upon the quantum of work. The storage facility of computers provides statistician the immense help to use the stored data whenever it required. Innumerable data can be processed and analyzed with greater ease and speed. Moreover, the results obtained are generally correct and reliable. Not only this, even the design, online data collection, pictorial graphing and report are being developed with the help of computers.

Hence, researchers should be given computer education and be trained in the line so that they can use computers for their research work.

Researchers interested in developing skills in computer data analysis, must be aware of the following steps:

(i) data organization and coding;

(ii) storing the data in the computer;

(iii) selection of appropriate statistical measures/techniques;

(iv) selection of appropriate software package;

(v) execution of the computer program.

First of all, researcher must pay attention toward data organization and coding prior to the input stage of data analysis. If data are not properly organized, the researcher may face difficulty while analyzing their meaning later on. For this purpose, the data must be coded. Categorical data need to be given a number to represent them. Once the data is coded, it is ready to be stored in the computer. Input devices may be used for the purpose. After this, the researcher must decide the appropriate statistical measure(s) he will use to analyze the data. Researcher will also have to select the appropriate program to be used. MS Excel, SPSS, SAS, STATA, R, Python etc. are the special statistical packaged program whereas Microsoft Excel can be used for simple statistical analysis. SPSS, SAS, STATA are the windows-based user-friendly program where as R, Python are the freeware where we have to write code to run the mathematical and statistical calculation. Among this all MS Excel is easily accessible and user-friendly software which mostly we are familiar with it. So, this data analysis workshop program is designed with MS Excel software.

Data analysis using Microsoft Excel

Microsoft Excel

Microsoft Excel helps you to organize, attractively present and analyze data.

A spreadsheet is the computer equivalent of a paper ledger sheet. It consists of a grid made from columns and rows. It is an environment that can make number manipulation easy and somewhat painless.

The statistics that goes on behind the scenes on the paper ledger can be overwhelming. If you change the any amount, you will have to start the calculation all over again (from scratch). The nice thing about using a computer and spreadsheet is that you can experiment with numbers without having to RE-DO all the calculations.

NO erasers! NO new formulas! NO calculators!

Excel has many applications:

- Sorting and organizing data
- Creating visual representations of the data
- Addition, Subtraction, Division, Multiplication, percentage of Cells
- Statistical analysis
    - Average (Mean)
    - Median
    - Quartile
    - Standard deviation
    - Estimation
    - Test of hypothesis (parametric and non-parametric)
    - Correlation and regressions…
- Matrix Operations
    - Addition/Subtraction
    - Multiplying
    - Inverse
- Optimization
    - Linear programming
    - Transportation
    - Assignments
- And many more…