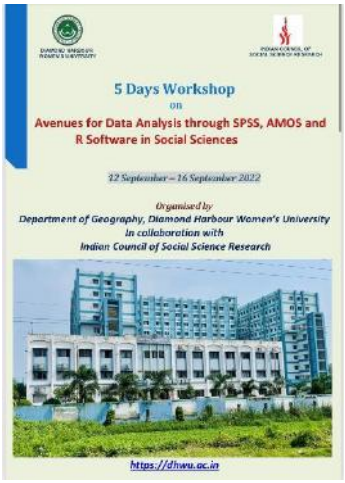
Background of the Workshop

5 Days Workshop on "Avenues for Data Analysis through SPSS, AMOS and R Software in Social Sciences" was organized by Department of Geography, Diamond Harbour Women's University in collaboration with Indian Council of Social Science Research (ICSSR) from 12.09.2022 to 16.09.2022. With the release of the brochure, several enquiries about the programme were received which denoted the enthusiasm among the fraternity and soon the seats got filled-up.



Front page of the Brochure

A total of 42 participants attended the workshop, out of which 16 were scholars of the department and rest 26 were from several universities like University of Calcutta, Jadavpur University, Bankura University, University of Burdwan, University of Kalyani and Kazi Nazrul University. 2 participants were from outside state from Atal Bihari Vajpayee University and OPJS University.

Table 1: Overview of participants across the institutions

Affiliation (College/ University/ Institution)	Frequency	Percent
Atal Bihari Vajpayee Viswavidyalaya	1	2.4
Balarampur College, Purulia	1	2.4
Bankura Sammilani College	1	2.4
Bankura University	1	2.4
Jadavpur University, Centre for Studies in Social Science Calcutta	1	2.4
Diamond Harbour Women's University	17	40.5
Kazi Nazrul University	2	4.8
Midnapore College Autonomous	1	2.4
OPJS University	1	2.4
The University of Burdwan	4	9.5
University of Calcutta	3	7.1
University of Kalyani	1	2.4
Vivekananda College for Women, Calcutta University	8	19.0
Total	42	100.0

Representatives were from 10 different districts of West Bengal. 47.6% of participants are from Kolkata followed by South 24 Parganas, Paschim and Purba Bardhhaman and

Bankura. There are only 2.4% of participants each from the Birbhum, Hooghly, Howrah, Murshidabad and Purulia Districts.

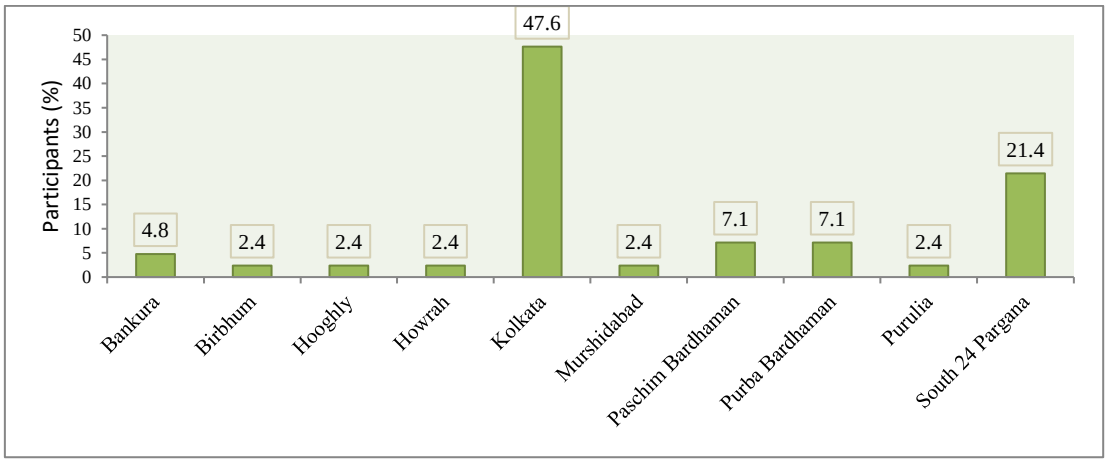


Fig 1: District-wise participation at a glance.

Majority of the participants were from geography background; one was from English department and another from social science. More than 75% of participants are female while only a few (9 participants) were male. Another fact that deserves a mention is that nearing 60% participants were from backward community (SC, ST and OBC categories).

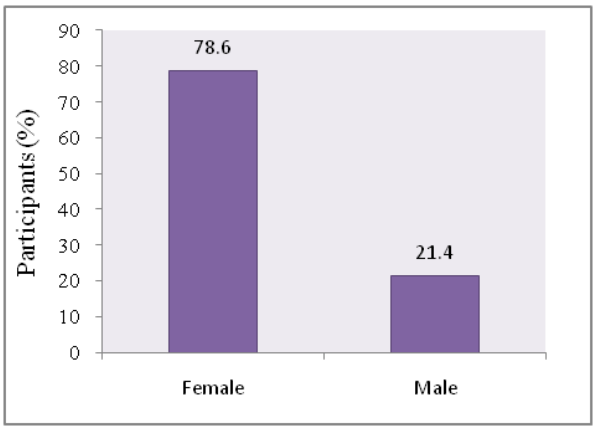


Fig 2: Gender composition of the participants.

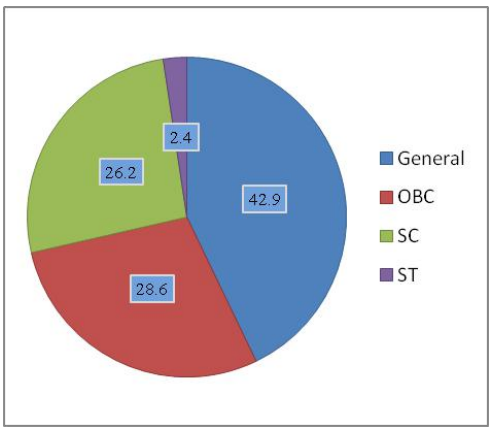


Fig 3: Social category of the participants.

So, the workshop was able to have a wide reach geographically and socially as well. Apart from venue management, the organizing committee arranged for local hospitality, hostel accommodation for the outstation participants and pick-up – drop service for the participants coming from different parts of the city.

Day 1 (12.09.2022)

The programme started with the inaugural speech delivered by the Chief Patron of the webinar Prof. Soma Bandhyapadhyay, Hon'ble Vice Chancellor of Diamond Harbour Women's University through audio message and congratulated the Department of Geography for organising this sort of relevant and appropriate workshop. The Hon'ble Vice Chancellor Madam explained briefly the importance of the three software i.e., SPSS, AMOS and R in analysing data in social science research.

Inauguration by lighting of lamp

In the inaugural session, the stage was graced by the Chief Advisor of the workshop and Dean of Science, Prof. Dilip Das and IQAC Co-ordinator, Prof. Raj Kumar Kothari. The Convener of the workshop and Head of the Department of Geography, Dr. Shovan Ghosh delivered the welcome address. He warmly welcomed the guests, resource persons and all the participants. He acknowledged the hard work and the efforts of the collaborating institute ICSSR and the faculty members of the department.

Welcome address The Convener Dr. Shovan Ghosh

Prof. Das, started his special address with a thanking note for IQAC Co-ordinator, HOD and the resource persons for organising the workshop. He emphasised the importance of the advanced data processing and analysis by modern software in current day researches. He also

hoped that the workshop becomes a platform for sharing knowledge and experiences among the participants and the resource persons. He also conveyed best wishes to the participants.

A section of audience at the inaugural session

IQAC co-ordinator, Prof. Raj Kumar Kothari mentioned his observations regarding the new scientific advancement in the field of geography and other social sciences. He explained how behavioural methodology gained importance after early 1960s. He also put forward the other important methods of scientific research like empiricism. He wished for the grand success of the workshop.

After the ceremonial felicitation of the dignitaries the session was wrapped up with a formal Vote of Thanks by Organising Secretary, Dr. Anindya Basu. She thanked the Hon'ble

Vote of Thanks by Organising Secretary, Dr. Anindya Basu

Vice Chancellor Madam, Registrar Prof. Dr. Sayeedur Rahaman, ICSSR-ERC Director Prof. Saibal Kar, Head of the Department Dr. Shovan Ghosh, along with Prof. Sujit Mondal and other faculty members of the department for their constant effort and support. She also thanked the research scholars and students of the department for handling the nitty-gritty regarding the workshop arrangements. She expressed her immense gratitude to the two resource persons Dr. Baidya Nath Pal, Associate Scientist, Indian Statistical Institute and Prof. Rahul Bhattacharya, Professor and Head, Department of Statistics, University of Calcutta. She also expressed

happiness seeing the presence of all the participants in spite of the bad weather conditions on the very first day.

The inaugural session was followed by light refreshment and then the two technical sessions by Dr. Baidya Nath Pal, Associate Scientist 'A', Biological Anthropology Unit, Indian Statistical Institute, Kolkata and his associates ensued.

Technical Session IA

In the first technical session Dr. Baidya Nath Pal gave a brief introduction about the SPSS software. Before entering to the data entry part for this particular software platform, Dr. Pal unfolded the elements of data. The discussion started with qualitative and quantitative data type and the session was continued by revelling the pros and cons of discrete or categorical data and continuous data. Dr. Pal also unveiled the domain of different scales used in this software i.e., nominal scale, ordinal scale and interval scale for different data type. After a brief conceptual deliberation, the participants were getting used to with the SPSS software (Trial Version 26.0), provided by the team and the research scholars of the Department of Geography, DHWU. IBM-SPSS Statistics is a powerful statistical software platform. It offers a user-friendly interface and a robust set of features that lets researcher quickly extract actionable insights from the dataset. According to the experienced researchers, advanced statistical procedures help ensure high accuracy and quality decision making. All facets of the analytics lifecycle are included, from data preparation and management to analysis and reporting.

Technical Session by Dr. Baidya Nath Pal

The resource team explained each and every basic features of the software to the participants. They got to know about the research jargons like data view, variable view, ID, width, variable type, labels, values, string, alpha characters, numeric etc. along with its hands

on examples. After the conceptual session, the participants got the entry to the domain of statistics using SPSS. The coding-recoding, descriptive statistics, split cells, frequency distribution, the methods of deriving mean, median, mode etc. and other basic statistical knowledge had been shared before the lunch hour.

Technical Session IB

After lunch break the second technical session started. Emphasis was given on the finding the distribution, normality check and outliers of a data. The outlier concept was thoroughly explained by the resource team. It was followed by explaining the method of computation of different parametric tests and non-parametric tests. Among the parametric test, three types of t-Tests i.e. One-Sample t-Test, Two-Sample t-Test and Paired Sample t-Test along with brief introduction of ANOVA, had been explained using SPSS but the weightage was given to the non-parametric test as these tests are used for those data that are not normally distributed.

Technical session by Dr. Baidya Nath Pal

The chi-square test was performed using SPSS and the explanation of each output table had been clarified. The conceptual idea of odd ratio was given to the participants and simultaneously deriving odd ratio while calculating chi-square had also been performed.

The hands-on training was supervised by the resource person Sanchari Ghosh and Nairita Majumder, two team members of Dr. Pal. Finally, after an eventful day and interactive session, the Convener and Organising Secretary called for the closing of the 1st day of workshop.

Day 2 (13.09.2022)

The second day of workshop had lectures and hands on training sessions on SPSS covered by Dr. Baidya Nath Pal, Associate Scientist 'A', Biological Anthropology Unit, Indian Statistical Institute, Kolkata and his associates. The topics ranged from ANOVA one way, ANOVA two-way, Correlation, Bi-variate Regression, Logistic Regression, Factor Analysis, Principal Component Analysis, Cluster Analysis, ANCOVA, to MANOVA and MANCOVA.

Technical Session IIA

The session commenced in the morning after a light refreshment. Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship. The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two. One-way ANOVA uses one independent variable, while a two-way ANOVA uses two independent variables. One way ANOVA is a hypothesis test, used to test the equality of three of more population means simultaneously using variance. Two-way ANOVA is a statistical technique wherein, the interaction between factors, influencing variable can be studied. Three or more levels of one factor two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables.

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Correlation is a statistical measure that expresses the extent to which two variables are linearly related (means they change together at a constant rate). It is a common tool for describing simple relationships without making a statement about cause and effect. Different Types of Correlation:

- Positive and negative correlation.
- Linear and non-linear correlation.
- Simple, multiple, and partial correlation.

As a rule of thumb, a correlation greater than 0.75 is considered to be a “strong” correlation between two variables. However, this rule of thumb can vary from field to field. For example, a much lower correlation could be considered strong in a medical field compared to a technology field.

Statistical modelling regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables, in combination, affect a dependent variable. Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.

Second, in some situation regression analysis can be used to infer causal relationships between the independent and dependent variables. The p-value indicates if there is a significant relationship described by the model. Essentially, if there is enough evidence that the model explains the data better than would a null model. The R-squared measures the degree to which the data is explained by the model. High p-values indicate that evidence is not strong enough to suggest an effect exists in the population. An effect might exist but it is possible that the effect size is too small, the sample size is too small, or there is too much variability for the hypothesis test to detect it.

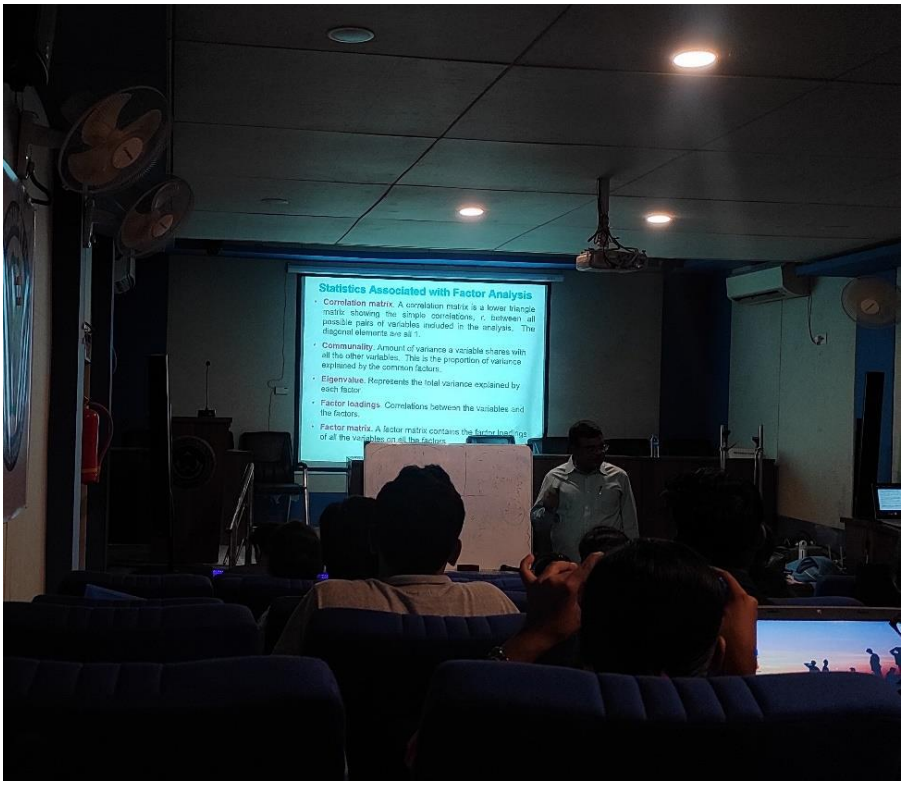
In the morning session the discussion about the statistical analysis was very elaborative where firstly, classification of ANOVA, type of data which would be used to perform such analysis were described. Besides on what basis one chooses T test over ANOVA was also analysed. Interpretation of output layers with significant or not significant values as well as why choosing continuous or categorical data and dependent or independent variable also exemplified with ample of information. Secondly, for modification of analysis different estimates described, beside that correlation, partial correlation assumption, conditions, processes, examples, use of 0 and 1 to measure identity matrix profusely explained. Lastly, different parts of regression elaboratively discussed with the concepts like normality check, collinearity, homoscedasticity, autocorrelation, standardized predicted value etc. and validity checks like WATSON TEST, COOK' ranges.

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analysing the relationship between one or more existing independent variables. Logistic regression is applied to predict the categorical dependent variable. In other words, it is used when the prediction is categorical, for example, yes or no, true or false, 0 or 1. The predicted probability or output of logistic regression can be either one of them, and there's no middle ground. The output of a logistic regression model is the

probability of our input belonging to the class labelled with 1. And the complement of the model's output is the probability of the input belonging to the class labelled with 0. Logistic Regression should not be used if the number of observations is fewer than the number of features; otherwise, it may result in overfitting. Because it creates linear boundaries, one will not obtain better results when dealing with complex or non-linear data.

Technical Session IIB

After the lunch break afternoon session started. It covered the areas like Factor Analysis, Principal Component Analysis, Cluster Analysis, ANCOVA, MANOVA and MANCOVA.



Technical session by Dr. Baidya Nath Pal

Factor analysis is a powerful data reduction technique that enables researchers to investigate concepts that cannot easily be measured directly. By boiling down many variables into a handful of comprehensible underlying factors, factor analysis results in easy-to-understand, actionable data. The purpose of factor analysis is to reduce many individual items into a fewer number of dimensions. Factor analysis can be used to simplify data, such as reducing the number of variables in regression models. Most often, factors are rotated after extraction. In statistics, factor analysis of mixed data or factorial analysis of mixed data is the factorial method devoted to data tables in which a group of individuals is described both by quantitative and qualitative variables. The eigenvalue is a measure of how much of the common variance of the observed variables a factor explains. Any factor with an eigenvalue ≥1 explains more variance than a single observed variable.

Principal Component Analysis is a tool for identifying the main axes of variance within a data set and allows for easy data exploration to understand the key variables in the data and spot outliers. Properly applied, it is one of the most powerful tools in the data analysis tool kit. Principal Component Analysis (PCA) is very useful to speed up the computation by reducing the dimensionality of the data. Plus, when you have high dimensionality with high correlated variable of one another, the PCA can improve the accuracy of classification model.

Cluster analysis is a statistical method for processing data. It works by organizing items into groups, or clusters, based on how closely associated.

Analysis of covariance (ANCOVA) is a method for comparing sets of data that consist of two variables (treatment and effect, with the effect variable being called the variate), when a third variable (called the covariate) exists that can be measured but not controlled and that has a definite effect on the variable. ANOVA is a process of examining the difference among the means of multiple groups of data for homogeneity. ANCOVA is a technique that remove the impact of one or more metric-scaled undesirable variable from dependent variable before undertaking research. Both linear and non-linear model are used.

The Multivariate analysis of variance (MANOVA) procedure provides regression analysis and analysis of variance for multiple dependent variables by one or more factor variables or covariates. The factor variables divide the population into groups. The general purpose of multivariate analysis of variance (MANOVA) is to determine whether multiple levels of independent variables on their own or in combination with one another have an effect on the dependent variables. MANOVA requires that the dependent variables meet parametric requirements.

Technical session by Dr. Baidya Nath Pal

Multivariate analysis of covariance (MANCOVA) is an extension of analysis of covariance (ANCOVA) methods to cover cases where there is more than one dependent variable and where the control of concomitant continuous independent variables – covariates – is required. A MANCOVA is identical to a MANOVA, except it also includes one or more covariates. Similar to a MANOVA, a MANCOVA can also be one-way or two-way.

The main issues which covered were –

- Difference between logistic and linear regression, interpretation of both the regressions
- Assigning the binary codes into categorical data
- Association with data and validity check models like NAGELKERKE R SQUARE, DUMMY R SQUARE, WALD TEST
- Classification of Factor Analysis (EFA and CFA), goals, variance error, BARTLETT'S TEST and Kaiser-Meyer- Olkin measure

After a fruitful interactive session, the workshop schedule of that day ended with high tea.

Day 3 (14.09.2022)

Both the sessions were conducted by Dr. Baidya Nath Pal, Associate Scientist 'A', Biological Anthropology Unit, Indian Statistical Institute, Kolkata and his associates. The first session was covered theoretical aspects of Structural Equation Modelling (SEM). The second session revolved around hands-on training session of AMOS.

Felicitation of Dr. Md. Sayeedur Rahman, Hon'ble Registrar,DHWU

Technical Session IIIA

SEM is an extension of the general linear model (GLM) that enables a researcher to

test a set of regression equations simultaneously. In other words, the purpose of SEM is to examine a set of relationships between one or more exogenous Variables (independent variables) and one or more endogenous variables (dependent variables).SEM software can test traditional models, but it also permits examination of more complex relationships and models, such as confirmatory factor analysis and time series analysis. Moreover, through SEMs the structural relations can be modeled graphically to enable a clear understanding of the theory under study.

Benefits of using SEM

When compared to old multivariate procedures, several advantages can be seen in SEM usage -

- It conducts a confirmatory rather than exploratory approach to the data analysis (Exploratory approach also can be conducted through SEM).
- SEM estimates error variance parameters but traditional multivariate procedures are incapable of estimating the measurement error.
- SEM can incorporate both observed and latent variables, whereas formermethods are based on observed measurements only.
- Researcher can get a unifying framework that fit numerous linear models by using SEM.
- SEM programs provide overall tests of model fit and individual parameterestimate tests simultaneously.
- Regression coefficients, means, and variances may be compared simultaneously, even across different groups.
- Longitudinal data, databases with auto correlated error structures (time series analysis), databases with non-normally distributed variables and incomplete data can be handled.

Because of these advantages of SEM, it has become a popular methodology in non-experimental research.

Common terms related with SEM

- *Observed and latent variables*

Especially in behavioral and social sciences, researchers are often interested in studying two types of theoretical constructs, namely observed (manifest) and latent variables. Observed variables can be observed directly (e.g., income, blood sugar etc.).But researches very often have to deal with latent variables that cannot be directly measured such as personality,

perception, buying behavior etc. Researches use observed variables to measure the latent variables. The observation may include, for example, self-report responses to attitudinal scale, coded responses to an interview question and the like. These measured scores or in other words observed or manifest variables are used to measure the latent variables.

- *Exogenous and endogenous latent variables*

Exogenous latent variables are synonymous with independent variables and endogenous latent variables are synonymous with dependent variables. Endogenous variables are influenced by exogenous variables directly or indirectly.

- Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) Factor analysis is conducted to investigate the relations between sets of observed andlatent variables. If the links between the observed and latent variables are unknownor uncertain, exploratory factor analysis is conducted. Exploratory factor analysis isconducted to determine how, and to what extent, the observed variables are linked totheir underline factors. Confirmatory factor analysis is appropriate, when theresearcher has some understanding (through theory, empirical research or both) ofthe latent variable structure.

- The path diagram Path diagram is a visual representation of relations among variables which are assumed to hold in the study. Basically, four geometric symbols are used in the path diagrams; circles or ellipses represent unobserved latent variables, squares or rectangles represent observed variables, single- headed arrows represent the effect of one variable on another variable, and double-headed arrows represent covariance or correlation between two variables. Figure 4, is the simple model used to explain the meanings of the symbols of a path diagram.

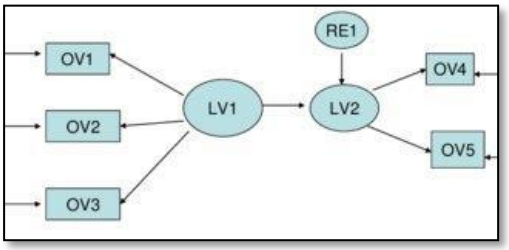


Figure 4: Simple path diagram

ME- Measurement Error RE- Residual Error OV- Observed Variable LV- Latent Variable

Technical Session IIIB

Before starting of this session, the required software was provided with installation tutorial and during session power point presentation on SEM and sample data were given for hands-on practice.

Steps Involved in SEM

i. Open IBM SPSS Amos and save the file by selecting File > Save. A window will open.
ii. Import the SPSS dataset by selecting “Data Files” from the menu. A dialogue box will appear. Select File Name > location of file > file > open > Ok
iii. Draw the path diagram using the draw latent or its indicator icon. For moving a figure, select moving the object icon and then move the variable as per the requirement. You can also duplicate the model by selecting the duplication of the object icon. For rotating the diagram click of rotating the latent variable icon and for moving the drawn path diagram click on the symmetrical movement icon and then move the figure. Finally, to draw the dependent variable, the observed variable is drawn using the draw the observed variable icon and in order to include the measurement error in the computation of the value of perceived performance, click on the draw unique variable icon and then on the drawn variable. Link the constructed variables.
iv. Specify each variable using the imported dataset. For this select the icon presenting a list of the dataset. A below-shown dialogue box will appear. Drag each variable from this dialogue box on the drawn observed variable boxes. After observed variables specification, state the latent variables by double-clicking on the latent variable. A dialogue box will appear. Enter the name of the variable. Similarly, specify each latent variable.
v. Name all the unobserved variables i.e., residual and measurement error by clicking on Plugins > Name Unobserved Variables.
vi. Finally click on the calculate estimates icon to calculate the estimates. A dialogue box will appear.
vii. Click on Proceed with the analysis. Results of the analysis will appear.
viii. Further, a new result file will be created at the location where you saved the Amos file. Open the file.

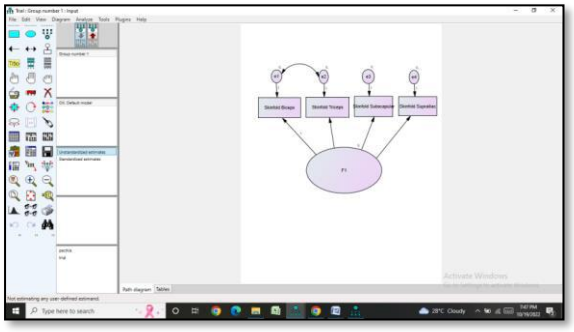


Fig 5: Path diagram representation in Amos window

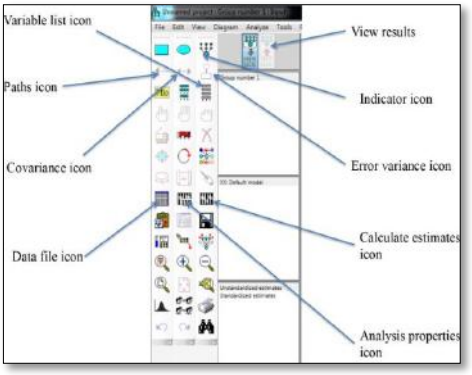


Fig 6: Various icons related with AMOS

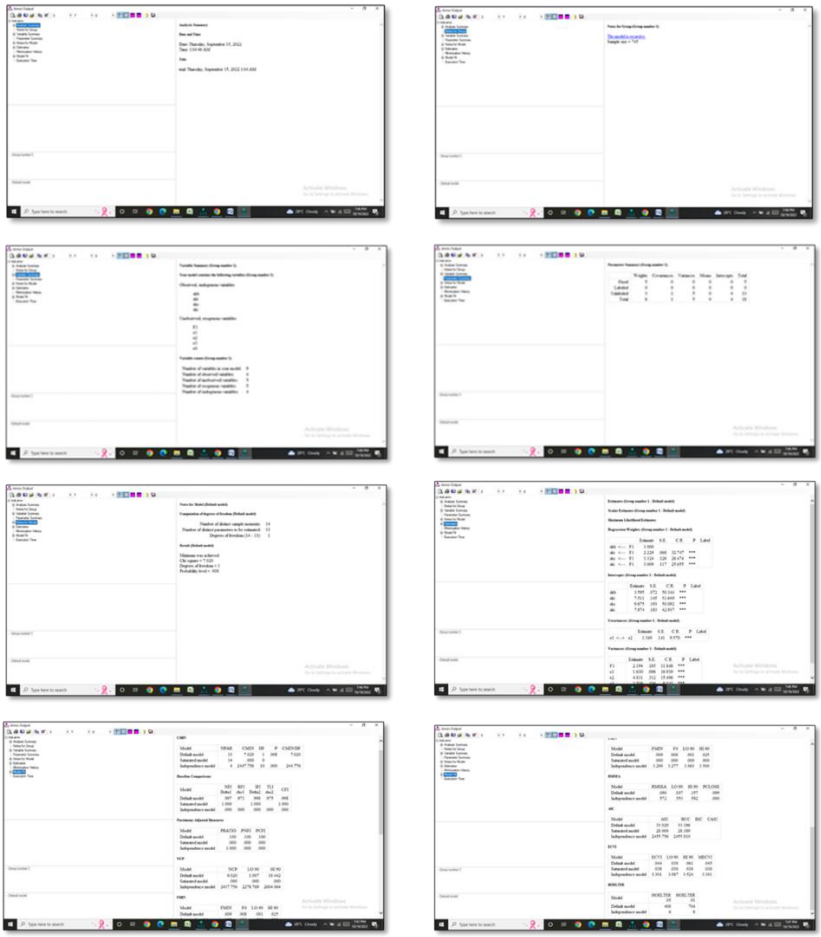


Fig.7: Different results of Amos output.

There are several steps in SEM analyses; model specification, model identification, model testing, and model modification. In model specification, researcher specifies the model determining every relationship among variables relevant to the researcher's interest. 30 SEM programs require an adequate number of known correlations or covariances as inputs in order to generate a sensible set of results. Identification refers to the idea that there is at least one unique solution for each parameter estimate in a SEM model. Models in which there is only

one possible solution for each parameter estimate are said to be just-identified. Models for which there are an infinite number of possible parameters estimate values are said to be under-identified. Finally, models that have more than one possible solution (but one best or optimal solution) for each parameter estimate are considered overidentified. Model is considered as identified if the model is either just- or overidentified. If a model is identified only, the parameter estimates can be trusted.

Group photo of perticipants with Dr. Pal

In SEM modelling researcher can use three main approaches to test whether the data fit the model; Confirmatory approach, Alternative model approach, and Model generating approach If data does not fit with the model generated by the researcher, model modification is done to receive a final best model. Amos generates alternative models by specifying optional and/or required paths in a model. Hence, in AMOS, researcher needs not to generate and delete paths to find out the best model. In order to test the model fit, absolute model fit, test of relative fit, Parsimonious Fit Indices can be used. Absolute model fit criteria commonly used are chi-square (x2), the goodness-of-fit index (GFI), the adjusted goodness-of-fit index (AGFI), and the root-mean-square residual (RMR) and the Root Mean Square Error of Approximation (RMSEA). Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI) compare the

absolute fit of your specified model to the absolute fit of the Independence model and used as the criteria to test the relative fit. The Parsimony Goodness-of-Fit Index (PGFI) and the Parsimonious Normed Fit Index (PNFI) are good examples to test the Parsimonious Fit.

Day 4 (15.09.2022)

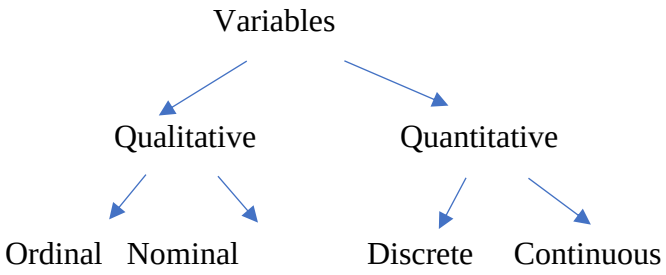
Prof. Rahul Bhattacharya, Head, Department of Statistics, University of Calcutta and his team was entrusted with explaining the nuances of R software for the final two days.

Topics covered on this day were: basics of statistics and types of data; brief overview about R software; graphical functions in R software and correlation and regression – in the first session, along with hands-on training of the software in the ensuing second session.

Technical Session IVA

A detailed discussion on data type was done.

Data types depending on variables -



Prof. Rahul Bhattacharya, Head, Department of Statistics, University of Calcutta

Qualitative variable: Variables that can't be measured numerically (e.g., hair color, gender)

• Ordinal variable: A natural ordering is possible (e.g., grades (A+, A, B etc.) in an examination).

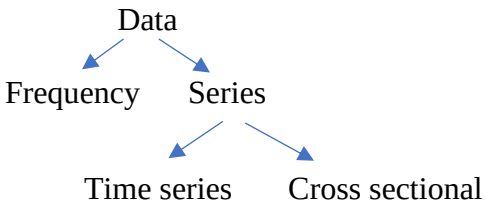
• Nominal variable: No such ordering exists. (e.g., blood group, response to a post on Facebook).

Quantitative variable: Variables that can be measured numerically (e.g., height, weight)

• Discrete variable: Variables taking only integer values. (e.g., number of births in a region).

• Continuous variable: Variables taking all possible type of numerical values. (e.g. height, weight).

Data types depending on the recording –



Frequency: The number of times of occurrence of a particular value in the data set.

Frequency Data: Data recorded in the form of frequencies (e.g., Number of misprints in a book)

Series Data: Data recorded in the form of a series.

Time series Data: Data collected on the same unit for the same variable but for different time points (e.g., Number of accidents for the last 10 days, price of an item over the years)

Cross sectional data: Data collected on different units for the same variable at the same time point (e.g., Current prices for 10 cars).

Hands-on practice session supervised by resource persons

If cross-sectional data relate to different places, it is called Spatial series data (e.g., Maximum temperature of different cities in India on a particular day)

A brief overview about the open-source R software was given. R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. So mainly it is a programming language with a statistical package and its open worldwide. R is a Programming language, Statistical package, and open source software that starts with a letter (A-Z or a-z) and must contain letters, digits (0-9), and/or periods "." This software is developed by Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland in 1995.

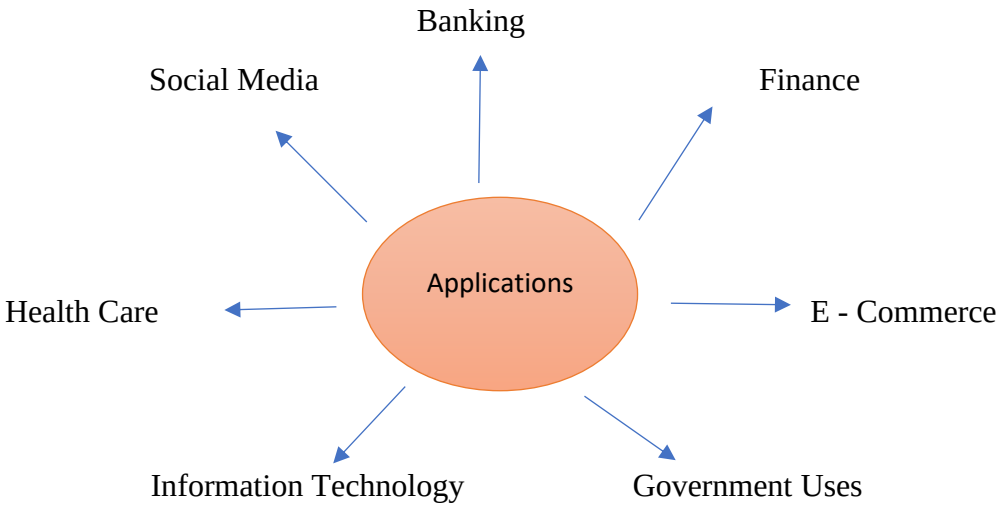


Fig 9: Various applications of R software.

Data Manipulation involved entry of data to R software and data entry from excel.

To start R software and entry of data first of all own directory must be set by following steps-

Create a blank folder in Desktop with name "R files" - Open RStudio - Go to Session-Set Working Directory-Choose Directory - Your directory is set at the folder "R files".

The steps which include to entry data from excel –

First select add file ⟶ import dataset ⟶ from excel

Typing any existing file name in console of R software the file can be seen, and for any particular element or factor calculation one can use dollar $ sign for particular statistical calculation.

The one and foremost function of R has many inbuild data sets from which one can type command data () and can use any representation.

Technical Session IVB

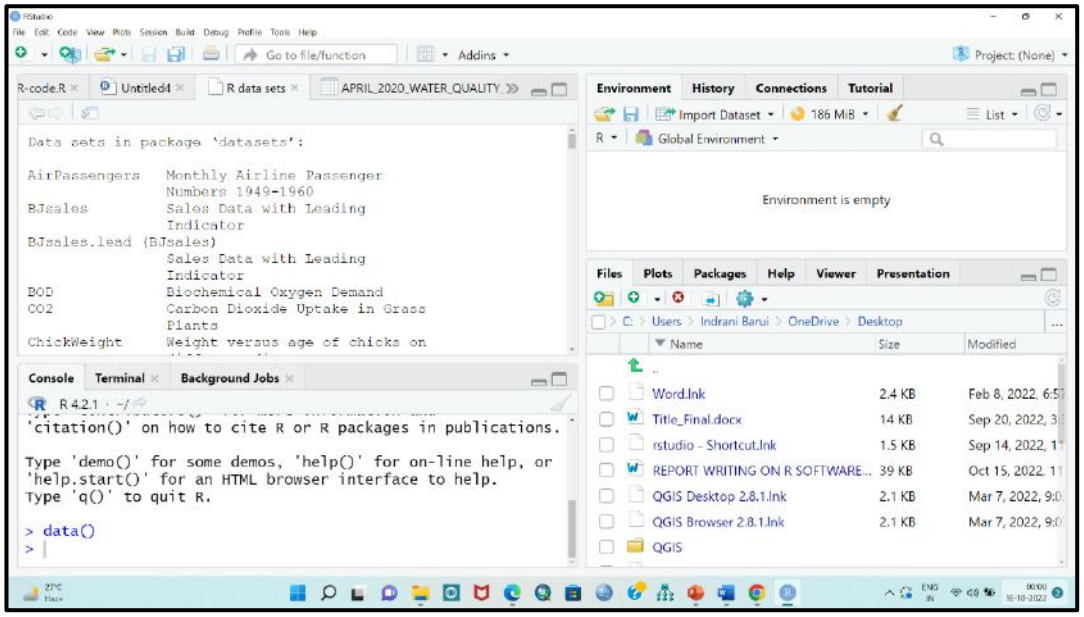


Fig 10: R software interface.

Multiple statistical techniques can be done by typing exact formula with correct programming language and existing file name. The session also clarified the use of mean, median, SD, along with examples from data sets. Help tool is very useful for memorizing the sets of data availability.

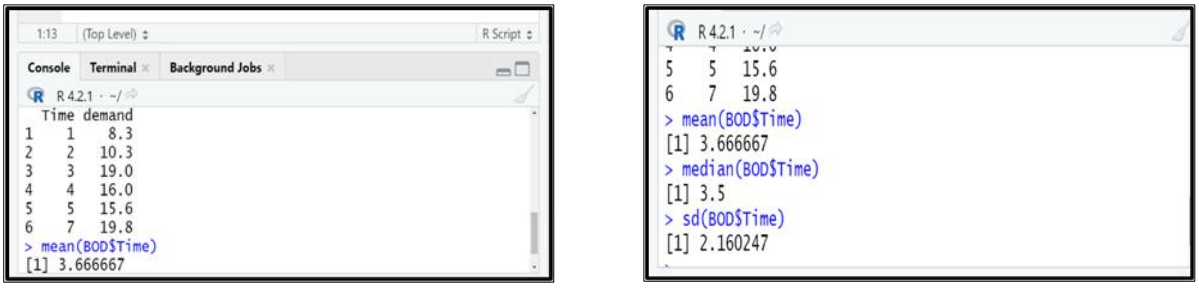


Fig 11: Basic statistical analysis through R software.

Various graphical function has been presented and described in this session were, plotting of Simple Bar diagram, Line diagram, grouped Bar plot, Pie diagram.

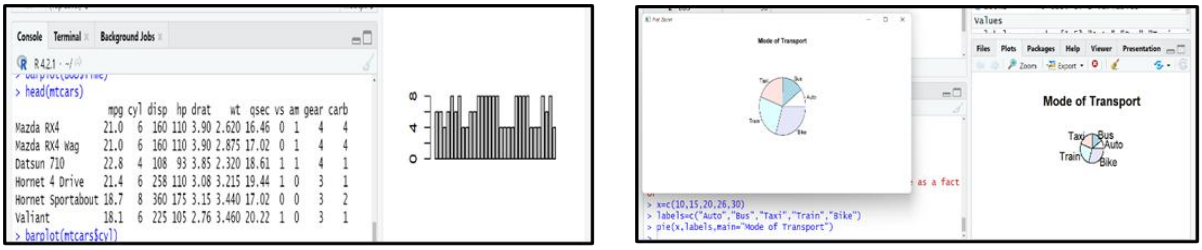


Fig 12: Graphical representations through R software.

Frequency distribution has been plotted for qualitative data using Dot plot, and histogram using quantitative data. Various other mathematical representation follows stem leaf diagram, discrete frequency distributions, dispersion, deviation and its measurements.

The topics which covered lastly were basic essence of scatter diagram plotting and corelation and regression in R software. Scatter diagram is the simplest diagrammatic representation of bivariate data. Scatter diagram tells how the values of one variable change with change in the values of another variable. The session also discussed the correct interpretation of r value.

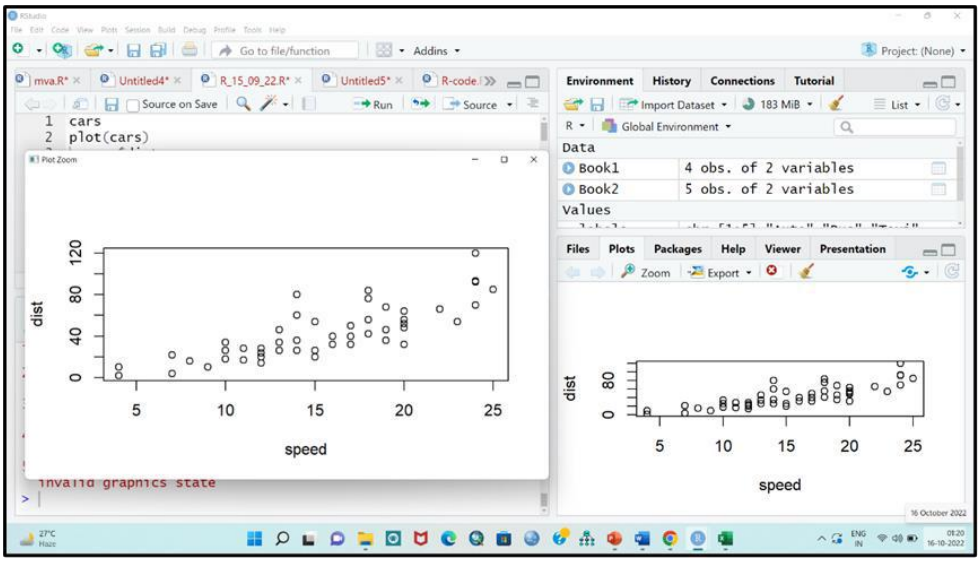


Fig 13: Corelation and Regression, R software

The session also introduced corelation types its interpretation, linear regression, estimation of parameters using least square method, R2 and Adjusted R2 and their role in checking model adequacy, non-linear regression and Regression Line Plotting on Scatter Diagram.

LM (Linear Model) and GLM (Generalized Linear Model) were introduced and each participant run the models on their own. The session was totally interactive one.

The session included hands-on training which helped to have a clearer idea on the applicability part of the concepts and procedures discussed.

Day 5 (16.09.2022)

The last day of the workshop had two segments: a special lecture session on quantification and a session on R software.

Technical Session VA

This day was divided into two halves. In the first half, the special lecture was delivered by Prof. Achin Chakraborty, the Director, Institute of Development Studies Kolkata. At the beginning of this day, an inauguration session had been arranged for him, and then he started

an outstanding lecture about the “Quantitative Aspects of Development**”**. He mainly focused on the importance of value judgement in every aspect of evaluation process and highlighted exploratory and evaluative research methodology. He stated the application of ordinal scale and cardinal scale in framing of research questions.

Prof. Achin Chakraborty, the Director, Institute of Development Studies Kolkata

Development is broad and all-inclusive a person's total transformations which can be evaluated through careful observation. While some children do not grow in terms of height, weight, or size, they do experience functional improvement or development. According to him, research is like an idea development and it is a thinking process that frequently combines elements of engineering, computer science, social science, cognitive science, and learning sciences.

Research work should be based on the author’s idea from the research topic to the way the author responds to it. A researcher should take a moment to think things over before starting an outline. He gave an outstanding example of census data. During the census, most of the people used to talk about their age in a round figure. That time surveyors believe him/her and exactly write down the same thing and after that, all the tables, charts, and diagrams have been made depending on the subject matter. But no one bothers to check its authenticity. Most of the researchers think the data from different sources fit with their research question or not but they don’t think about new crucial issues or ideas in resolving them.

Prof. Achin Chakraborty during his outstanding lecture on "Quantitative Aspects of Development"

Quantitative aspects look into how statistical techniques like modelling, time-series data, and various multivariate techniques are affected by realistic data. Researchers have to look into how methods can be changed and new methods developed and explore methods to enhance the assessment of research. While talking about development, he spoke about the improvement of society by reducing gender discrimination in the 21st century. The

development of society is an indicator of a reflection of human thinking. The lecture was extremely thought proving.

Technical Session VB

This session was started by Prof. Rahul Bhattacharya who is a Professor at Calcutta University in the Department of Statistics. In the first half, he and his team gave a hand session practice on R software and also gave a valedictory lesson about 'Basics of Statistical Inference in R'.

Analysis of Variance (ANOVA) is a statistical test for determining how the levels of one or more categorical independent variables affect the levels of a quantitative dependent variable. It is used to investigate relations between categorical variables and continuous variables in R Programming.

ANOVA is two types- One-way and Two-way Anova, which analyze the variance in the data to look for differences. It does this by considering two sources of variance, the between-

group variance and the within-group variance. For both types of variances, a sum of squares (SS) is the numerical metric used to measure them and this metric simply sums the distances of each point to the mean. The ratio of these SS (between SS divided by within SS) results in an F-statistic, which is the test statistic for ANOVA. The p-value is then calculated by combining the F-statistic with the degrees of freedom (df). The F-statistic is the test statistic for ANOVA. The associated p-value can assist in interpreting the importance of the F-statistic. The p-value of less than 0.05 indicates that there is at least one group's mean differs from another at the $\alpha=0.05$ level of significance. We can run our ANOVA in R using different functions. The most basic and common functions we can use are aoy() and lm().ANOVA is a type of linear model, we can use the lm().

Procedural Details of ANOVA

After downloading R and R Studio we have to go to R studio and click on File > New File > R Script.

Step 1: Load the data into the R

Only quantitative or categorical variables are used here when importing a dataset into R. We can use the read.csv() command to read the data.

A section of audience at the technical session

Step 2: Perform the ANOVA test

One can perform an ANOVA in R by using aov() function. This will calculate the test statistic for ANOVA and determine whether there is significant variation among the groups formed by the levels of the independent variable.

Step 3: Find the best-fit model

There are various ANOVA models to explain the data among them Akaike information criterion (AIC) is a good test for model fit.

Step 4: Check for homoscedasticity

To check whether the model fits the assumption of homoscedasticity or not in R by using the plot() function.

Step 5: Do a post-hoc test

ANOVA tells us if there are differences among group means. To find out which groups are statistically different from one another, you can perform a Tukey's Honestly Significant Difference (Tukey's HSD) post-hoc test for pairwise comparisons.

Step 6: Plot the results in a graph

Step 7: Report the results

ANOVA test includes:

- A brief description of the variables you tested
- The f-value, degrees of freedom, and p-values for each independent variable
- What the results mean.

Technical Session VC

In this post-lunch session, Time Series Modelling in R: AR, MA, and ARMA were covered.

A time series data is created by any metric that is measured throughout time with regular intervals. Eg. Information about the weather, stock prices, industry forecasts, etc. R has a powerful inbuilt package to analyze the time series. Here, a function is built to take distinct process components.

Time series forecasting is the use of a model to predict future values based on previously observed values. However, it is crucial to confirm that the time series remains stationary during the span of the historical observational data.

$$Yt=Tt+St+\varepsilon t$$

AR (autoregression) modelling

The autoregression (AR) model shares the very familiar interpretation of a simple linear regression, but here each observation is regressed on the previous observation.

$$\text{AR (p) model: } Yt=\varphi 1Yt-1+\varphi 2Yt-2+\ldots+\varphi pYt-p+\varepsilon t$$

MA (moving average) modelling

The simple moving average (MA) model is used to account for very short-run autocorrelation. It does have a regression like form, but here each observation is regressed on the previous innovation, which is not observed

$$\text{MA (q) model: } Yt=\varepsilon t + \theta 1\ \varepsilon t\text{-}1+ \theta 2\ \varepsilon t\text{-}2+\ldots+ \theta q\ \varepsilon t\text{-}q$$

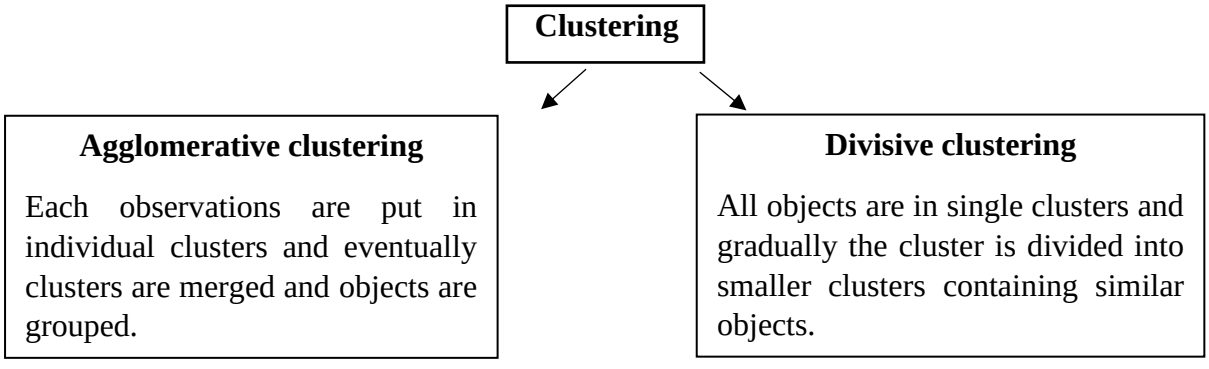
ARMA modelling

Combining both AR (p) and MA (q) models. In this model, residuals and the effects of earlier lags are taken into account for predicting future values of the time series.

Classification and Cluster Analysis in R

In clustering techniques, objects are grouped or classified in such a way similar objects are clustered in a single group. Hierarchical clustering in R Programming Language is an unsupervised non-linear algorithm in which clusters are created in a hierarchical manner. The standard R function plot. hclust() can be used to draw a dendrogram from the results of

hierarchical clustering analyses. Eg. Consider a bank hierarchically. A clerk, officer, and manager at a bank all are grouped into the same work and same field which form a hierarchy. Two Clustering can be achieved either in an agglomerative way or a divisive way.



K Means Clustering is a divisive clustering where objects are classified into a pre-selected number of clusters. The number of clusters can also be guided by Sugar and James' (1994) algorithm.

So, R software offers a wide range of graphical and statistical tools, including time-series analysis, classification, clustering, and linear and nonlinear modelling. It is also very extensible. Due to its adaptability as an effective language that bridges software development and data analysis, the statistical software R has gained popularity day by day.

Valedictory Session

The valedictory session started at the end of this course. It was co-ordinated by Dr. Anindya Basu, Organising Secretary. The valedictory session was adorned by Dr. Tania Chakravertty, Dean of Students´ Welfare and Jagadish Chandra Ghosh, Finance Officer who took the onus of certificate distribution too. The summary of the workshop was provided the Convener of the workshop and Head of the Department of Geography, Dr. Shovan Ghosh. Prof. Rahul Bhattacharya, the eminent resource person for the workshop appreciated the arrangement and lauded the smooth organization. The participants shared their experience about the workshop. Chaitali Roy, a Research Scholar at Diamond Harbour Women's University; Sudeb Pal, a Research Scholar of University of Calcutta spoke highly about the effectiveness of the workshop, expressed satisfaction about local hospitality workshop and suggested organizing more such workshops in the future. All participants were also very pleased with the transportation arrangement (pick-up and drop from city proper) as they could reach the remote venue smoothly.

Valedictory session

A participant sharing her experience about the workshop

A moment of certificate distribution

The vote of thanks was delivered by Dr. Kapil Ghosh, the Co-organising Secretary, who thanked the participants, resource persons, scholars and the teaching and non-teaching members of the university.

Dr. Kapil Ghosh, the Joint organising secretary

After the formal certificate distribution, the workshop ended on a positive note with hope for such future endeavours. Photography and partial video-recording of the event was done for documentation.

Feedbacks

Feedback on any event is the right way to judge the level of the success of the event and it is not an individual viewpoint but rather a holistic view of the total participants. All the registered participants provided their valuable feedback regarding the 5 days long workshop.

Group photo of the participants with Resource persons, faculty members of Dept. of Geography and Dean of students' welfare, DHWU

The responses of the workshop have been captured from the participants and categorized into 5 groups. (1= Least Satisfied; 2= Less Satisfied; 3 = Moderately Satisfied; 4 = Highly Satisfied; 5 = Very Highly Satisfied). 64.95% and 67.6% of participants were very highly satisfied with the overall workshop and specifically with the content of the workshop while 32.4% and 29.7% were highly satisfied and there is no participant from the least and less satisfied categories. 45.9% of participants were very highly satisfied with the hands-on-training of SPSS-AMOS while 59.5% were with R software. 2.7% of participants were less satisfied with the hands-on-training of SPSS-AMOS while there is no such participant who is less satisfied with the hands-on-training of R software. 51.4% of participants considered that the workshop had a very highly satisfying hands-on-training part while 32.4% considered it as highly satisfying. 67.6% of participants considered that the hospitality was very good. 94% opined that overall, the workshop was a fruitful, informative one and will help in their further research. Maximum participants said that they would recommend this type of workshop in near future to others.

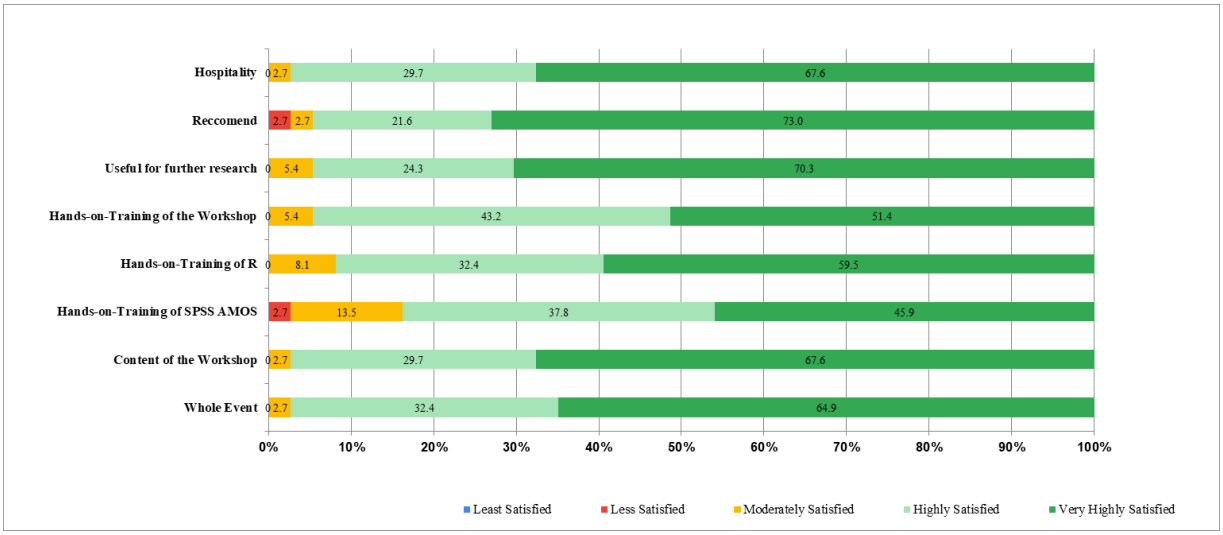


Fig 14: Level of satisfaction amongst the participants