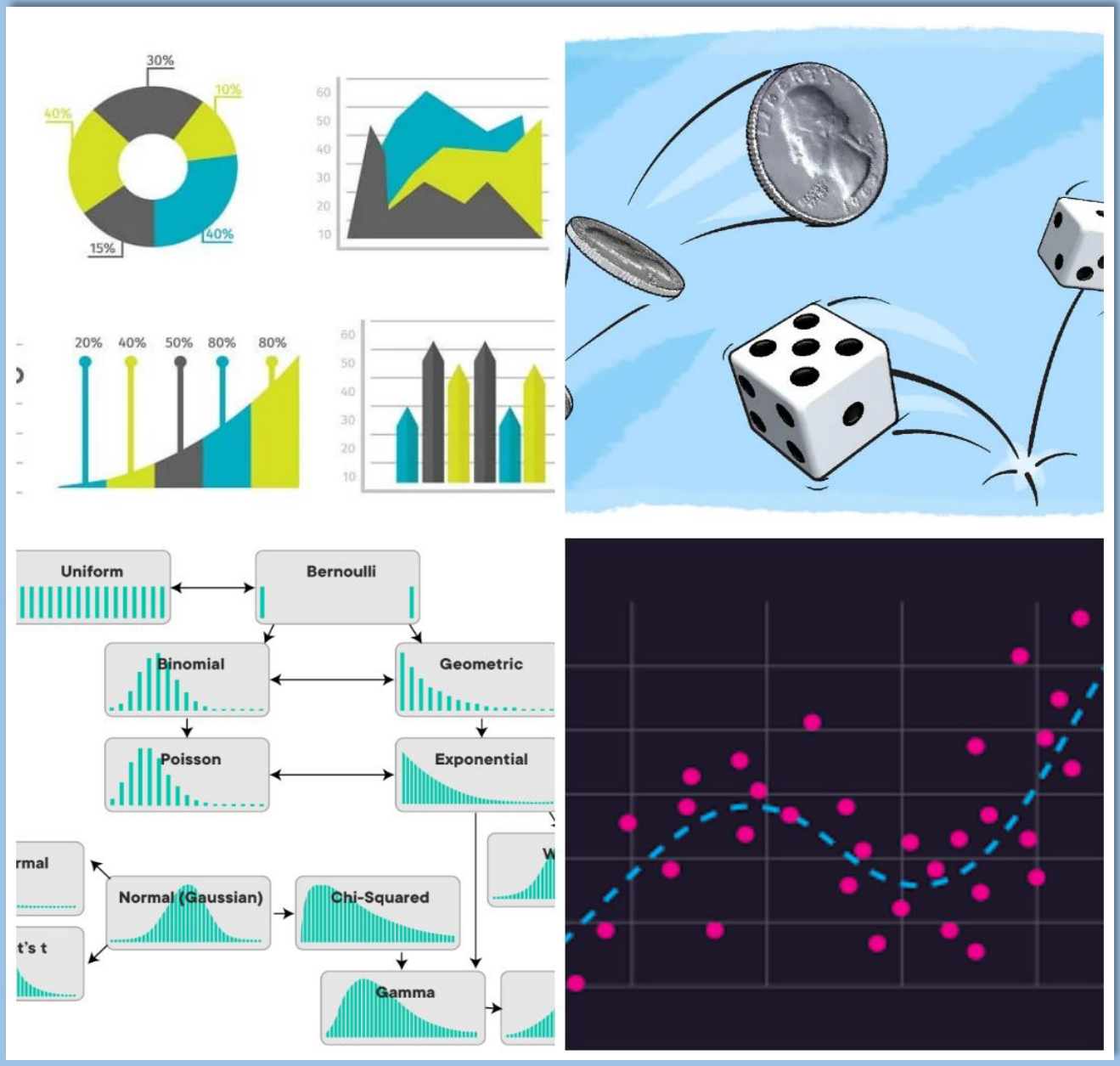
Divergence

Newsletter of Department of Statistics

Haldia Government College

June 2024



About the Department

The Department of Statistics has started its journey in 1988 and since then it is one of the college's leading departments. There are very few colleges in West Bengal offering statistics at the undergraduate level outside of Kolkata, and Haldia Government College is one of them. Currently, the Department of Statistics of the college has a good strength of teaching staff. The department has separate computer laboratories for the teachers and students. The department also has a seminar library consisting of an excellent range of text and reference books, which are accessible to the students. The faculties of the department are engaged in various UGC-sponsored Minor Research Projects along with their involvement in the teaching of all normal courses. The department also holds seminars or workshops at regular intervals. Presently, the head of the department is Dr. Shyamsundar Sahoo.

Statistics is an emerging scientific discipline that focuses on learning from data, which involves collection, presentation, analysis, and interpretation utilising inductive reasoning based on the Theory of Probability. It has wide applications in almost every field of science and technology, and consequently, students with statistics have job or career opportunities in several government and private sectors. The course structure at Vidyasagar University is very well-designed with components from the disciplines of Statistics, Mathematics, Economics and Computer Programming. After completing the undergraduate course, the students of this department get an opportunity to pursue a Master’s degree at several prestigious universities and institutions, including IIT-Kanpur, Calcutta University, Banaras Hindu University, Presidency University, Delhi University, Pondicherry University and others. The students of the department have established themselves in various academic and business sectors.

Only those students who pass the H.S. examination with Mathematics as a main subject are eligible to apply for the B.Sc. Honours in Statistics or B.Sc. Honours in Statistics with Research program in the college.

Present Faculty members:

- Dr. Shyamsundar Sahoo, Associate Professor & HoD, M.Sc., Ph.D.
- Sibsankar Karan, Assistant Professor, M.Sc.
- Tanmay Kumar Maity, Assistant Professor, M.Sc., M.Phil.
- Bijitesh Halder, Assistant Professor, M.Sc., M.A. (Population Science)

Present students of the department (2024-25)

Semester 2

1. Hirak Mahapatra
2. Kuntal Kumar Acharya
3. Lipika Das
4. Moumita Goswami
5. Mrinal Kanti Bej
6. Pravas Roy
7. Rajesh Singha
8. Sanchita Maji
9. Santanu Samanta
10. Santanu Samanta
11. Satarupa Das
12. Snehasis Bhunia
13. Soumitra Middya
14. Srabani Mondal
15. Subhasis Halder

Semester 4

1. Ananya Gupta
2. Anshuman Jana
3. Arpan Jana
4. Basanti Bera
5. Debabrata Sahoo
6. Debjyoti Ghorai
7. Mala Samanta
8. Pampa Patra
9. Sabyasachi Bag
10. Saini Pradhan
11. Shreyasi Maiti
12. Sourish Halder
13. Sudip Mandal
14. Sumana Jana
15. Supriya Midya

Semester 6

1. Sanjay Kumar Jana
2. Diparna Santra
3. Debjyoti Mukherjee
4. Nilanjan Manna
5. Trisha Ghosh
6. Tanushree Dey
7. Jayita Jana
8. Kartik Jana
9. Sougata Maity
10. Pritimoy Jana
11. Utsha Dutta

A Message from the Head of the department

Welcome to the first edition of our Department's Newsletter. It gives me immense pleasure to write about our department in this Newsletter. There are very few colleges in West Bengal outside of Kolkata that offer Statistics Honours and General at the undergraduate level, and Haldia Government College, affiliated with Vidyasagar University, has been providing this service to students since its establishment in 1988. I am pleased to announce that the Department of Statistics of this college continues to grow and is becoming more active. We currently have 4 permanent faculty members: 3 Assistant Professors and 1 Associate Professor, and 42 undergraduate students. The Department has separate computer laboratories for the students and teachers, equipped with the basic and modern computing facilities, as well as a seminar library with an excellent range of text and reference books. The Department has the facility to access the reputed international journals "SANKHYA" and "CSA Bulletin." The college also provides e-Journals and e-Books through INFLIBNET-NLIST. On a regular basis, the Department organizes academic activities for the betterment of the students, such as Seminar Lectures, Workshops and Webinars, with speakers from different prestigious institutions across India. Many of our faculty members received research grants from the University Grants Commission and are actively involved in their research work. The faculty members are engaged in delivering lectures in different national and international conferences/symposiums/workshops.

The course structure for Statistics Honours and General at Vidyasagar University is very well-designed, including key sections of Mathematics, Economics and Computer Programming (such as C, Statistical Package-R). The Department extends modern statistical courses such as Survival Analysis, Actuarial Statistics and Econometrics to the honours students as the Discipline Specific Elective (DSE) papers. The Department also assists the students by guiding them through the theory and methods of Statistics, setting up relevant questionnaire and Google Forms for data collection, and using the statistical package R or MS-Excel for computational resources to complete their independent research-oriented project work in their 6th semester course.

Students from various regions of West Bengal, as well as from beyond West Bengal, enroll in Statistics honours at our college. The students of this department graduate with a sound knowledge in not only the core subject Statistics, but also Mathematics, Computer Programming and Economics. The students of this department qualify national level different entrance examinations such as IIT-JAM, CUET, GATE examinations. Our students regularly participate in various statistical events at the state and national levels organized by the reputed academic institutions, and have received the prestigious award of the FIRST position in some events. I am very proud of our students for their outstanding achievements. The present course structure, combined with the department's hybrid teaching-learning process, also strengthens our students' ability to pursue their Master's degree at several prestigious universities and institutions, including IIT-Kanpur, Calcutta University, Banaras Hindu University, Presidency University, Delhi University, Pondichery University, Vivekananda University (Belur), Central University of Punjab and others. Finally, the department's students have established careers in various academic and corporate sectors.

I would like to express my heartfelt thanks to my departmental colleagues for their enormous contributions for the Department's enrichment. I also wish my loving students a great success in their life.

List of students who passed from our departments in the last 5 years

Session: 2018-19

1. Pradip Rana
2. Chandra Sekhar Maity
3. Pathajit Das
4. Sourav Bera
5. Swarnava Samanta
6. Debasish Maity
7. Meghnath Jana
8. Sayan Bhowmik
9. Subhasish Roy

Session: 2019-20

1. Souvik Bal
2. Reshma Khatun
3. Animesh Guchhait
4. Puspasourav Panda
5. Sakuntala Manna
6. Sayani Pradhan
7. Shibananda Jana
8. Somnath Karan
9. Sourav Pal
10. Subhabrata Adak
11. Supriya Sahoo
12. Suryakanta Ghanta

Session: 2020-21

1. Akash Maity
2. Animesh Mandal
3. Anubhab Maity
4. Biswadip Das
5. Debasish Manna
6. Romiyars Mondal
7. Sagnik Patra
8. Shreya Panda
9. Soumen Khatua
10. Arnab Manna
11. Vivekananda Mallik
12. Ankit Dhara

Session: 2021-22

1. Biswajit Mandal
2. Debojyoti Sahu
3. Krishanu Bakshi
4. Rudra Samanta
5. Samapti Das
6. Soumyadeep Rudra
7. Subhadip Ghosh
8. Sudin Jana
9. Tanmoy Manna
10. Tirtharaj Sasmal
11. Udit Narayan Sahoo
12. Upam Acharya
13. Sampreeti Sahoo

Session: 2022-23

1. Akash Das
2. Bikash Gayen
3. Kousik Ghara
4. Nilanjan Samanta
5. Pratyay Mondal
6. Raj Sekhar Das
7. Riddhiman Ghosh
8. Sayak Kanti Jana
9. Shatakshhe Das
10. Snehasis Haldar
11. Somnath Samanta
12. Sourav Jana
13. Sudipta Mondal
14. Swagata Karmakar

Value Added Courses organised by the department

Session: 2023-24 (Ongoing course)

On Multiple Linear Regression Modelling Using Statistical Software R

Course duration: 32 hours

Objectives of the course:

1. To discuss the various stages of regression modelling.
2. To assess the underlying assumptions of the multiple linear regression model, and to implement corrective measures if needed.
3. To discuss parameter estimation and testing in a k-variable linear equation.
4. To provide understanding of the geometry of the least squares.
5. To review linear algebra required for comprehending the inferential part of a multivariate setup.
6. To provide hands-on training for the use of multiple linear regression models with the statistical package R.

Session: 2022-23

1. ***Statistical Computing using Ms-Excel and Python***

Course duration: 38 hours

Aim of the course:

- To aware the participants about basic of MS-excel toolbar, functions, graphs and charts.
- To provide knowledge about some advance excel techniques such as macro and solver.
- To provide knowledge about handling a dataset and some statistical computing using MS-excel
- To discuss the fundamental programming using python
- To demonstrate about real life data handling using python
- To explain some advance techniques such as python library and graphs
- To demonstrate some data analysis technique using python

No of participants: 26

Course outcomes:

- The participants have gained knowledge about basic to advances features of MS-Excel. The hands-on session helps the students to incorporate their theoretical knowledge in day-to-day data handling as well as analysing and presenting complex feature of a dataset
- The python programming enables the students to understand basic structures of any programming language. The hands-on session helps the students to summarize a dataset using python programming.
- The knowledge of MS-excel and python will definitely help the students to cope with the present competitive scenario deals with big data analytics and machine learning.

2. ***Model Diagnostics for Multivariate Data***

Course duration: 32 hours

Aim of the course:

- To provide basic knowledge on various stages of regression modelling.
- To check the underlying assumptions of the linear regression model, and to take remedial measures if needed.
- To provide insight into the phenomenon of collinearity from a geometric perspective.
- To select a suitable subset of regressors on the basis of certain criteria.
- To identify the discordant and the influential observations.
- Real data analysis with hands-on training using the statistical package R.

No of participants: 17

Course outcomes:

Students, after completion of the course, would be able:

- To comprehend several types of real-world data and how to analyse some of them, with a focus on diagnostics for regression models;
- To build a working stochastic model to explain quantitatively the nature of association between the response and the regressors;
- To assess quantitatively the nature of the dependence of the response on a few regressors in the presence of other regressors;
- To predict the response for a new case using the observed set of regressor values;
- To understand the various stages of regression modelling, namely, development, fitting, and validation;
- To examine whether the model assumptions are being violated by using the basic tools for diagnosis, including formal statistical tests of hypotheses;
- To comprehend whether regressors have a collinear relationship;
- To choose a best subset of predictors
- To use the statistical tool R to do statistical analysis on actual data sets, with a focus on regression modelling.

Session: 2021-22

Financial Time Series Analysis by Artificial Neural Networks

Course duration: 34 hours

Aim of the course:

- To aware the participants about financial time series and their different features
- To provide knowledge about different financial time series models
- To provide knowledge about neural computing
- To discuss the possibility of capturing the complexity of financial time series using automated modelling such as neural networks
- Real life financial data handling using artificial neural networks

No of participants:14

Course outcomes:

- The participants have gained knowledge about different complex feature of financial time series. They also gain knowledge about the statistical modelling of financial time series. This will help them to understand the features of financial market.
- The knowledge about neural computing will help the participants to understand automated modelling. This type of automated financial forecasting can help them to understand financial market risk, investment, trading etc.

Project works done by the final semester students in last two years

Academic session: 2022-23

Sl.	Name of the Student	Title of the Project Work
1	Akash Das	Perception Of Student Mental Stress
2	Bikash Gayen	A Statistical Study on Fast Food Consumption Pattern and Its Effect On Health
3	Kousik Ghara	Analysis Of Data Oon Death by Road Accident in India: A Comparative Study
4	Nilanjan Samanta	The Competing Analysis of Stock Market Dynamic In Indian Perspective: Before, During And After Covid-19
5	Pratyay Mondal	Constructing An Index to Compare Rural Health System Condition Between Different States of India
6	Raj Sekhar Das	Statistical Analysis on Impact of Several Factor on Lung Cancer
7	Riddhiman Ghosh	A Study of Waiting Time to Judgement of Court Cases: A Survival Analysis Approach
8	Sayak Kanti Jana	Health Care : Heart Attack Possibility
9	Shatakshee Das	Brain Drain
10	Snehasis Haldar	Predicting The Hotel Booking Cancelation Behaviour of Two Europian Hotel
11	Somnath Samanta	Determining The Important Factors for Choosing Rare Subject as Career In UG
12	Sourav Jana	Analysis Of Survivorship of Cricketer's in Home Ground and Overseas Ground
13	Sudipta Mondal	The Impact of Education on Economic Growth: The Case Of India
14	Swagata Karmakar	A Comparison Between Two Methods of Estimating Population Proportion

Academic session: 2021-22

Sl.	Name of the student	Title of the Project Work
1	Biswajit Mandal	The project work on effect of some factors on growth of vennamei prawn.
2	Debjyoti Sahu	Analysis of survivorship of cricketer's in home and overseas ground.

Sl.	Name of the student	Title of the Project Work
3	Krishanu Bakshi	An analysis of petroleum price of several Indian metro cities (2015-2022).
4	Rudra Samanta	Comparison of different approaches for Fisher- Behrens problem.
5	Samapti Das	Statistical analysis on the use of contraceptive method.
6	Soumydeep Rudra	Inference on probability of success under binomial and geometric sampling.
7	Subhadip Ghosh	Study the importance of four bus stops using the Google page ranking method.
8	Sudin Jana	Prediction of GDP of India using different factor: A cross-correlation approach.
9	Tanmoy Manna	A study on effects of some factors in paddy production.
10	Tirtharaj Sasmal	Analysis on satisfaction, after purchasing the product for different ways of shopping.
11	Udit Narayan Sahoo	A project work on survey regarding thought about causes of different diseases and uses of social media.
12	Upam Acharya	Forecasting the rainfall of different regions of India in monsoon: A stochastic modelling approach.
13	Sampriti Sahoo	Statistical analysis on crimes of juveniles and senior citizens from 2016-2020.

Some glimpses of selected projects done by our students:

A Study of Waiting Time to Judgement of Court Cases: A Survival Analysis Approach

-Riddhiman Ghosh, 2023

The path to justice in the context of the Indian legal system is oftentimes a lengthy one, often stretching for a decade or two. The time to disposal of a case depends on many factors, such as whether it is a criminal or civil case, in which court the case is being filed, the judge or the bench, availability of evidence to continue hearing and so on. All these factors contribute to the waiting time to judgement of a particular case. If one looks at the status of pending cases in but High Courts, there will be some homogeneity regarding the material conditions as compared to that of the lower courts or when compared to the Supreme Court.

The objective of this study was to estimate the survival function for the data on waiting time to disposal of court cases filed in Indian High Courts. Inferences about the results obtained by the two approaches mentioned will be drawn, leading to a comparison of the two approaches and their validity for the given problem. Furthermore, the mean and median waiting time to disposal of a case have been estimated, rendering some insight on the efficiency and shortcomings on the working of the legal institutions of the country.

Source of Data: The data was obtained from National Judiciary Data Grid's website.

Results and Discussions:

The time to judgement of court cases concerning the data, is therefore found to follow the LogLogistic distribution.

The mean time to survival for the waiting time to judgement has been found to be 4.39 years.

The median time to judgement was found to be about 3 years in both the parametric and nonparametric setup. The standard deviation of the Loglogistic distribution was found to be 3.77 years.

Thus, we find a 95% confidence interval for the time to judgement random variable as (2.88,15.99), using large sample normal approximation and thus the percentage points of the standard normal distribution. Hence, it could be inferred that the time to judgement of court cases, both civil and criminal fall in the interval (2.88,15.99) 95% of the time.

Constructing An Index to Compare Rural Health System Condition Between Different States of India

- *Pratyay Mondal, 2023*

India's rural health system is facing significant challenges that are affecting the health and well-being of millions of people living in rural areas. Despite some progress in recent years, there are still significant disparities in healthcare access and quality between urban and rural areas. This is due to a lack of adequate infrastructure, shortage of healthcare professionals, and limited access to essential medicines.

Real-life examples illustrate the challenges faced by the rural health system in India. For instance, many people living in rural areas do not have access to basic healthcare services, and maternal and infant mortality rates are significantly higher than in urban areas. The prevalence of chronic diseases such as diabetes and hypertension is also rising in rural areas, but access to specialist care and medication remains limited.

To better understand the condition of the rural health system in India, we have constructed an index that measures key indicators such as healthcare infrastructure, availability of essential stuffs, and quality of healthcare services in a particular state. This index is unique because there is no such well-known index previously constructed, and we hope that it will be helpful in visualizing the condition of rural health system in India. By identifying areas of improvement, policymakers can develop targeted interventions to improve rural health outcomes and ensure that all people have access to quality healthcare services, regardless of where they live

The objectives of the study were:

1. To assess the current state of rural health systems in India using our constructed index. 2. To identify areas of improvement in rural health systems and develop recommendations for policymakers. 3. To raise awareness about the challenges faced by the rural health system in India and the need for targeted interventions to improve health outcomes. 4. To contribute to the development of evidence-based policies and strategies to improve rural health outcomes in India. 5. To

compare rural health system condition between different state and UTs of India
6. To compare it with an existing reputed index.

Data Collection:

The entire data is collected from Annual Rural Health Statistics report (2020-21) published by Ministry of Health and Family Welfare.

The factors which were considered for this work:

1. Average Population Covered by a Sub Centre 2. Average Population Covered by a Primary Health Centre 3. Average Population Covered by a Community Health Centre 4. Average Rural Area Covered by a Sub Centre 5. Average Rural Area Covered by a Primary Health Centre 6. Average Rural Area Covered by a Community Health Centre 7. Average Radial Distance Covered by a Sub Centre 8. Average Radial Distance Covered by a Primary Health Centre 9. Average Radial Distance Covered by a Community Health Centre 10. Average Number of Villages Covered by a Sub Centre 11. Average Number of Villages Covered by a Primary Health Centre 12. Average Number of Villages Covered by a Community Health Centre 13. Average Rural Population Covered by a Health Worker 14. Health Workers (Female) / ANM at Rural Areas (In Position/ Required Ratio) 15. Doctors at Primary Health Centres in Rural Areas (In Position/Required Ratio) 16. Total Specialists at CHCs in Rural Area (In Position/Required Ratio) 17. Radiographers at CHCs in Rural Area (In Position/Required Ratio) 18. Pharmacists at PHCs and CHCs in Rural Area (In Position/Required Ratio) 19. Laboratory Technicians at PHCs and CHCs in Rural Area (In Position/Required Ratio) 20. Nursing Staff at PHCs and CHCs in Rural Areas (In Position/Required Ratio) 21. Statewise Rural Infant Mortality Ratio.

Conclusions:

In this study, a rural health index was constructed to evaluate the health infrastructure and outcomes in Indian states. Despite acknowledging the scope for improvements in our index, it is notable that the results obtained align closely with the existing NITI Aayog health index, which primarily focuses on overall health performance without specifically emphasizing rural areas. This similarity suggests that the constructed rural health index is moving in the right direction.

The fact that our rural health index and the NITI Aayog health index demonstrate similar findings indicates that the existing index effectively captures the overall health performance of states, encompassing both rural and urban areas. However, the development of a dedicated rural health index is crucial to shed light on the specific challenges and disparities faced by rural populations in accessing healthcare.

While there are opportunities for improvement in our index, such as refining indicators, data sources, and methodology, the alignment with the NITI Aayog health index signifies that our efforts have been in the correct direction. By focusing on rural areas, our index

provides a more nuanced understanding of the state of rural healthcare in India, emphasizing the need for targeted interventions and resource allocation.

The findings of this study hold significant implications for policymakers, healthcare organizations, and other stakeholders. The constructed rural health index can serve as a valuable tool in identifying areas for improvement and guiding the development of tailored interventions to address the unique healthcare needs of rural communities. By leveraging the insights provided by the rural health index, policymakers can work towards bridging gaps in healthcare infrastructure, enhancing primary healthcare services, and reducing health disparities in rural India.

Future research should aim to further refine the index by incorporating additional relevant indicators, improving data quality and availability, and engaging in ongoing collaboration with experts and stakeholders. Longitudinal studies can also be conducted to monitor the progress of states over time and assess the impact of interventions aimed at improving rural healthcare.

In conclusion, the construction of a rural health index represents a vital step towards understanding and addressing the healthcare challenges faced by rural populations in India. While improvements are possible, the similarity between our index and the NITI Aayog health index validates our approach and underscores the importance of prioritizing rural healthcare to build a more inclusive and equitable healthcare system for all

Study the importance of four bus stops using the Google page ranking method

- *Subhadip Ghosh, 2022*

Page rank (PR) is an algorithm used by Google Search to rank websites in their search engine results. Page rank was named after Larry Page, one of the founders of Google. Page Rank is a way of measuring the importance of a website pages. According to Google, page rank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. This is done using the concepts of the eigen vector and eigen value.
Here the objective was to rank four bus stops (Nimtouri, Nandakumar, Brajalalchak, Ranichak) in terms of their importance by the method of "GOOGLE PAGE RANKING".

From the eigen vectors it was found that the most important bus stop is NANDAKUMAR and nd the less important bus stop is BRAJALALCHAK.

Events organized by the Department in recent past:

- State-sponsored workshop on "**Applications and Career Aspects of Statistics**" held on 7th October, 2013.

  The objective of the workshop is to provide an overall idea on the subject 'STATISTICS' with its applications in several important fields and its career opportunities, to the 10+2 students with Mathematics/Statistics as major subject in the neighbouring Higher Secondary schools. The lectures were given by 1. Mr. Ujjwal Kanti Manna, Assistant Adviser (AGM),, Department of Statistics and Information Management, Reserve Bank of India, Mumbai. 2. Mr. Kaushik Maity, Associate Manager – Biostatistics, Clinical Programming & Data Management, MMS Holdings, Bangalore. 3. Dr. Abhijit Mandal, Visiting Scientist, Indian Statistical Institute, Kolkata and 4. Mr. Sujan Chandra of this department.

- One-Day seminar on "**An Overview of Clinical Research: Career Opportunities in Biostatistics & SAS Programming**" held on 31st October, 2014.

  In this seminar the participants were given an idea about the field of clinical research leading to career opportunities in Biostatistics & Clinical SAS programming. The seminar would be beneficial not only for the Honours students of Statistics & Mathematics but also for other students with Statistics or Mathematics as one of their general subjects. The lectures were given by 1. Mr. Kaushik Maity, Associate Manager – Biostatistics, Clinical Programming & Data Management, MMS Holdings, Bangalore and 2. Mr. Shyamsundar Sahoo, Assistant Professor (formerly), Dept. of Statistics, Haldia Govt. College.

- Seminar lecture on "**Accelerating Science and Skilling for the Next-Generation of Jobs: A Global Exposure**" held on 1st March, 2017.

  In this lecture Dr. Sushanta Tiwari, Ph.D. (Pune University) an alumnus of this department briefly pointed out the scope and career options of Statistics. He also discussed about different scopes of interdisciplinary work in the field of data science.

- One-Day Seminar on "**Probability and Statistics**" in collaboration with the **Calcutta Statistical Association** on 5th March, 2018.

In this seminar Prof. Bikas K Sinha of Indian Statistical Institute give his fascinating lectures on probability and statistics. He also demonstrated some real-life paradox in statistics. Students enjoyed the session very much.

- Science Festival "**PARADOX-2019**" on 27th November, 2019.

  The main objective of this fest was to aware the participants regarding different popular topics from Physics, Chemistry, Mathematics and Statistics. The program was held in two sessions. In the first session there are four presentations by the students from Physics, Chemistry, Mathematics and Statistics department. In this session there is also two presentations, one by a research scholar from Department of Statistics, University of Calcutta and another one from Dr. Susanta Tewari, Amity University. In 2nd session there is an Inter-Departmental Quiz competition organized by the students of Department of Statistics.

- Webinar on "**An Introduction to Probability and Statistics**" on 3rd August, 2020.

  In this webinar Professor Arnab Chakraborty, Indian Statistical Institute has delivered this lecture on some basic and interesting aspects of Probability and Statistics. The program was steamed on YouTube platform.

- Webinar on "**The Theory of Percolation**" on 14th August, 2020.

  In this webinar Prof. Kumarjit Saha has delivered this lecture on an emerging field of probability percolation theory and possibility to apply this theory to some real life examples.

- Webinar on "**Application of Statistics in Marketing Industry**" on 30th November, 2021.

  In this webinar Mr. Bappaditya Mondal has delivered his lecture on application of Marketing Mix Model in Marketing Industry. Before discussing about Marketing Mix Model Mr. Mondal briefly discussed about predictive modelling and mix modelling. He also discussed the utility of using mix modelling present scenario. The session started with some introductory speech of Dr. Dipankar Sadhukhan, Coordinator, IQAC, Haldia Government College. After the speech of Mr. Mondal there was an interactive session with the departmental students.

Some other programs:

Seminario de Statistica, 25/07/2021 & 01/08/2021, 3.30 pm – 5 pm

As in most of the schools, statistics is not offered as a subject, students are unaware of the subject, its importance in modern day and the diverse career opportunities it holds. This two-day programme was the brainchild of "StatSutram", an organization of the alumni of department of Statistics, Haldia Government College in order to raise awareness among the school students about this subject and its career opportunities and to encourage them to pursue higher studies in this field. In the two-day programme, the speakers and organizers successfully discussed the scope and prospects of studying Statistics honours as well as the intricacies involved in this subject.

Some glimpses of programs and activities

State-sponsored workshop on "**Applications and Career Aspects of Statistics**", 7th October, 2013.

One-Day Seminar on "**Probability and Statistics**" in collaboration with the **Calcutta Statistical Association**, 5th March, 2018.

*Science Festival "**PARADOX-2019**", 27th November, 2019*

Teachers' Day celebration at the department

Departmental Picnic 2022 at Tajpur

List of students progressing to higher education in the last 5 years (2019-23):

Name of student	Year of graduation	Name of institution joined	Name of program admitted to
Chandra Sekhar Maity	2019	Sardar Patel University, Gujarat	M.Sc. In Applied Statistics
Pathajit Das	2019	Karnatak University, Dharwad	M.Sc. In Statistics
Sourav Bera	2019	Karnatak University, Dharwad	M.Sc. In Statistics
Swarnava Samanta	2019	Karnatak University, Dharwad	M.Sc. In Statistics
Debasish Maity	2019	Sardar Patel University, Gujarat	M.Sc. In Applied Statistics
Meghnath Jana	2019	Aliah University, Newtown, Kolkata	M.Sc. In Statistics
Sayan Bhowmik	2019	IIT Kanpur	M.Sc. In Statistics
Subhasish Roy	2019	Ramakrishna Mission Vivekananda Educational and Research Institute, Belur, Howrah.	M.Sc. In Big Data Analytics
Reshma Khatun	2020	Aliah University, Newtown, Kolkata	M.Sc. In Statistics
Animesh Guchhait	2020	Central University of South Bihar	M.Sc. In Statistics
Puspasourav Panda	2020	Central University of South Bihar	M.Sc. In Statistics
Sakuntala Manna	2020	Aliah University, Newtown, Kolkata	M.Sc. In Statistics
Sayani Pradhan	2020	Aliah University, Newtown, Kolkata	M.Sc. In Statistics
Shibananda Jana	2020	Aliah University, Newtown, Kolkata	M.Sc. In Statistics
Somnath Karan	2020	Banaras Hindu University	M.Sc. In Statistics
Sourav Pal	2020	IIT Kanpur	M.Sc. In Statistics
Subhabrata Adak	2020	Central University of South Bihar	M.Sc. In Statistics
Supriya Sahoo	2020	Central University of Punjab	M.Sc. In Statistics

Name of student	**Year of graduation**	**Name of institution joined**	**Name of program admitted to**
Suryakanta Ghanta	2020	Central University of Punjab	M.Sc. In Statistics
Akash Maity	2021	Banaras Hindu University	M.Sc. In Statistics and Computing
Anubhab Maity	2021	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Arnab Manna	2021	Pondicherry University	M.Sc. In Statistics
Biswadip Das	2021	Aliah University, Newtown, Kolkata	M.Sc. In Statistics
Debasish Manna	2021	Aliah University, Newtown, Kolkata	M.Sc. In Statistics
Romiyars Mondal	2021	Banaras Hindu University	M.Sc. In Statistics and Computing
Sagnik Patra	2021	The International Film & Television School, Paris (Master of Fine Arts in Filmmaking Program)	Master of Fine Arts in Filmmaking Program
Shreya Panda	2021	Vellore Institute of Technology	M.Sc. In Data Science
Soumen Khatua	2021	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Vivekananda Mallik	2021	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Biswajit Mandal	2022	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Debojyoti Sahu	2022	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Krishanu Bakshi	2022	Maulana Abul Kalam Azad University of	M.Sc. In Applied Statistics and Analytics

Name of student	Year of graduation	Name of institution joined	Name of program admitted to
		Technology, West Bengal (MAKAUT)	
Rudra Samanta	2022	Presidency University	M.Sc. in Applied Statistics
Samapti Das	2022	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Soumyadeep Rudra	2022	University of Delhi	M.Sc. In Statistics
Subhadip Ghosh	2022	Symbiosis Statistical Institute, Pune	M.Sc. in Statistics
Sudin Jana	2022	University of Delhi	M.Sc. In Statistics
Tanmoy Manna	2022	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Udit Narayan Sahoo	2022	Maulana Abul Kalam Azad University of Technology, West Bengal (MAKAUT)	M.Sc. In Applied Statistics and Analytics
Akash Das	2023	Banaras Hindu University	M.Sc. In Statistics and Computing
Kousik Ghara	2023	University of Calcutta	M.Sc. In Statistics
Nilanjan Samanta	2023	Banaras Hindu University	M.Sc. In Statistics and Computing
Pratyay Mondal	2023	University of Delhi	M.Sc. In Statistics
Raj Sekhar Das	2023	University of Calcutta	M.Sc. In Statistics
Riddhiman Ghosh	2023	University of Calcutta	M.Sc. In Statistics
Snehasis Haldar	2023	University of Delhi	M.Sc. In Statistics
Somnath Samanta	2023	University of Calcutta	M.Sc. In Statistics
Sourav Jana	2023	Pondicherry University	M.Sc. In Statistics
Sudipta Mondal	2023	Pondicherry university	M.Sc. In Statistics

Name of student	Year of graduation	Name of institution joined	Name of program admitted to
Swagata Karmakar	2023	University of Delhi	M.Sc. In Statistics

Placement of outgoing students for the last 5 years:

Name of student who has been placed	Year of graduation	Placement details
Chandra Sekhar Maity	2019	Associate Data scientist at Sibia Analytics and Consulting Services Pvt Ltd
Pathajit Das	2019	Associate Data scientist at Sibia Analytics and Consulting Services Pvt Ltd
Sourav Bera	2019	Guest faculty at Department of Statistics, Bangalore University
Swarnava Samanta	2019	Data Analyst at CirrusDB LLC
Debasish Maity	2019	Associate Data scientist at Sibia Analytics and Consulting Services Pvt Ltd
Subhasish Roy	2019	Junior Data Scientist at Aadhar Housing Finance Limited
Sayan Bhowmik	2019	Pursuing Ph.D. at IIT Kanpur
Somnath Karan	2020	Associate Biostatistician at INFERENCE CLINICAL RESEARCH SERVICES PRIVATE LIMITED
Sourav Pal	2020	Data Scientist at Involead Services Pvt. Ltd.
Subhabrata Adak	2020	Associate Data Analyst at Involead Services Pvt. Ltd.
Akash Maity	2021	Analyst at Ipsos Research Pvt. Ltd., Mumbai
Arnab Manna	2021	Intern at Arogya AI Innovations Pvt. Ltd.
Shreya Panda	2021	Programmer Analyst Trainee at Cognizant
Soumen Khatua	2021	Analyst at Ipsos Research Pvt. Ltd., Mumbai

ON THE DURATION OF POSTPARTUM AMENORRHOEA

Dr. Shyamsundar Sahoo
Head and Associate Professor
Department of Statistics

A closed birth interval consists of three key components: (i) the waiting period until conception, (ii) gestation, and (iii) the duration of postpartum amenorrhoea (PPA). Couples can manage the waiting time to conceive with adequate contraception, although gestation is uniformly constant in duration. However, the length of PPA varies due to its complicated nature. The duration of PPA is the time between the end of the pregnancy due to a live birth, stillbirth, or late term abortion and the first ovulation after the pregnancy. In other words, it is the postpartum anovulatory period. Because ovulation is difficult to detect, the onset of the first menstruation after delivery is considered as the end of PPA, and the timing of the first menstruation after delivery is known as the duration of PPA.

For many years, the duration of PPA has been a major source of concern in the general public because it tends to increase the inter-live birth interval. It has a direct effect on natural fertility by lengthening the period of conception, and women are expected to be in a safe phase for probable conception. Many key variables, referred to as intermediate variables, have a direct impact on human fertility. In various studies, PPA has been identified as a key intermediate variable among several proximate factors of fertility. As a result, in communities where contraception is not widely used, the duration of PPA has a significant impact on fertility reduction by increasing the inter-live birth interval. Other associated variables, such as socio-economic, demographic, and biological characteristics, influence fertility via intermediate variables. These variables are referred to as explanatory variables. For example, the PPA period directly influences natural fertility, and nursing has a significant impact on the duration of PPA.

In research investigations, the length of PPA is often used as a response variable. It is determined by several socio-economic and demographic factors. The duration of postpartum amenorrhoea is thought to be influenced by socio-economic variables such as place of residence, caste, religion, mother's education level, occupation, and family income, as well as demographic variables such as duration of breastfeeding, frequency of nursing, mother's age, sex of the previous child, parity, child living status, and use of contraceptive devices.

To investigate the nature of the duration of PPA and to examine the relative influences of the socio-economic and demographic variables on it, the researchers first conduct a survey over the target population. The survey is normally carried out among ever-married women who have given birth at least one, using adequate sampling techniques. Data are collected from the women respondents regarding the response variable and a set of explanatory variables. In the survey, if women report the end of amenorrhea, their durations of PPA, which are the time period between the end of conception and the return of the first menstrual period, provide complete information and are considered as uncensored observations. On the other hand, women experiencing amenorrhoea at the end of the study or on the date of interview provide incomplete information about the duration of PPA and are treated as censored cases, with their event history of interest being the time from the termination of conception to the end of the study or the

date of interview. As a result, one will get the censored amenorrhoea data, which can then be analysed using the statistical techniques of survival analysis.

Univariate statistical analysis can be used to understand the pattern of postpartum amenorrhoea for women with specific background characteristics. We can get estimates of the median durations of PPA for different levels or categories of the selected background characteristics by estimating the survival functions for each category using the Kaplan-Meier method, and we can compare them to see how the different levels of the characteristics affect PPA. The survival function refers to as the probability of not resuming the menstruation. It can also be called as women's amenorrhoeic probability. The log-rank test can also be used to examine the survival distributions of different groups under investigation, i.e., the distributions of the PPA period for the various categories of the given background socioeconomic or demographic factors.

Almost all research has shown that the duration of breastfeeding has a significant impact on the duration of PPA. Women who nurse their children for a shorter period of time resume menstruation more quickly, and thus the duration of breastfeeding has an increasing effect on the PPA period. Furthermore, parity, education, and mother's residence all have a substantial impact on the duration of PPA. Increased parity has been shown to have a beneficial influence on PPA length, whereas education level has an inverse effect. Women with higher education levels are more likely to offer their infants shorter periods of nursing, lowering the duration of PPA. Furthermore, women in rural areas tend to have a longer duration of PPA than those in cities. Numerous studies have also shown that women with a higher BMI restart menstruation faster than those with a lower BMI. That instance, undernourished nursing mothers will have longer periods of PPA than well-nourished mothers. Furthermore, the frequency of nursing has a significant effect on the amenorrhoea period. Other factors may also be associated with the duration of PPA. Based on the amenorrhoea data from the sampled women in a given community, a quantitative conclusion can be drawn about the possible influences of several factors on the duration of the PPA.

From a statistical point of view, variability in the length of the amenorrhoea period or, equivalently, the resumption period of menses across groups as well as within a group relating to the different characteristics of the mother and her child occurs due to the influence of other socio-economic and demographic factors. These factors may be related with mother's certain biological characteristics. For example, prolonged lactation supresses the secretion of certain types of hormones in mother's body, thereby extending the postpartum anovulatory period. Multivariate analysis can explain a portion of the variability by including a number of relevant characteristics into the model.

When studying postpartum amenorrhoea, the mother's breastfeeding duration has been found to be the most relevant predictor. Breastfeeding is also linked to other explanatory variables due to the diversity of socioeconomic culture. As a result, multivariate regression modelling may be useful in developing a working model to explain quantitatively the nature of the relationship between PPA duration and breastfeeding duration and other predictors, as well as to assess quantitatively the nature of the dependence of PPA duration on a few explanatory variables of interest in the presence of the other explanatory variables. This method can help us understand the partial effect of breastfeeding length on PPA duration while adjusting for other variables. Univariate analysis, on the other hand, analyses the influence of each explanatory variable on PPA duration without taking into account the effects of other relevant

variables. Before beginning regression analysis on data, it is obvious that one needs be familiar with the various processes of regression modelling. Otherwise, it will lead us misleading results, thereby misleading interpretations.

When performing a multivariate regression analysis on the PPA data, there are some difficulties in using the traditional statistical methods such as the multiple linear regression model, ANOVA, etc. because of the censorship and non-normality nature of the PPA duration data. One can then undertake the most popular and immensely useful Cox's proportional hazards (PH) regression model to determine the impact of breastfeeding, as well as other demographic variables and socio-economic variables, on the duration of PPA. Since a longer duration of PPA correlates to a smaller risk of resuming menstruation, and a shorter duration of PPA is associated with a higher risk of returning to menstruation, the risk, which is represented by the hazard rate of change, can be used to characterise the survival distribution. The Cox's PH regression model uses the hazard rate function or cumulative hazard rate function to connect the distribution of PPA duration to explanatory variables.

The hazard rate or the risk of resumption of menstruation of a mother with covariate profile $\boldsymbol{z}$ according to the Cox's PH regression model is given by

$$\lambda(t:\boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{z}),$$

where $\lambda_0(\cdot)$ is an unspecified nonnegative function of time t (but not on $\boldsymbol{z}$) known as the baseline hazard function, and $\boldsymbol{\beta}$ is a vector of regression parameters measuring the effects of the covariates. If a mother's covariate profile is split as $\mathbf{z} = (x, \boldsymbol{y})'$ and x is a scalar covariate, then the hazard rate of the mother is written as

$$\lambda(t:x,\boldsymbol{y}) = \lambda_0(t)\exp(\beta_1 x + \boldsymbol{\beta}_2'\boldsymbol{y}).$$

Consider a simple case in which x is an indicator variable taking value 0 for the reference group (say, a mother living in a rural area) and 1 for the target group (in an urban area). The hazard ratio of the two groups with other covariate profiles remain unchanged,

$$\frac{\lambda(t:x=1,\boldsymbol{y})}{\lambda(t:x=0,\boldsymbol{y})} = e^{\beta_1},$$

is called the relative risk of the target group compared to the reference group. A relative risk value greater than unity ($\beta_1 > 0$) indicates a higher risk of resumption of menstruation for mothers living in urban areas compared to rural areas, while a relative risk value less than unity ($\beta_1 < 0$) indicates a lower risk of return to menstruation for mothers living in urban areas compared to rural areas. When β_1 equals to zero, there is no difference in the risk of menstrual resumption between rural and urban areas. The Cox's proportional hazard model is used to assess the impact of each category of each explanatory variable on the hazard function while controlling for the effects of other factors. The popularity of this model stems from the existence of a simple semi-parametric estimating approach that may be utilised when the form of the baseline hazard function is unknown.

Because prolonged postpartum amenorrhoea has several benefits, including reducing mother's fertility, and improving mother's health as well as her baby's health, it is important to take appropriate initiatives to improve the other factors associated with prolonged breast feeding so that eligible mothers can extend the duration of postpartum amenorrhoea.

Forecasting Monsoon Rainfall in India

Tanmay Kumar Maity, Assistant Professor, Dept. of Statistics, Haldia Govt. College

Indian economy is still considered mostly based on agriculture. The production of a large part of agricultural crops is dependent on the amount of monsoon rains. Good monsoon always means a good harvest and a weak or bad monsoon is considered as a huge effect on India's economy and results in a big loss in the country GDP levels. Though, southwest monsoon rainfall over the country as a whole is more or less stable, even a small fluctuation in the seasonal rainfall can have devastating impacts on agricultural sector. Therefore, long-range forecasting (LRF) of southwest monsoon rainfall is a high priority in India. An accurate forecast of monsoon performance is very useful for better macro level planning of finance, power and water resources.

The India Meteorological Department (IMD) has been issuing long-range forecasts (LRF) for the southwest monsoon rainfall over India (ISMR) for more than 100 years based on statistical methods.

Background

After the great countrywide drought and famine of 1876 - 78, the Government of India called Sir H. F. Blanford, the first Chief Reporter of India Meteorological Department (IMD) for estimating the prospective rains. Blanford issued tentative forecasts from 1882 to 1885 using the indicators provided by the snowfall in Himalayas. In 1885, it was decided that a monsoon rainfall forecast would be issued annually as a matter of routine. The first of the regular series of forecasts was given on the 4th June 1886. This is continuing practically till date with modifying format, content and methodology. In 1892, long range forecast (LRF) for the rainfall for the second half of the monsoon season (August-September) was also started. After Blandford, Sir John Eliot as the Head of India Meteorological Department (IMD) applied subjective methods such as analogue and curve parallels for the LRF of ISMR in 1895. After that Sir Gilbert Walker, Director General of IMD started the forecasts based on correlation and regression techniques for preparing long range forecasts. From 1886, the monsoon forecasts were issued for the entire India and Burma. In 1988, India Meteorological Department introduced the 16 parameters power regression and parametric models and started issuing forecasts for the southwest monsoon rainfall over the country as a whole. Using the power regression model, quantitative forecasts were prepared and using the parametric model, qualitative forecasts (whether normal/excess or deficient) were issued. After the failure of forecast in 2002, IMD introduced a new two stage forecast strategy in 2003. During the period 2003-2006, the first stage quantitative and 5 category probabilistic forecast for the season rainfall over the country as a whole were issued using 8 parameter power regression (PR) model and Linear Discriminant Analysis (LDA) model respectively.

Present scenario

At present, the forecast for the South-West monsoon rainfall is issued in two stages. The first stage forecast for the seasonal (June to September) rainfall over the country as a whole is issued in April and the update of the April forecast is issued in June. For issuing the forecast for the seasonal rainfall over the country, a new statistical forecasting system based on the ensemble technique is introduced in 2007. The 8 predictors considered for the new ensemble forecast system are given in *Table 1*. For the April forecast (*Set I*), the first 5 predictors given in this table are used. For the updated forecast in June (*Set II*), 6 predictors that include 3 predictors (first 3 predictors

in this table) are used. For developing the models, two different statistical techniques namely, Multiple Regression (MR) and Projection Pursuit Regression (PPR) were considered. It is observed that Ensemble methods based on Multiple Regression outperforms nonlinear models such as Projection Pursuit Regression (PPR). The model error of the April forecast system is 5% and for the June forecast system, it is 4%.

Ensemble multiple linear regression (EMR) model

For the p predictors $x_1, x_2, \ldots, x_p$ multiple regression (MR) model is given as follows:

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_px_p$$

where y is the response variable and a_i' s are model parameters. If there are p predictors, it is possible to build $(2^p - 1)$ MR models relating the predictors and response considering all possible combinations of the predictors. In the ensemble method, instead of relying

Sl. No.	Predictor (Period)	Used for the forecasts in
1	North Atlantic Sea Surface Temperature (December + January)	April and June
2	Equatorial SE Indian Ocean Sea Surface Temperature (February + March)	April and June
3	East Asia Mean Sea Level Pressure (February + March)	April and June
4	NW Europe Land Surface Air Temperatures (January)	April
5	Equatorial Pacific Warm Water Volume (February + March)	April
6	Central Pacific (Nino 3.4) Sea Surface Temperature Tendency (MAM-DJF)	June
7	North Atlantic Mean Sea Level Pressure (May)	June
8	North Central Pacific Wind at 1.5 Km above sea level (May)	June

Table 1: Details of the predictors used in the forecast

on a single model, all possible models based on all the combination of predictors are considered. For April (June) forecast with 5 (6) predictors, 31 (63) different models were constructed. To find out the optimal length of training period, the prediction performance of each of the all possible models during a fixed common period 1981–2004 was examined. For this, a sliding window training period technique for all the

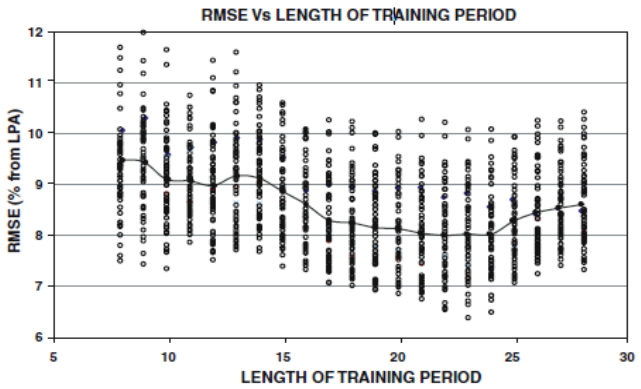


Fig. 1: RMSE vs. Length of Training period

possible lengths of training period has been used. Root mean square error (RMSE) of the predictions for the period 1981–2004 is used as the measures of prediction performance. Fig. 1 shows the diagram of RMSE of all the possible models (63 models) for April forecast (SET-I) plotted against different model training period lengths (from 8 to 28 years). The solid line shows the mean value of the RMSE obtained by taking average across all the 63 models. It is observed that the RMSE of the models decreases with increase in the length of training period and reaches the minimum value around 23 years and then again increases with increase in the length of training period. A similar analysis with all the possible MR models derived from the June forecast (SET-II) also showed that the optimal length of training period is around 23 years.

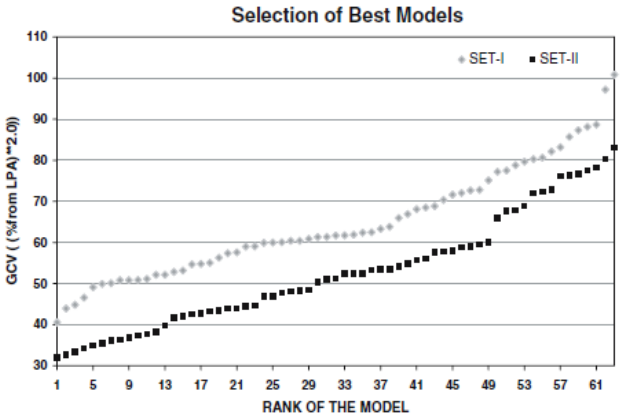


Fig. 2: Selection of best models using GCV criteria

Now, the best models for the ensemble average were obtained in two steps. In the first step, models were ranked based on generalized cross-validation (GCV) function given by: $GCV = \sum_1^n (y_i - \hat{y}_i)^2 / n(1 - \frac{p}{n})^2$, where $\hat{y}_i$ is the model forecast obtained using sliding training period method with an optimal window

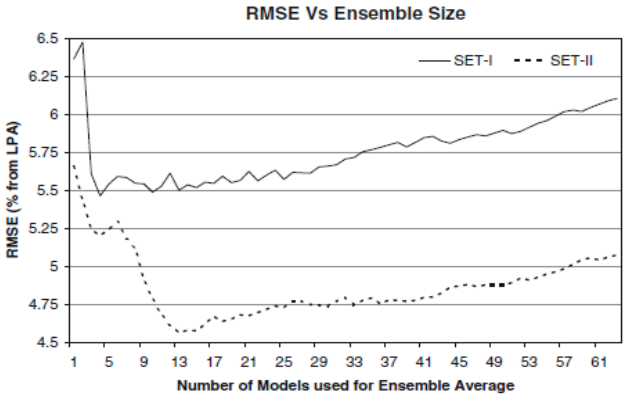


Fig. 3: RMSE vs. Ensemble size

period of length 23 years. GCV is nearly equal to the square of the RMSE with a correction for the number of predictors used in the model. The model with lowest value of GCV was ranked first and model with highest value of GCV was ranked last. The scatter plot of GCV values against the rank of the model is shown in Fig. 2. In the second step, ensemble average of forecasts from the models ranked based on GCV values was computed for the period 1981–2004 by using first 1 model, first 2 models, first 3 models and

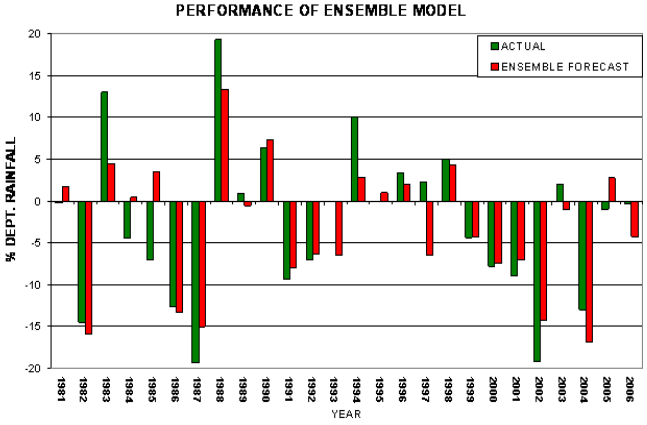


Fig. 4: Performance of EMR model for 1981-2006

so on up to all the possible 63 models in the rank list as the ensemble members. The ensemble average for each year of the independent period 1981–2004 was computed as the weighted average of the forecasts from the individual ensemble members. The weighted average is calculated as: $\hat{y}_E = \frac{\sum_{i=1}^{k} w_i \hat{y}_i}{\sum_{i=1}^{k} w_i}$, where $\hat{y}_E$ is the ensemble forecast for a given year, $\hat{y}_i$ is the corresponding forecast by i^{th} of the k ensemble member models and w_i's are the weights. Here, we have used the adjusted R^2 of the model during the training period as the weights. The RMSE shows a decrease in the RMSE with increase in the ensemble size to reach its lowest value (4.56% of long period average) for the 13-member ensemble for SET-II as shown in Fig. 3. Further increase in the ensemble size shows increase in the RMSE. Thus, in the case of models derived from SET-II, the natural break point is reached at ensemble size of 13. Therefore, these 13 MR models were selected to build the ensemble model, EMR-II. Similarly for SET–I forecast; the breakpoint is reached at ensemble size of 4.

Conclusions

IMD also uses nonlinear models such as Projection Pursuit Regression (PPR) alongwith ERM model. Verification of the results (Fig. 4) with the past data showed that the ensemble method performed better than the individual models.

For the forecast of July rainfall over the country as a whole, a statistical model with 6 predictors was developed using Principal Component Regression (PCR) technique. The predictors used are: North Atlantic Sea surface temperature (December of previous year), NINO 3.4 Sea Surface Temperature (May +June), North Pacific mean sea level pressure (March), East Asia mean sea level pressure (February + March), North Atlantic mean sea level pressure (May) and Equatorial Indian Ocean mean sea level pressure (November of previous year). The model error of the model for July rainfall is 9%.

In addition, IMD prepares operational long-range forecasts for the Winter Precipitation (Jan to March) over Northwest India and Northeast Monsoon rainfall (October to December) over South Peninsula. For this purpose, separate statistical models have been developed.

References

- Rajeevan M, Pai DS, Anil Kumar R (2005) New statistical models for long range forecasting of southwest monsoon rainfall over India. NCC Research Report No 1/2005, India Meteorological Department, Pune, India
- Rajeevan M, Pai DS, Dikshit SK, Kelkar RR (2004) IMD's new operational models for long range forecast of south-west monsoon rainfall over India and their verification for 2003. Curr Sci 86:422–431
- Rajeevan M, Guhathakurta P, Thapliyal V (2000) New models for long range forecasts of summer monsoon rainfall over Northwest and Peninsular India. Meteorol Atmos Phys 73:211–225

Program on Career Trends in Statistics in 7-Day Workshop on Recent Trends in Career Opportunities in Higher Education in the Light of N.E.P. 2020

Seminar lecture on "**Accelerating Science and Skilling for the Next Generation of Jobs: A Global Exposure**", 1st March, 2017.

ANALYSIS OF SURVIVORSHIP OF THREE LEGENDS IN TEST CRICKET

Soumydeep Rudra & Debjyoti Sahu, Final Semester students of 2022

1 Introduction

In the history of World cricket, many Indian batsman have contributed a lot. Sachin Tendulkar, Sourav Ganguly,Rahul Dravid are some memorable names. No doubt all of them are legends of Indian cricket but when the question came to our mind who can survive more time on crease, that is, who is expected to score more runs there arises a lot of arguments among their fanbase. Here I have tried to find answer this to some extent using the concept of life table.

WHAT IS LIFE TABLE ? Life table is used to compare mortality over two distinct situation.

It also answers

1. For each age what the probability is that a person of that age will die before their next birthday ("probability of death").
2. How long a person can survive.
3. How long a person can live after reaching age x.

A life table (also called a mortality table or actuarial table) is a table which shows, for each age, what the probability is that a person of that age will die before their next birthday ("probability of death"). In other words, it represents the survivorship of people from a certain population.

COMPONENTS OF LIFE TABLE

1. l_x : The number of persons who attain exact age x out of an assumed number of births l_0 .

2. ${}_nd_x$: The number of persons among the l_x persons reaching age x, who die before reaching age x+n .

$$ {}_nd_x = l_x - l_{x+n} $$

3. ${}_nq_x$: The probability that a person of exact age x will die before reaching age x+n.

$$ {}_nq_x = \frac{{}_nd_x}{l_x} $$

4. ${}_np_x$: The probability that someone aged exactly x will survive for n more years, i.e. live up to at least age x+n years .

$$ {}_np_x = 1 - {}_n\, q_x $$

5. ${}_nL_x$: The number of years lived by cohort of l_0 persons between ages x and x+n.

$$ {}_nL_x = \int_x^{x+n} l_t dt $$

6. T_x : The number of years lived by l_0 persons after attaining age x.

$$ T_x = \int_0^{\infty} l_t dt $$

7. e_0^x : The average number of years lived after age x by each of the l_x persons who attain that age.

$$ e_0^x = T_x / l_x $$

$\frac{l_x}{l_0}$ is proportions of new born surviving upto age x. These proportions are called survival probabilities.

2 Methodology

Our objective is to use the components of life table to compare the survivorship of three legends of Indian cricket Sachin Tendulkar, Rahul Dravid and VVS Laxman.

Here we use the following interpretations of the life table functions :

(a) l_x : Number of times the batsman scored runs more than or equal to x.

(b) ${}_nd_x$: Number of times the batsman get dismissed between runs x to x+n .

(c) ${}_nq_x$: The probability that the batsman get dismissed in the scoring interval x to x+n.

(d) ${}_np_x$: The probability that the batsman will survive in the interval x to x+n.

(e) $s(x)$: The probability that a batsman will score run more than or equal to x.

$$s(x) = \frac{l_x}{l_0} = \text{Survival probability}$$

3 DATA ANALYSIS

For the ease of comparison here runs of first 200 innings played by the batsman is considered and the not out innings are ignored.VVS Laxman have played a total of 197 not out innings. Now the data collected on the three batsman is as follows:

Rahul Dravid

Interval	X	${}_nd_x$
0 – 10	0	50
10 – 20	10	29
20 – 30	20	20
30 – 40	30	22
40 – 50	40	12
50 – 60	50	13
60 – 70	60	9
70 – 80	70	6
80 – 90	80	9
90 – 100	90	9
100 – 110	100	2
110 – 120	110	4
120 – 130	120	1
130 – 140	130	2
140 – 150	140	4
150 – 160	150	0
160 – 170	160	2
170 – 180	170	0
180 – 190	180	1
190 – 200	190	1
200 – 210	200	0
210 – 220	210	1
220 – 230	220	1
230 – 240	230	2

Sachin Tendulkar

Interval	X	${}_nd_x$
0 – 10	0	54
10 – 20	10	31
20 – 30	20	16
30 – 40	30	18
40 – 50	40	13
50 – 60	50	12
60 – 70	60	9
70 – 80	70	8
80 – 90	80	6
90 – 100	90	7
100 – 110	100	3
110 – 120	110	5
120 – 130	120	3
130 – 140	130	2
140 – 150	140	3
150 – 160	150	1
160 – 170	160	2
170 – 180	170	5
180 – 190	180	0
190 – 200	190	1
200 – 210	200	0
210 – 220	210	1
220 – 230	220	0
230 – 240	230	0

VVS Laxman

Interval	X	${}_nd_x$
0 – 10	0	14
10 – 20	10	39
20 – 30	20	33
30 – 40	30	24
40 – 50	40	18
50 – 60	50	8
60 – 70	60	12
70 – 80	70	15
80 – 90	80	10
90 – 100	90	3
100 – 110	100	13
110 – 120	110	2
120 – 130	120	0
130 – 140	130	1
140 – 150	140	1
150 – 160	150	1
160 – 170	160	0
170 – 180	170	1
180 – 190	180	1
190 – 200	190	0
200 – 210	200	0
210 – 220	210	0
220 – 230	220	0
230 – 240	230	1

SURVIVAL PROBABILITIES

Rahul Dravid

Interval	X	l_x	${}_nd_x$	${}_nq_x$	${}_np_x$	$s(x)$
$0-10$	0	200	50	0.25	0.75	1
$0-10$	10	150	29	0.193333	0.806667	0.75
$20-30$	20	121	20	0.165289	0.834711	0.605
$30-40$	30	101	22	0.217822	0.782178	0.505
$40-50$	40	79	12	0.151899	0.848101	0.395
$50-60$	50	67	13	0.19403	0.80597	0.335
$60-70$	60	54	9	0.1666767	0.833333	0.27
$70-80$	70	45	6	0.133333	0.866667	0.225
$80-90$	80	39	9	0.230769	0.769231	0.195
$90-100$	90	30	9	0.3	0.7	0.15
$100-110$	100	21	3	0.142857	0.457143	0.105
$110-120$	110	19	4	0.201526	0.789474	0.095
$120-130$	120	15	1	0.066667	0.933333	0.075
$130-140$	130	14	2	0.142857	0.857143	0.07
$140-150$	140	12	4	0.333333	0.666667	0.06
$150-160$	150	8	0	0	1	0.04
$160-170$	160	8	2	0.25	0.75	0.04
$170-180$	170	6	0	0	1	0.03
$180-190$	180	6	1	0.66667	0.833333	0.03
$190-200$	190	5	1	0.2	0.8	0.025
$200-210$	200	4	0	0	1	0.02
$210-220$	210	4	1	0.25	0.75	0.02
$220-230$	220	3	1	0.333333	0.666667	0.015
$230-240$	230	2	2	1	0	0.01

Sachin Tendulkar

Interval	X	l_x	${}_nd_x$	${}_nq_x$	${}_np_x$	$s(x)$
$0-10$	0	200	54	0.27	0.73	1
$0-10$	10	146	31	0.212329	0.787671	0.73
$20-30$	20	115	16	0.13913	0.86087	0.575
$30-40$	30	99	18	0.181818	0.818182	0.495
$40-50$	40	81	13	0.160494	0.839506	0.405
$50-60$	50	68	12	0.176471	0.823529	0.34
$60-70$	60	56	9	0.160714	0.839286	0.28
$70-80$	70	47	8	0.172013	0.829787	0.235
$80-90$	80	39	6	0.153846	0.846154	0.195
$90-100$	90	33	7	0.212121	0.787879	0.165
$100-110$	100	26	3	0.115385	0.884615	0.13
$110-120$	110	23	5	0.217391	0.782609	0.115
$120-130$	120	18	3	0.166667	0.833333	0.09
$130-140$	130	15	2	0.133333	0.866667	0.075
$140-150$	140	13	3	0.230769	0.769231	0.065
$150-160$	150	10	1	0.1	0.9	0.05
$160-170$	160	9	2	0.222222	0.777778	0.045
$170-180$	170	7	5	0.714286	0.285714	0.031
$180-190$	180	2	0	0	1	0.01
$190-200$	190	2	1	0.5	0.5	0.01
$200-210$	200	1	0	0	1	0.005
$210-220$	210	1	1	1	0	0.005

VVS Laxman

Interval	X	l_x	${}_nd_x$	${}_nq_x$	${}_np_x$	$s(x)$
$0-10$	0	197	14	0.071066	0.928934	1
$0-10$	10	183	39	0.213115	0.786885	0.928934
$20-30$	20	144	33	0.229167	0.834711	0.730964
$30-40$	30	111	24	0.216216	0.782178	0.563452
$40-50$	40	87	18	0.206897	0.793103	0.441624
$50-60$	50	69	8	0.115942	0.884058	0.350254
$60-70$	60	61	12	0.196721	0.803279	0.309645
$70-80$	70	49	15	0.306122	0.693878	0.248731
$80-90$	80	34	10	0.294118	0.705882	0.172589
$90-100$	90	24	3	0.125	0.875	0.121827
$100-110$	100	21	13	0.619048	0.380952	0.106599
$110-120$	110	8	2	0.25	0.75	0.040609
$120-130$	120	6	0	0	1	0.030457
$130-140$	130	6	1	0.166667	0.833333	0.030457
$140-150$	140	5	1	0.2	0.8	0.025381
$150-160$	150	4	1	0.25	0.75	0.020305
$160-170$	160	3	0	0	1	0.015228
$170-180$	170	3	1	0.333333	0.666667	0.015228
$180-190$	180	2	1	0.5	0.5	0.010152
$190-200$	190	1	0	0	1	0.005067
$200-210$	200	1	0	0	1	0.005067
$210-220$	210	1	0	0	1	0.005067
$220-230$	220	1	0	0	1	0.005067
$230-240$	230	1	1	1	0	0.005067

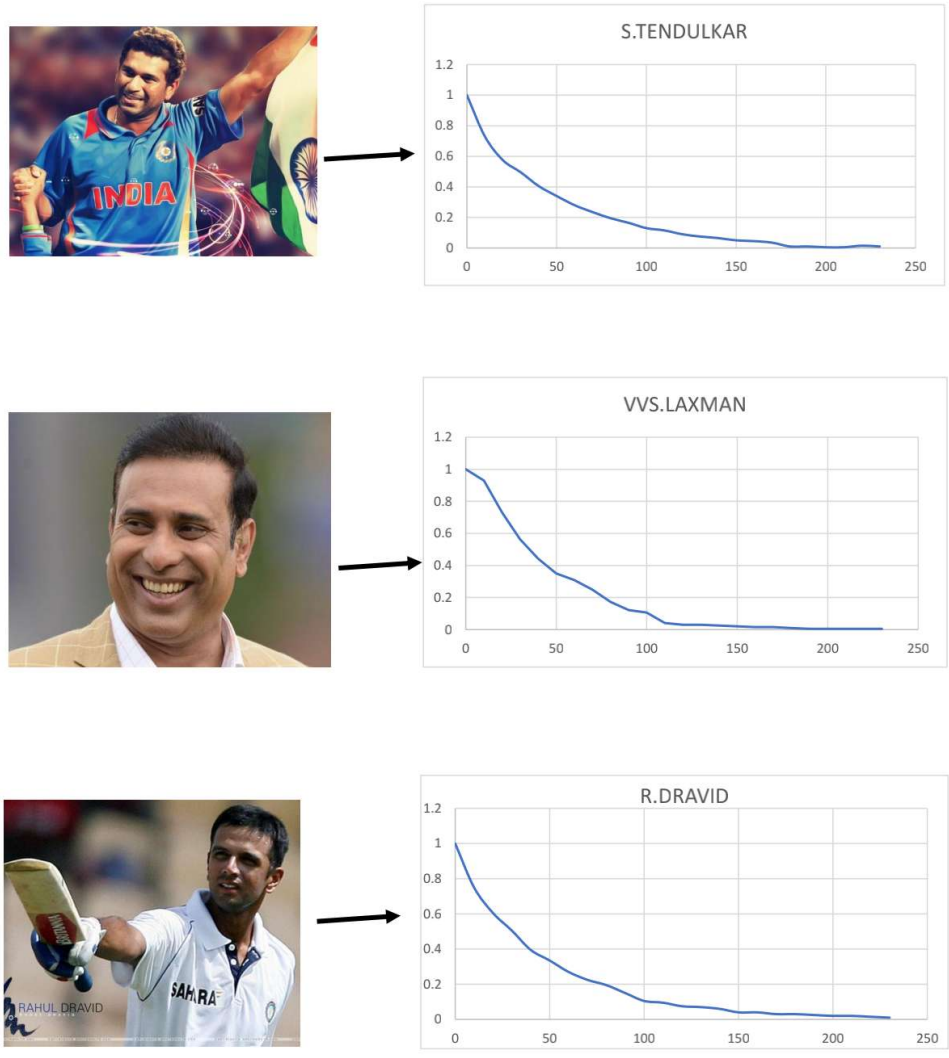


Figure 1: SURVIVAL PROBABILITY

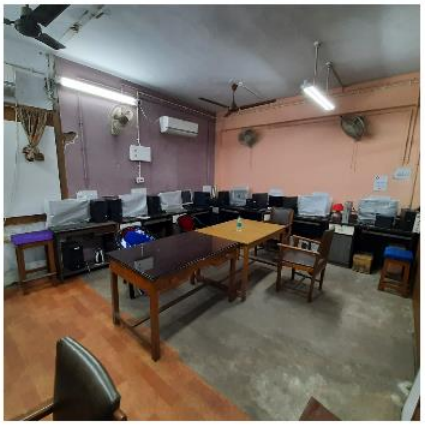
Figure 2: COMPARISON OF SURVIVAL PROBABILITIES

4 CONCLUSION

On comparing the survival curve of the three batsman we see that the curve of VVS Laxman was above the survival curve of the others in the early run scoring intervals but it goes down the others around run 75. The curve of Rahul Dravid and Sachin Tendulkar are more or less same. We can conclude that VVS Laxman is expected to get a decent start in every match, but once getting a start both Sachin Tendulkar and Rahul Dravid is expected to convert it to higher runs as compare to VVS Laxman. Around 100,Sachin Tendulkar has the highest survival probability though that of Rahul Dravid and VVS Laxman arev almost same. After 175, the survival probability of Rahul Dravid is more.

Acknowledgement: I acknowledge my friends for helping me in data collection.


SEMINAR LIBRARY




3. Moving Average Method
GPS Map Camera
Haldia, West Bengal, India
335M+6R7, Haldia, West Bengal 721631, India
Lat 22.057646°
Long 88.084224°
19/07/23 11:32 AM GMT +05:30
Google