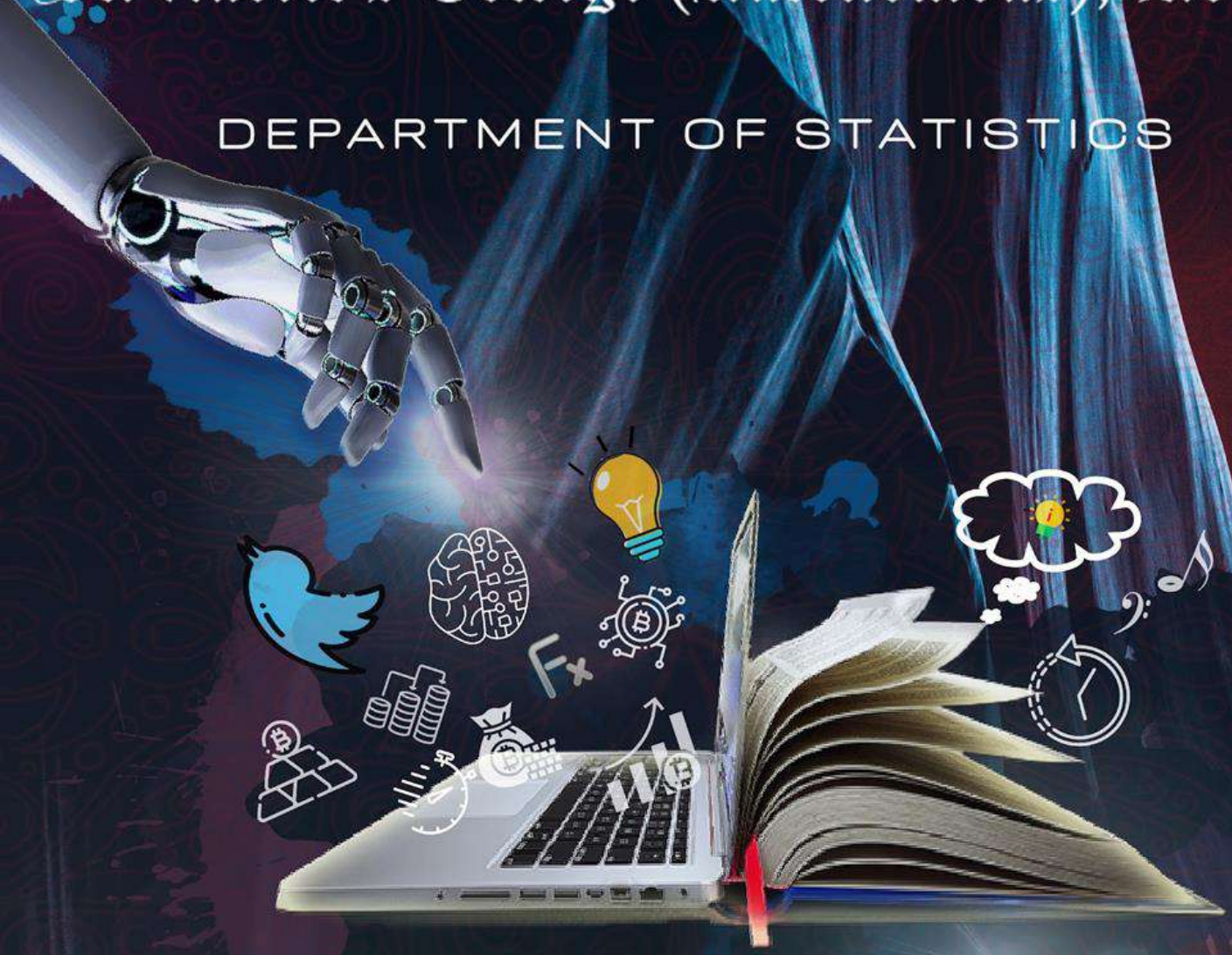
St. Xavier's College (Autonomous), Kolkata

DEPARTMENT OF STATISTICS



PRAKARSHO

VOL XIV

2022

ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA

DEPARTMENT OF STATISTICS

PRAKARSHO

VOL XIV

2022

PRAKARSHO

VOL XIV

2022

Designed By

Arka Roy

Cover Art and Illustrations

Rachit Yadav

EMAIL: stsa@sxccal.edu

PHONE: 2255-1270

CONTENTS

CONTENTS

CONTENTS

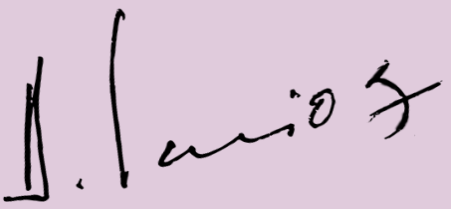
Message from the PRINCIPAL

Rev. Dr. Dominic Savio, SJ
Principal
St. Xavier's College (Autonomous), Kolkata

"I take great pride in congratulating the Department of Statistics of St. Xavier's College (Autonomous), Kolkata, for continuing its legacy on the path of excellence and publishing the 14th edition of its Annual Departmental Magazine, 'PRAKARSHO'.

The magazine is a unique platform for the students to bring forth their ideas, interpretations and research in the field of Statistics. This carefully curated magazine successfully upholds the spirit of innovation and encourages young minds to continue on their quest for knowledge.

I wish the Department of Statistics success in all its future endeavours."



Principal

Message from the VICE-PRINCIPAL

Prof. Bertram Da Silva
Vice-Principal
St. Xavier's College (Autonomous), Kolkata

"The Department of Statistics of St. Xavier's College (Autonomous), Kolkata, has been instrumental in furthering the cause for research and innovation for decades. Its commitment to the spirit of learning is reflected in its latest edition of the departmental magazine, 'PRAKARSHO 2022'.

The magazine brings together students and teachers from all over the country to share ideas and intriguing theories on the subject of Statistics. It is a creative compilation of articles, research papers and other interesting segments, which motivates students to learn, grow and explore the field of science.

I congratulate the students and the faculty members of the Department of Statistics for successfully publishing the 14th edition of Prakarsho."

Vice-Principal

Message from the DEAN OF SCIENCE

Dr. Tapati Dutta
Dean of Science
St. Xavier's College (Autonomous), Kolkata

"The Annual Departmental Magazine of the Department of Statistics of St. Xavier's College (Autonomous), Kolkata, 'PRAKARSHO', has always been an extremely innovative compilation of articles in the field of Statistics. It has successfully upheld the legacy of this college with its commitment to excellence, innovation and originality.

I extend my heartiest congratulations to the Department of Statistics for successfully publishing the 14th edition of 'PRAKARSHO'. I am confident that this edition will be extremely influential in encouraging young minds to explore the world of science. I wish them good luck for another year of ingenuity and brilliance."

Dean of Science

Message from the HEAD OF THE DEPARTMENT

Dr. Durba Bhattacharya
Head, Department of Statistics
St. Xavier's College (Autonomous), Kolkata

"I take immense pride in presenting the 14th edition of our Annual Departmental Magazine, 'PRAKARSHO'. The Department has always strived to encourage young and budding statisticians to explore their discipline and voice their thoughts and interpretations in an innovative medium.

'PRAKARSHO' is our humble attempt to provide an inclusive platform to students and teachers alike from all over the country and even abroad to present their ideas, theories and research in the form of scientific articles.

I extend my heartfelt gratitude to Father Principal, Vice-Principal and the Dean of Science for their perennial guidance and encouragement, and acknowledge sincere thanks to my colleagues for making it a success. I congratulate the Student Editorial Committee for successfully compiling this edition through their hard work and diligence."

Head of the Department

Message from the EDITOR'S DESK

The quintessence of normality is that it changes with time. The world as we knew it, is and shall not be the same. After a long halt, when it is time again to step back into life, we redefine and adapt to new ways. Thus, walking on the path of "Redefining Normality", we present to you the 14th Edition of our magazine, PRAKARSHO 2022.

PRAKARSHO has always broken barriers and this year's edition is no exception. But this would not have been possible without all those people who have dedicated their most sincere efforts to this journey. We would like to express our heartfelt gratitude to our respected professors, the entire student working committee and the various authors who have contributed to this magazine.

Hence, with immense pride and pleasure, we present to the readers of this magazine a carnival of statistical ideas and thoughts – PRAKARSHO 2022.

Editorial BOARD

Patron

Rev. Dr. Dominic Savio, SJ
Principal

Advisory Board

Prof. Bertram Da Silva
Vice-Principal

Dr. Tapati Dutta
Dean of Science

Dr. Argha Banerjee
Dean of Arts

Prof. Surabhi Dasgupta
Prof. Surupa Chakraborty
Prof. Debjit Sengupta
Prof. Pallabi Ghosh

Prof. Ayan Chandra
Prof. Durba Bhattacharya
Prof. Madhura Das Gupta

Rajnandini Kar
Student Editor

Abhinandan Bag
Associate Student Editor

Editorial BOARD

Student Editorial Committee

Anuroop Roy
Anirban Ghosh
Achena Sengupta
Saheli Datta
Saikat Datta
Sayan Bera
Sweata Majumder
Srishti Lakhotia
Sayan Das
Rhritajit Sen
Kaulik Poddar
Hrithik Sen
Yenisi Das
Himagno Roy
Tathagata Banerjee
Abhay Ashok Kansal

Student Designing Committee

Shamie Dasgupta
Sambit Ghosh
Tanishi Parasramka
Arka Roy
Svadhaa Agarwal
Rachit Yadav

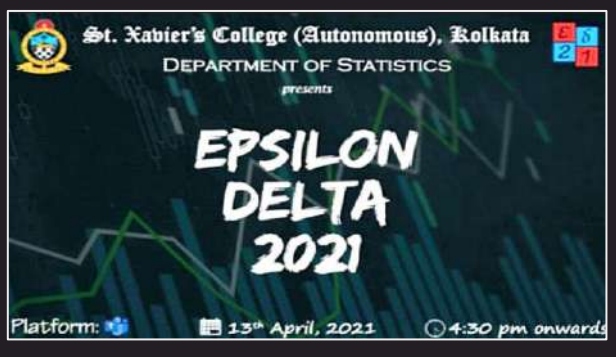
Departmental Report: 2021 - 2022

Departmental Activities:

EPSILON DELTA – THE ANNUAL DEPARTMENTAL EVENT

The Department conducted the 2021 version of its Annual Event "Epsilon Delta" on 13th April, 2021, on MS Teams, for the first time owing to the pandemic. The event commenced with the launch of the 13th Edition of the Departmental Magazine "Prakarsho". Organized throughout the day were –

- Webinar on 'Robust Estimators' by Prof. Gaurangadev Chattopadhyay, Department of Statistics, Calcutta University.
- Paper presentations by students – 'Proectura'.
- A Cultural Program arranged by the students of the Department.

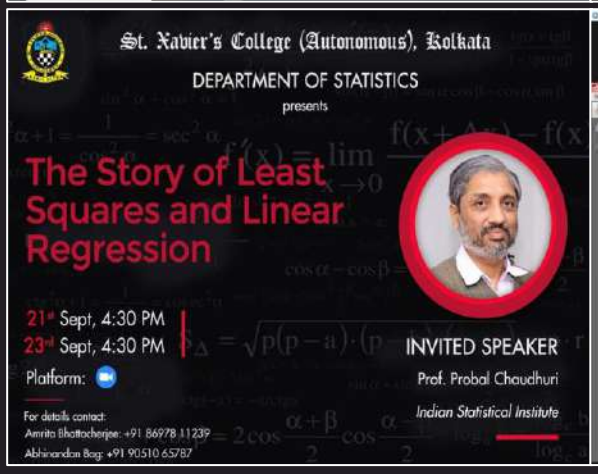


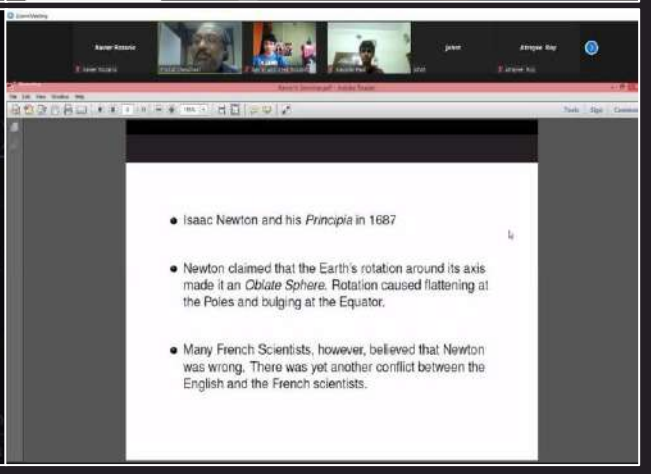
Departmental Report: 2021 - 2022

Departmental Activities:

WEBINAR ON "THE STORY OF LEAST SQUARES AND LINEAR REGRESSION":

The Department had organized a Webinar, streamed on Zoom, on "*The Story of Least Squares and Linear Regression*" on 21st and 23rd September, 2021. Prof. Probal Chaudhuri from the Indian Statistical Institute, India, was the speaker.





Departmental Report: 2021 - 2022

Students' Achievements:

Awards and Recognitions (3rd Years):

1. Anuroop Roy -
   - Second Place at Coupe de-Mandeville Season 2 Football, representing Mandeville FC Club, 10th April, 2021.
   - Best Goalkeeper Award at Coupe de-Mandeville Season 2 Football, Mandeville FC Club, 18th November, 2021.

2. Adrija Saha -
   - Published an article on 'Newton-Pepys Problem: From a Different Viewpoint' in Beacon-9th Edition, Department of Mathematics, St. Xavier's College (Autonomous), Kolkata on 20th September,2021.

3. Shrayan Roy-
   - Published an article on 'Newton-Pepys Problem: From a Different Viewpoint' in Beacon-9th Edition, Department of Mathematics, St. Xavier's College (Autonomous), Kolkata on 20th September,2021.

Departmental Report: 2021 - 2022

Students' Achievements:

4. Saptarshi Chowdhury-
    - Winner of 'Sphinx Quiz Event' in Analytica 2021, Department of Mathematics, St. Xavier's College (Autonomous), Kolkata, 21st September, 2021.

Awards and Recognitions (2nd Years):

1. Xavier Abhishek Rozario-
    - Made and launched an application on Google Play Store on 28th May, 2021. App Link: https://play.google.com/store/apps/details?id=com.xavi.linearAlg

2. Yenisi Das-
    - Winner of 'Sphinx Quiz Event' in Analytica 2021, Department of Mathematics,St. Xavier's College (Autonomous), Kolkata, 21st September, 2021.

    - Second Runner Up in X-Tra Innings (3-Day Mathematics Contest), Mathematics Society of St. Stephen's College, Delhi University, 29-31st October, 2021.

Departmental Report: 2021 - 2022

Students' Achievements:

Awards and Recognitions (1st Years):

1. Ankita Dey-
   - First position in an inter-district online quiz competition, George Telegraph, 5th March, 2021.

2. Sohini Mandal-
   - Gold award in an art exhibition, Manikarnika Art Gallery, 20th December, 2021.

   - First Position in Drawing Competition and First Position in pencil sketch competition, Fairy Kids, 16th January, 2022.

In Conversation with Prof. Nitis Mukhopadhyay

Prof. Nitis Mukhopadhyay has been a full-time professor in the Department of Statistics, University of Connecticut-Storrs, USA, since 1985. Prior to joining the University of Connecticut, Prof. Mukhopadhyay was a faculty member at the University of Minnesota-Minneapolis, University of Missouri-Columbia and Oklahoma State University-Stillwater. He received his PhD (1975) from Indian Statistical Institute, Calcutta. He made prolific contributions in a number of areas in Statistics including Statistical Inference - parametric and nonparametric, Sequential Analysis, Multiple Comparisons, Clinical trials, Applied Probability, Econometrics, and others. As an author or co-author, Prof. Mukhopadhyay published 6 books, 18 book chapters, nearly 300 peer-reviewed research papers, and edited or co-edited more than 7 special volumes.

Prof. Mukhopadhyay spent his career in a pool of extraordinary knowledge in the field of Statistics and worldwide exposure. Hence, the students decided to have a candid conversation with him about Statistics as a discipline and how it can structure the future. This interview was conducted by Amrita Bhattacherjee, Rajnandini Kar, Abhinandan Bag and Xavier Abhishek Rozario, on behalf of the editorial committee, on 16th and 23rd February, 2022 via Google Meet.

In Conversation with Prof. Nitis Mukhopadhyay

INTERVIEW:

Q: Sir, firstly, would you like to share something from your early schooling days?

NM: Sure! I was born a long time ago. I started my schooling directly from grade 7. I never went to school before that. My parents are from what is now known as Bangladesh. Back then, it was known as East Bengal. They moved from Bangladesh with my brothers and sisters on around 13th August 1947. We were refugees with nothing apart from the clothes on our backs. I was raised in a ghetto (or, a slum) in Calcutta. Since I did not attend school, my father used to teach me Mathematics at home and my mother used to teach me Languages, History, Geography, etc. We did not have any chalkboard, paper or pen, so we used the floor. Six or seven of us used to stay in one room, under one roof. We used to have an area on the floor in one corner of the room which was full of Mathematics and Geometry, and my mother was forbidden to clean that area.

When I started school in grade seven, I considered myself to be a great Mathematician. Being a freedom fighter, my father never believed in working for anyone for a salary. So, he gave free tuition to all the people near our ghetto. That is how the headmaster of the school knew us. You may not be familiar with it, but in ghettos you see gambling, prostitution, stealing, murders; you see death from diseases like cholera, typhoid and burns. But nobody touched us, because of Mathematics.

In Conversation with Prof. Nitis Mukhopadhyay

Even our landlord did not care to take rent from us because he knew we didn't have any money. Many people were kind to us, only because of education. That's what moved us forward. Even when we went to school, we didn't have any books. Our headmaster donated his books to me, and he did that for all my other siblings. I used to go to his office to share his lunch. That's a privilege that we had. Such a magnificent life in so many ways! The same headmaster would punish me for something else I did in the school because I was always getting in trouble.

After being admitted to school, I did well in grade seven. In eighth grade, we had a great teacher named Hari Mohan Chatterjee. He was known to be very strict. In his course, I scored 28 in Math. That is when I realized that it was time for me to wake up. Everybody in my locality knew I got 28 and I was almost beaten up by my family. They went, "How could you do this? 28 in Math and we thought you were a genius!" Our school was not very well known - Salkia Anglo-Sanskrit High School. It was a good school, but not famous - not like your institution, or like Calcutta Boys. But we had excellent teachers. So that's how it all started.

Q: What drove you towards the subject?

NM: In school, Physics, Chemistry (and Math of course) were impressive subjects for me. I was not very good at languages at that time because I did not have a whole lot of social skills when I started going to school.

In Conversation with Prof. Nitis Mukhopadhyay

I think my own passion and the passion that was created by my family was what drove me to this subject. Even though we had no money, we were always passionate about education. I am the youngest in the family. There were six of us. My 'chorda' (one of my elder brothers) was a statistician. His name was Anish Chandra Mukhopadhyay. He was 11 years older than I was. When he was in university, I was still playing in the ghetto with cow dung, mud and marbles. He was a gold medalist in B.A, M.A. and I.A from Calcutta University. I did not realize who he was because he was so much older than I was.

My brother was already a Mathematician and then I went to Presidency College to study Honours in Mathematics. That was in 1967. In Presidency College's Math Honours curriculum, there was a paper in Statistics – the eighth paper. By that time my elder brother was already famous in his own way, and I heard about Statistics. I almost didn't know anything at all except for that eighth paper. That was my contact with statistics at that time. While studying math, I learned number theory, calculus, differential equations, trigonometry, three-dimensional trigonometry and so on. However, I had a lot of friends in the Statistics department. At that time, it was in the Baker's Laboratory. . I used to go there very frequently to meet my friends. Again, I knew a couple of professors at that time, because of my brother. I'm sure you know them by name, like Professor Gun, Professor Arijit Chowdhury, Prof. Sujit Basu, Prof. Utpal Dasgupta. Prof. Dasgupta was one of the founders of your program at St. Xavier's College. Utpal Dasgupta and Arijit Chowdhury were classmates of my Chorda. Utpal Da and Arijit Da were known to me since I was a kid.

In Conversation with Prof. Nitis Mukhopadhyay

When I was a little kid, I used to climb on their shoulders, and they would carry me around. That was a big connection, why I eventually moved to statistics. I wanted to be a pure Mathematician. My passion was to work in number theory because my number theory professor, Professor Biren Bose, was just phenomenal.

Then the Naxal Movement started. Almost every day, the Naxalites would come, and they would break things, beat people up, throw chairs at them. That's when, in 1970, I decided to take the exam in ISI. I was luckily selected and that's what actually changed my life. If the Naxalite movement had not been there, I would probably still be in Calcutta University doing number theory or something. So, different influences in life take stance.

Q: How has your perception of statistics evolved through your educational and teaching life?

NM: No subject, no discipline is static. Every discipline is a dynamic process. So, Statistics has changed drastically. From Presidency College when I moved to Indian Statistical Institute to do M.Stat in the fall of 1970, Statistics was also Mathematics at that time. So, most of us learnt all sorts of Mathematical Statistics and not so much of Applied Statistics. There were some Applied Statistics, but there was no hands-on experience. We learned computer programming as a couple of command lines on the chalk board. But the teachers were fantastic. In ISI, for the first time, I found great thinkers. They taught me how to think. Then, by the time I finished Master's and started doing PhD, it was a similar situation.

In Conversation with Prof. Nitis Mukhopadhyay

I saw the evolution after I moved here in USA. Here, I saw how Applied Statistics was different than the concept of Applied Statistics in India at that time. In the last 20 years, evolution has been more drastic. Suddenly, Financial Mathematics, having a big cross-section with Statistics, became a big area. Non-statisticians in business schools, chemists, psychologists started using Statistics. This is the time of big data – large parameters, low observations. It's called HDLSS, High Dimension Low Sample Size. We see very high dimensional data.

If you look at the history of Statistics, you will see that the people who have developed Statistics, almost all of them were non-statisticians. Like, Carl Pearson. He was in Anthropology, Sociology, Psychology, but he laid some of the foundations of Statistics. R.A. Fisher was a geneticist. In our country, Mahalanobis was an Applied Physicist. Just by sheer chance, he became interested in Statistics after he went to England and saw Biometrica journals, which was founded by Carl Pearson. You see the connection?

But the later generation, for example, C.R. Rao was straight in Mathematics. He came to Calcutta and met Mahalanobis at Presidency College. Mahalonabis hired him as a technician. That's how his journey to Statistics started. Then, when he went to Fisher as a doctoral student, he learned some Applied Statistics. You will see that there are some pictures of Rao, very young, as a kid, cleaning the cages of guinea pigs that R.A. Fisher used to have in his laboratory as a geneticist. Even machine learning was started by Rao in those days.

So, evolution goes in steps. You have to just keep swimming. Some people can keep up with the pace, and some people cannot. And those who can, shine eventually.

In USA, the environment was different. In this country, what I told you about Dr Rao, that trend had started long before Dr Rao's encounter. The best statisticians in this country are in the Economics department, in Engineering schools and Business schools. Not in the Math or Statistics department. Zellner was an Economist; Ingram Olkin (pioneer of multivariate analysis) was in Psychology. So, that trend has been here for a long time. It took a while to go to India, but now you are swimming in it. It's everywhere, which is good.

Q: Would you tell us a bit about your graduation and post-graduation days?

NM: In Presidency College, I was studying Math honours because I did not want to do Statistics at that time. But in Statistics, I had great friends. Dibyen Majumdar is world-famous in combinatorics. Siddheswar Ray did work on pattern recognition. Dilip Roy was in Burdwan University. They were all my classmates, but they were all in Statistics. I was in Mathematics. I really wanted to take Statistics pass course because my friends were there. It was a pity that I had to take Physics and Chemistry. You know why? It's funny. In Baker's Laboratory, in the Statistics department, there were 12 facet machines. So, they could take 12 students in the pass course. Normally 7 or 8 will come from Physics, 3 or 4 will come from Chemistry and that exhausted the system. Maybe there would be room for just one guy from Mathematics.

In Conversation with Prof. Nitis Mukhopadhyay

But in our time, six students each from Physics and Chemistry took statistics as their pass subject. I had to pick Chemistry. So, I ruined their labs! During titration, I would put things in and then go chat with other people while my solution's colour changed from red to white and then to green. So, I was very disliked by my professors in Chemistry.

Then I went to Indian Statistical Institute. I had great teachers, great friends. In my second-year specialization, I picked Advanced Probability and Mathematical Statistics. At that time, that's what it was called - APMS. Professor Malay Ghosh became my advisor. Professor J.K. Ghosh became one of my mentor. Professor Ashok Mitra, a student of David Blackwell, did his PhD at Berkley. Dr Basu (of the Basu's Theorem) changed my way of thinking. He is often referred to as the master of column space. There is no problem in linear algebra that Prof. Mitra could not solve with column space. He could explain anything at any level. Professor Sujit Mitra was another noteworthy professor. He never used to write anything on the chalkboard. He would just walk up and down for a 50-minute lecture and walk out at the end. The class is over and there is not one single thing on the chalkboard that you could substantially see. So, my job (and I took that job myself) was to go back that night and recreate the notes. Every night. Every class. And my friends would come and borrow those notes. But why that was important was that it made me think, and re-think. Where did this training come from? This training came from my father. From doing mathematics only on the floor. And, we had to remember all those formulae and concepts.

In Conversation with Prof. Nitis Mukhopadhyay

One day, Sujit Mitra taught us something that caught my interest. So, that night I started working on what he was possibly saying. And I got something. I was just a first-year Masters student. So, I went to Prof. Mitra's office, and I showed it to him. I was standing there hoping that he will say something. He looked at it for a while without saying anything. Then, he finally said, " I have written some of these things in my paper. It is being published soon at Sankhya". He wrote it with his prized student Bhimasankaram, and I managed to touch parts of it without knowing anything. That gave me confidence. After my Masters, I wanted to come to the US and do my PhD, but I didn't have any money. Even to apply, you had to pay 25 USD application fee. What I did was, I already knew many professors in the US. Some of them are very famous, like Prof. Pranab Sen. So, I met him at ISI and asked him, "Would you please pay my application fee?" He paid it for me. Whenever I had some money, I went back to his house and paid him back. Eventually, I did not leave. I decided to work right there under Prof. Malay Ghosh.

Q: From your editorial experience, what advice would you give to young researchers?

NM: Since you referred to my editorial experience, from editor's point of view, writing is very important. You may know the best result in the world, but if you cannot express that in very well-written form, it's going to remain unknown. This ability takes a long time to develop. What I tell my students is that, if you cannot sell it, it will remain under your bed until you die, and nobody will know.

In Conversation with Prof. Nitis Mukhopadhyay

If I give seminars for general audience, I will give one kind of seminar. The same seminar will change drastically if I am talking to only five people who are experts in the area. So, these lines have to be drawn. I am not talking about formatting, I am talking about how you present things to Biometrica, JASSA, Sankhya, JRSS or other journals.

Let me tell you a very nice story about Rao-Blackwell theorem. Doctor Rao had told this; it is in his memoir. A long time ago, he was a very junior Statistician, and he had a paper in the Calcutta Mathematical Society. It had Rao-Blackwell theorem, Cramer-Rao theorem - everything you want to know about Statistics was in that paper. So, Dr. Rao was attending a seminar. The speaker kept talking about 'Blackwell theorem'. But he noticed that the result is exactly similar to his result. Rao did it before Blackwell. Rao's paper was in '45 and Blackwell's paper was in '46. So, Dr. Rao stood up and said very politely, "Sir, I have that result in my paper, in my thesis". The speaker did not know that, and he said, "Dr. Rao, you did not even state it in your paper as a theorem. Blackwell stated this result as a theorem in his paper". To this, Dr. Rao replied, "Sir, I was only a student at that time". Lucky for Doctor Rao, Jerzy Neymann was present in that session. And this led to the result being called 'the Rao-Blackwell theorem'. So, writing is very important. You have to sell.

In Conversation with Prof. Nitis Mukhopadhyay

Q: You have authored and co-authored some of the most crucial books and chapters in the field of statistics. Are there some literary resources you wish you had when you were a student?

NM: If I want to be politically correct, I should say yes. But I would rather be politically incorrect. I would say no. If we had all the resources back then, that is sixty years ago, I wouldn't be whatever I am today. Nowadays, many students use Google for solutions. That does not teach you anything. Suppose, someone has solved it and I get the answer from Googling. What do I gain from it? My solving would be different, my answers, descriptions, motivations and references would be different. Right after a lecture, some students will come asking questions. I say, "Spend at least 10 hours on it and them come back to me". Then you think about it for 10 hours. You will see that you are getting better. It will help you stretch your imagination. So, I was lucky in many ways. I had good resources and great teachers in Presidency College.

Q: As a professor, what is something that you believe should be given more priority while studying this subject?

NM: What distinguishes it from other subjects is very simple. Other subjects, non-statistics, are mainly closed subjects. Like when you go to Chemistry, Physics, or Computer Science, it is restricted. But in Statistics, if you are a professional, you can work in any field. In that sense, I am very lucky. I can work in agriculture, I can be a teacher, I can work on writing, in education sector, in air quality - I can work in any of these areas.

In Conversation with Prof. Nitis Mukhopadhyay

Statistics involves a lot of thinking and a lot of perception. Somebody asked Stephen Fienberg, “What is different in Statistics? What do you like about it?” He said, "I can play with so many thoughts. I can talk with an engineer, a cancer researcher, even an Orthopedic surgeon. It’s like a playground where kids share different pieces of toys, but they have to leave it there. I do not have to leave it there. I can bring it home”. Therefore, some things that I believe should be given more priority while studying this subject are your thinking, originality and novelty.

Q: What are some changes that you see in the educational landscape as compared to how it was when you were a student? What are your opinions about these changes?

NM: I do not think that the teachers have changed much. Students have evolved, because students are more resilient. Teachers have no time to be resilient, to learn anything new, to challenge you. The university syllabi is tied up. There is a missing link. So, they are teaching you like how they used to teach us, when I was a student. I know they have changed one thing here and two things there, but that is not enough. I am talking about a general movement. So, that makes the teaching part hard for them and uninteresting to you.

Take the current testing procedure, for example. It is not ideal at all. If I give an assignment to the entire class, they will collaborate.

In Conversation with Prof. Nitis Mukhopadhyay

Q: You have been involved in this discipline for decades and have collaborated with students and teachers from all over the world. What is something that distinguishes India from other countries when it comes to the subject of Statistics?

NM: I have a number of Indian PhD students and a number of non-Indian PhD students (from America and various other countries). I think, you have to understand that the countries' trainings are very different from one country to the other. India is no longer what it used to be, in terms of core education. But some countries, for example France, Israel and Germany, have maintained their edge more than India. If you go further north in Europe to Norway or Sweden, you will find that they are more religious about their core. In India what has happened is, we have followed the market economy. This market economy is driven by the United States. Remember, earlier I had said that you have to sell. But there are always limits. If nobody uses Rao-Blackwell Theorem, then after 50 years it will be extinct. So, you have to market things, but then how much? In what depth? Those limitations are some things that you set. You cannot let them sell you short. You are being sold short now because people are not preparing themselves to prepare you well. Now, most of the times, people who come from India, they are mostly very good in Mathematics and Statistics, despite the fallacies we talked about. India is very good in handling 'math' things. But the tendency of mugging things up and writing them down in the exam is very prominent in India. This leads to the students being very mechanical, and then they struggle during interpreting an experiment, in analysing a data.

In Conversation with Prof. Nitis Mukhopadhyay

Applied Maths and Design of Experiments are a few courses in which students from India have trouble. This is something they do not expect – they think that they are good at math, so they should be okay. They think that they will not have any problem in analysing a BIBD. But first, they need to think "Is the experiment a BIBD? Does it make sense to assume this?" That's where you need to rationalise. That's where everybody falls behind. If the experiment is BIBD, I will analyse it; if it's Latin Square, I will analyse it; if it is a normal distribution, I will be able to do it. But should I assume normality, or should I use non-parametric methods? Which way should I go? That link is missing. Another thing that I have seen in India for many years is that the students have very little freedom to change their discipline. For example, you started with Chemistry and now you want to study Psychology – that's a big no. You should be allowed to change your discipline almost at any time, though it shouldn't come for free. Instead of 3 or 4 years, it may take 5 years to graduate. But it is getting better. Because nowadays we see lots of data structures, lots of analysis methods, lots and lots of data. If nothing else, at least you are seeing how to handle data, how to be friendly with data. The difficulty is, if you learn something more, then something else has to give. Because your time is limited right? Therefore, the question is about balancing.

In Conversation with Prof. Nitis Mukhopadhyay

Q: According to you, what are some promising job/research prospects for the current generation of statisticians?

NM: This question is very difficult to answer. It depends on your goal, your satisfaction, your imagination and passion. I don't think everyone should join the academia. Everybody should not join the financial market, everyone should not go and work in the pharmaceutical industry, everybody should not join a genetic lab. But as a statistician, remember the toys? So, we can bring the toys home! You can think of any job market, and you can go there. In this discipline, that is true. Some of my students have gone to the pharmaceutical industry without having taken one course in pharmaceuticals. Some of my students have gone to financial institutions. That doesn't mean you have to be an expert in financial math. But you should have the right motivations. You cannot say, "my teacher has to teach me financial math so that I can go and work in Wall Street". Who told you that you will like it in Wall Street? Have you thought about that? You know, people who work in Wall Street, work for 3-4 years and they get completely burned out. They make a lot of money and then they retire after 4 years to go and lie down in Miami beach for the rest of their lives. They are in their thirties, and they make so much money. But they have no energy to move even a couch; they are so burned out. Think about that! It is not all rosy. But how can you get jobs in different sectors? With the basic training being very strong. There is no substitution.

In Conversation with Prof. Nitis Mukhopadhyay

Q: Apart from theoretical and practical concepts of statistics, what additional skills do you think are important for the students of this generation?

NM: I think two most important things are communication and the ability to write. One has to be able to communicate not only with Statisticians but also with other people. At the same time, you should be very good in writing. I am not talking about the absolutely essential skills of computing and core mathematical tools, because those are obvious. Most of the time, what we ignore are written and spoken skills. These two things should be cultivated very consciously. From an Indian perspective, we are very lucky that a large majority of students learn English from the very beginning. Communication has a style depending on who you are communicating with. Further, your writing skills should be superb - very precise writing and very precise expression of what you feel is essential. I think that is something that's in a lot of graduate students. However, students from India can write well. Many American students cannot write well. They are not good at the kind of conversation you will make in a professional conference. So, these are some skills on top of statistical and mathematical skills that should be emphasized. Writing your projects will definitely help but try to write it in your own way. You should develop your own style.

In Conversation with Prof. Nitis Mukhopadhyay

Q: You have organized, attended and taken part in countless conferences. Which conference is most memorable to you and why?

NM: The memory of a conference is a multi-variable thing. It is very difficult to order a multivariable vector. Some conferences come to mind. Many years ago, I was at Oklahoma State. I suddenly saw that there was a conference in Santa Monica, California. It was a conference on Information Theory and the keynote speaker was Wolfowitz. Jacob Wolfowitz was one of the foremost pioneers of Information Theory and Coding Theory. He collaborated with Abraham Wald. They proved Optimality of Sequential Designs. I saw he was the keynote, so I immediately submitted a paper for the conference and went there, just to see him and attend his talk. Then, there was a conference in Paris in Applied Probability. Albert Shiryaev, Russian Probabilist - one of the best of his century and a student of Kolmogorov - came to give a talk. We interacted and he was in Sequential Probability theory. We became good friends and I now know some of his students. Calcutta Triennial is always my favourite. I have attended all Calcutta Triennials so far. I even organized one of the conferences. In one Triennial, Amartya Sen gave the keynote opening address. Again, I organize a conference called International Workshop in Sequential Methodologies (IWSM). I started it in 2007. It is held every other year and the other alternate year we have International Workshop in Applied Probability. I go to all IWAP and IWSM conferences, because I'm the organizer. A very long time ago, I went to a conference in Uppsala. It is about 2 hours from Stockholm.

In Conversation with Prof. Nitis Mukhopadhyay

Uppsala university is a very prestigious institution and a lot of great Mathematicians worked there. Actually, there's a funny story. My wife and I went back a couple of other times; we have friends there. One time, a friend invited me to dinner. I readily accepted and said "Well, wherever you want to go, just come and pick us up from the hotel". They said, "Pick you up? We cannot pick you up, we do not bring our cars". So, they always travel in bikes. The parking lots have no cars - big parking lots filled with bicycles. Students, faculty, staff - everybody biked there.

Q: Sir, we know that you are a writer by passion, so who are your favourite authors and what are some books and stories that you think every student must read at least once?

NM: Whenever I get bored or whenever I get stuck, which happens many times, I read some other books. I really love to read biographies of scientists, artists and big personalities. I also read lots of musical books. I don't know if you are familiar with it, but in old times there was a very famous music critique named Dhurjati Prasad Mukhopadhyay. His son, Kumar Prasad Mukhopadhyay wrote books on music history. Allaudin Khan has a book named "Amar Kotha". I love to read these. When I am stuck with a problem and cannot proceed, I like to read other things. Obviously, we cannot drop Tagore or Nazrul. Our copy of Gitabitan has countless notes scribbled alongside the songs. Sometimes, I just casually read the Gitabitan. Have you heard of "Nakshikanthar math"? This book is a classic. This not a very hefty book, it's quite thin. This copy actually belonged to my father-in-law.

In Conversation with Prof. Nitis Mukhopadhyay

I was able to acquire this from his library after he passed away. Nevertheless, this is a novel that has been written as if it was a poetry. A novel written through a poem. The foreword of this book is also quite special. The foreword has not been written by the author. It has been written by Abanindranath Thakur. This small write up is a masterpiece. Sometimes, I sit with it just to read the foreword and end up memorizing it. The writing of Abanindranath is fluid, it seems to paint a picture. The foreword is of just 10 lines where he has introduced Jasimuddin, who was not that well known at that time. He introduced him in the Bengali Literature. Then, we have Jibananda Das. You should read some of his works. Subhash Mukhopadhyay is another one of my favourites. I love Joy Goswami's writing as well. His book, Ranaghat Local, is beautifully written. His grasp on the Bengali language is quite envious.

Another person we recently lost is Rituporno. He is the only person, in my opinion, belonging from Kolkata whose grasp on English and Bengali is equally good. Going back a little more, we had Samaresh Basu. There's this book he wrote that gained fame a few years back titled, "Dekhi Nai Phire". Have you read it? It's a huge book. It's the life of Ramkinkar Beij in the eyes of Samaresh Basu. That novel remained unfinished because Samaresh Basu passed away. Later, Subodh Ghosh thought he will finish it with somebody else as the writer. But then he decided against it. You cannot write over the writings of Samaresh Basu. How will you finish that? Unfinished works are better left unfinished. I read "Dekhi Nai Phire" a lot. When I bought the book, it cost around Rs. 2,500 to Rs. 3,000.

In Conversation with Prof. Nitis Mukhopadhyay

The reason was not the novel itself, but also the paintings in it. Do you know whose paintings those are? They are created by Bikash Bhattacharjee, a stalwart from Calcutta. His colour composition is marvellous.

I almost forgot to mention one more person – Bibhutibhushan Bandyopadhyay. Do you know there are two Bibhutibhushans - Bibhutibhushan Bandhopadhyay and Bibhutibhushan Mukopadhyay. Both are superb authors. Bibhutibhushan's "Pather Panchali" , "Ashani Sanket" and Bibhutibhushan Mukhopadhyay's "Ranur Pratham Bhag" are masterpieces.

Q: We would love to read your compositions. Which one should we start with?

NM: A series is released in India named "Kobita Probashi". I send some of my writings there and then I forget. I have a fascination about writing on Indian people. I cannot write songs, but I can write about songs. I take any song, for example a Rabindrasangeet, and start interpreting it. I am inclined towards critical writing. "Jete Jete ekla pothe" is one of my favourites. I had written an interpretation on it a long time ago. At the end of this song, there is a very unique and extraordinary changeover. From that changeover, it can be understood how much he owned the Bengali language. You'll find at the end, the line "Probhat hobe Raati" is at a very high scale. What is the meaning of this line? The song comprises storms - a lost poet; the storms are showing him the path to salvation, and so on. The line 'Prabhat hobe raati' is an exceptional construction of words by Tagore. In literal sense, it might be difficult to interpret.

In Conversation with Prof. Nitis Mukhopadhyay

But when you look at the message hidden underneath, "Raati hobe Probhat"- dawn arrives at the end of the night, just like the poem "Runner". It is an excellent construction of words and music.

Q: Tell us about “Sur - o - Chhanda”.

NM: It happened around 25 years ago; we were here in this house for around 20 years. At that time, the local Bengali Associations used to have cultural programmes. But in those programmes, there was no creativity. They had flashy artists and bands. The quality was slighted very heavily. So, we thought of doing something on our own. We started teaching dance and music, our sons played instruments. With my wife, my two sons and I, it was almost a full orchestra. We also had some very talented young boys and girls. The first adventure that we had was when we created a production of ‘Shyama’ on-stage with live music. It was mainly produced for non-Indians. It was done in English - we gave the literature to everyone, and I produced a magazine for that day with lots of articles from all sorts of Tagore Scholars. So, Shyama was one of our first big productions. It was very successful, very fulfilling.

We did quite a few singing events. We took various kinds of poetry and performed a song along with the poetries which matched thematically. We did many more events like, ‘Rupashi Bangla’, ‘Jiboner Jolsaghore’ and so on. Then slowly, the boys moved after graduating. One time, at a function in India, I noticed a tabla player. During one of the breaks, I found him and complimented him on his tabla skills.

I didn't compliment anyone else from the choir - only the tabla player. I asked him to apply to US. He applied for his PhD in our university and got accepted. His name was Debanjan Bhattacharjee. You will find him in my CV; he wrote a lot of research papers with me. He didn't play tabla with anyone, because his teacher forbade him to do so. But he agreed to play only for us. He played with us for 4 years. After that he completed his studies and moved on with a job.

Q: We would like to have a parting piece of advice from you.

NM: Be good human beings. Always try to help someone. When you give something to somebody, you might think that you are losing something - as if you are giving something away. But that is not true. Give people hope and try to help people. These days, people talk about 'brain drain' in India. They are always talking about how much the country is losing. I don't believe there is such a thing as brain drain. Consider Amrita and Rajnandini. After 2 or 5 years, one of you plan to come here and the other go to Germany, and you build up your own lives there. Is that brain drain from India's point of view? It is not. Yes, India raised you, India paid some money for your education, but the same way your parents did that. But when you are somewhere else, it is not like you have lost touch with your origin. For example, you can sometimes come back to St. Xavier's College and give lectures there; you can go to Calcutta University, participate in Calcutta Triennial; you can invite some students from Calcutta University or St. Xavier's College to come work with you, like I asked Debanjan. Many of my students have also gone back to India. Saibal Chatterjee, who became the director of IIM, returned back to India.

In Conversation with Prof. Nitis Mukhopadhyay

He did his PhD with me some years ago, and he decided to go back. Therefore, the idea of loss-gain needs to be more critically analysed. The registration fee at Calcutta Triennial is at least 250 USD. But when somebody comes to the same conference from University of Mumbai or University of New Delhi or Cochin, their maximum registration fee is 1000 INR. So, in a very selfish way, I can ask you who is actually funding the Calcutta Triennial conference? Where is the fund coming from? Is that brain drain? I never take any money from any college or university when I give lectures. Is that brain drain? Even Rabindranath Tagore needed supporters. The nomination for his Noble Prize was done by a British. Yeates nominated him for the prize. Why didn't Jagadish Bose get the Nobel prize? Marconi won it instead. Now those of us from Indian origin may complain that Marconi stole it from Bose. Bose gave a lecture in astronomical physical society, Marconi attended it, copyrighted his work and got the prize. Jagadish Bose was a student of St. Xavier's College. You may say that he didn't try and hence didn't get the prize. The question is, why didn't he? The basic reason behind it is that no one pushed Bose; the Indians never do. That is why it is important to help people. Indians never help anyone. That is why any Indian who has won a Nobel prize, has got it because of western influence. The only exception is C.V. Raman. Thinking that Amartya Sen's winning the Nobel Prize is a loss for us isn't correct. I think we should celebrate the fact that a simple Bengali gentleman has won the prize. You shouldn't remember that he got it when he was at Harvard. So, there are several viewpoints on which people need to think about before talking about brain drain. Being a good human is always necessary.

In Conversation with Prof. Nitis Mukhopadhyay

Some Interesting Stories Shared by Prof. Mukhopadhyay

1. In ISI, we had a regular visitor from Japan. His name was Prof. H. Morimoto. He was a very decent person. He was also fluent in Bengali. And you know, from Baranagar, you have different bus routes, and it goes through Shyambazar and to Sealdah markets. He knew everything, he didn't need an escort. He knew every road. He used to do his own groceries. So, one day he was telling us this story. One evening, he went to Sealdah. He took a bus and carried a small bag. He knew the local roads, so why pay for taxi. It was summer and mangoes were in season; and he loved mangoes. So, he goes from this vendor to that vendor on the streets of Sealdah market- sellers shouting and calling out for customers. Someone kept calling, "Four mangoes for 1 rupee". He liked the look of the mangoes at that particular store. So, he wanted to buy the mangoes from that store. He pushed through the crowd and went in front. The vendor saw that a foreigner was approaching his shop and instantly changed his rate to four rupees per mango. He increased the price 16 times. What did Prof. Morimoto do? He walked up to the vendor and very grimly said, "Ami Bangla jani" (I know Bengali), and walked away.

2. I talked about Prof Ashok Maitra, student of David Blackwell. He was very smart. One time, he was attending a seminar, where the presenter showed a function and said, "This is continuous". Now, Prof. Maitra had a bad habit of saying 'absolutely'. For example, if we required a book from his shelf and we asked him if we can borrow it for a day, he would say, "Absolutely". So, he used the term 'absolutely' after everything. So, in the seminar, when the person said 'continuous', Prof. Maitra said "Absolutely". And the speaker said "No, no - this function is continuous, and it is obvious" and Ashok Maitra again said, "Absolutely". This was repeated four to five times and it didn't get to a solution. No one was able to understand what happened. Later, we realised that the speaker thought that Prof. Maitra was suggesting that the function would be 'Absolutely Continuous' and he was saying the function was continuous but not absolutely continuous. So, this was the story of 'Absolutely'.

3. Norbert Wiener- you must have heard of him, from the Wiener process. He was a top Mathematician and regularly visited ISI. Outside the campus, we had ISI's post office. Now, one day Prof. Wiener was in the post office, trying to fill up a long form given by the P.O. At the same time, another research scholar from ISI went there for some work. He noticed Prof. Wiener on his way in. After finishing his work, which took around 15 minutes, he was about to leave. On his way out, he saw Prof. Wiener still standing there with his form. So, this research scholar asked him " Prof. Wiener, you are not done yet?" And then, he suddenly started writing. What had happened was, he had forgotten his own name.

In Conversation with Prof. Nitis Mukhopadhyay

The first thing to fill up in the form was his name, so he couldn't even start. The moment someone called him by his name, he remembered it and started filling up.

4. This story is about Kolmogorov. Kolmogorov used to travel to Stockholm from Moscow. He never liked flying by plane. He was very scared of flights. He had a phobia. As a result, he only visited India once. But he used to travel long distances via train. He used to visit Stockholm for Cramer. This one time, Kolmogorov was delivering a lecture and Cramer sat in the front row, attending it. Cramer was the head of the Mathematical program of Stockholm during that time. No one sat in the next five rows. Most students maintained a respectful distance and sat right at the back. During that seminar, a PhD student, Ulf Grenander was given the responsibility to attend to Kolmogorov and his every need – he would pick him up from the station, arrange meals for him and so on. Now, during one of their interactions, one day Kolmogorov asked Ulf, "Ulf, do you like my lectures?" He replied "Yes sir, you are Prof. Kolmogorov. We are lucky to have you." So, then he said, "But then, why do none of you ever ask me anything?" To this, Ulf replied, "Sir, how can we ask you any questions? We are merely PhD students. To us, you are God. You are the founder of Probability theory. We don't know enough to ask you anything. That doesn't mean the students don't love you, they appreciate your presence very much." Cut to a few days later, Ulf asked Prof. Kolmogorov who were his best teachers when he was a student himself.

In Conversation with Prof. Nitis Mukhopadhyay

Kolmogorov said that his geometry teacher was David Hilbert and that he was excellent. So, Ulf asked Kolmogorov, " Did you ask Hilbert any questions?" Kolmogorov replied, "Are you mad? That is David Hilbert! I was just a student. He was God to me."

5. Everyone is familiar with the book – Gun, Gupta and Dasgupta. But there is another very famous book – Hogg and Craig. This Hogg was a funny man. He was in University of Iowa. He was a very competitive man, but very jovial, very friendly and would talk to anyone anytime. Whenever he had basketball games or volleyball games in his backyard – he always has to win. So, that's the backdrop. One time, he had a visitor in their department. They took the visitor to a nearby restaurant for some food and drinks. Usually when faculty members start talking, do you know what we talk about? Any guess? We talk about how poor the students are prepared. I gave you the secret. We talk about how badly they have been taught before and how unprepared they are. So, they were also doing the same. But Bob (Robert) Hogg will not take that lying down. He has to prove that his University is better than all other places. He gets up mid conversation, goes to the washroom and on his way back, called this young girl, a waitress. He could see that she was probably a Fresher in the University of Iowa. He called her and said, "When you come to our table next time, one of my colleagues will ask you a question. Don't worry about the question. Just say 'x^3'. Can you do that? Practice it with me a couple of times". He then prepared the girl and gave her a $10 bill as an advanced tip for her trouble. Then, he comes back to his table.

The girl comes with some water. Bob then suggests to his colleagues, "Why don't you ask her a question? If she can answer, then that's it – your hypothesis is wrong". They say, "no no, I don't want to embarrass her in front of so many people.". So, Hogg said, "okay, then I will ask". And he asks, "What is integration of $3x^2\,dx$?" ?" She answers, "$x^3 + c$". Bob Hogg almost fell from his chair. He just lost 10 dollars for nothing!

6. There is a very young kid, very naughty, very aggressive and arrogant. He disturbs the class all the time, especially maths class. Whatever question you give him, he solves it while others haven't even started. He finishes his work first and then pokes other students. On one such day, the teacher is struggling to tackle him when his sister came. She said, "Give him a very tough problem and make him sit at the back and he will be busy with it for some time". At that time his classmates were just learning how to calculate 1+1 or 2+5. And the teacher gave him a problem, that went 'Calculate the value of 1+2+3+4+....+100'. These are 5–6-year-old kids, so he will never finish that problem. The teacher is safe now. As she walks back to her table, the kid raises his hand and says "Miss, I have done it. The answer is 5050". The teacher is stunned. It was the right answer. What he did was, he wrote 1, 2, 3, ..., 100. He then reversed it and wrote in the next line 100, 99, 98, ..., 1. He added the lines and each term added up to 101. So, 100 times 101 divided by 2, that is 5050. This is a 5-year-old kid. He discovered arithmetic progression in his own way. This kid was Carl Gauss.

Column of ALUMNUS

The Forecaster's guide to the Future:

Introduction to why, what and how to apply the forecasting techniques to predict real-life problems

- Smaranya Dey, Lead Data Scientist, Target

As human beings, we are constantly worrying about the future. While writing this article down, I am watching the 2022 financial budget and thinking about its possible implications. The students are concerned about what questions will come in their next exam. In general, humankind is thinking about when the Covid pandemic is going away. Many years ago, fortune tellers had unique ways to tell the citizens what would happen in the future. Even these days, the sight of people queuing up in front of the astrologers' chambers is not uncommon. Overall, there is no reason to doubt the far-reaching urge of being able to know what's coming is indomitable. The industries are not immune to this urge either. It is absolutely in their interest to have forecasting systems to make better decisions.

Dear Readers, in this article, we will look at three questions-

Why do the industries want to know about the future of their domains, products and the overall market?

What are the forecasting techniques which we can apply across different areas?

And how can we utilise such techniques to gather meaningful insights?

While delving deeper into the above questions, we will refer to some industry case scenarios or examples for better understanding.

Knowing the Why?

"All organisations start with why, but only the great ones keep their why clear year after year."

Forecasting is paramount for any business to take financial and operational decisions proactively. Having a sense of upcoming changes in the market and having short-term and long-term goals at their disposal helps the companies build better strategies. Forecasting techniques cannot be on point every time, but it is better to employ them rather than being blindsided about the future. Forecasting is applicable in a wide range of domains- retail, finance, agriculture, tourism, and other customer-facing businesses. The sectors which are starting to see the importance of prediction are- energy, higher distance, music festivals, human resources etc. [1]

Demand forecasting is one of the pivotal elements in any supply chain organisation (for example, any retail chain with brick and mortar and online presence). They use forecasting techniques widely to get away to avoid extreme scenarios. What are these extremes? Maintaining the proper inventory of items for a retail chain is crucial for their business.

Column of ALUMNUS

Too few products can compel the customers to shop elsewhere, and too many products can increase operational, storage costs, and even wastage of items in case of perishables. Using appropriate forecasting methods can minimise the cost and help the retailer place an optimised assortment in the store. [2] Accurate forecasting allows the retailers to decide on inventory management, forecasting demand of the items, maintaining proper assortment on shelves of the stores as well as in the back-room, out of stock detection, resource allocation and eventually maintaining customer satisfaction. [3] The primary purpose of the distribution centres is to temporarily hold the goods to maintain the flow of the products to different stores. If we take Walmart as an example, Figure 1 shows us how various centres they place strategically to make it easier for the item movement to and from the stores (the returned goods go back to manufacturing facilities). [4]. The movement of goods depends on the lead time for each supplier or product, freight transit time, storage cost, seasons, trends, global events, etc. This scheduling and planning process need accurate forecasting to consider the past and present demand of products. [5]

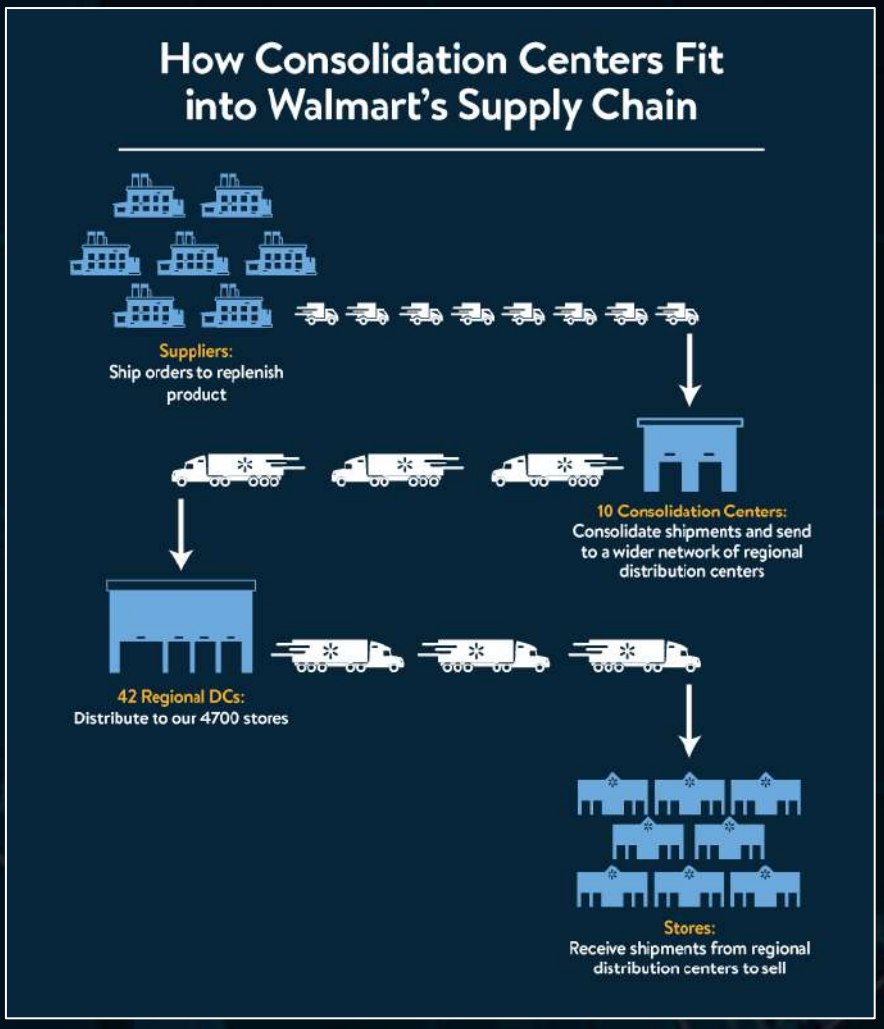


Fig 1: Importance of distribution centres in Walmart's Supply Chain

Planning the right assortment of products and seasonal items, promotional events etc. can be done properly if the retailers understand the customer demand and need. In 2017, the UK celebrated twenty years since the first book of the Harry Potter series, Harry Potter and the Philosopher's stone was first published. A successful large-scale campaign from their publishing house Bloomsbury saw a record-breaking sale in the house editions of the first book. [6]

Column of ALUMNUS

Keeping the labour demand optimised for e-commerce platforms or retailers with an omnichannel presence ensures the delivery of products to the customers' doorstep within the promised time. By omnichannel presence, we mean various shopping methods are accessible to the customers. Walmart in the United States has options of in-store shopping, shopping from website and mobile applications, placing an order and picking that order up from the nearest location etc. In 2019, Target reported an increase in pick-up sales of almost 50%. [7] To further the growth and customer satisfaction, maintaining the number of the in-store and delivery executives is of utmost importance.

Forecasting is equally crucial in management and supervision in the banking and finance sectors. Since banking assets and liabilities are influenced by economic and financial conditions, interest rates, prices of financial assets etc., proper econometric methods or models are employed to reduce forecast errors. [8] Forecasting techniques can also be helpful for investment decision making. Studies have shown that no single method can be applied uniformly across markets. Fundamental analysis to predict stock prices includes financial analysis of industries or companies. But, technical analysts use historical securities data and predict future stock prices based on the assumption that market forces determine stock prices and that history tends to repeat itself. [9]

Column of ALUMNUS

What's there in a Forecaster's toolbox?

While discussing forecasting, we need to understand time-series or time-stamped data. It is nothing but a sequence of data points gathered over time. For example, the monthly rainfall data in Kolkata for the last ten years, the weekly items sold in a store over the previous three years, the daily market closing price for a year, the daily step count captured by fitness trackers etc. Figure 2 illustrates the daily crude oil price for the last ten years.

$$y_t: \textit{Time series variable at the } t-\textit{th time point}, t = 0, 1, \ldots, N$$

Once we know what the time series is, we need to talk about its different components:

Trend

It results from long-term increases and decreases in the data.

Seasonal

Seasonal patterns are the short-term movements in the data which occurs because of seasonal factors. To think of examples, we see a sharp increase in apparels sales during Durga Pujo every year.

Column of ALUMNUS

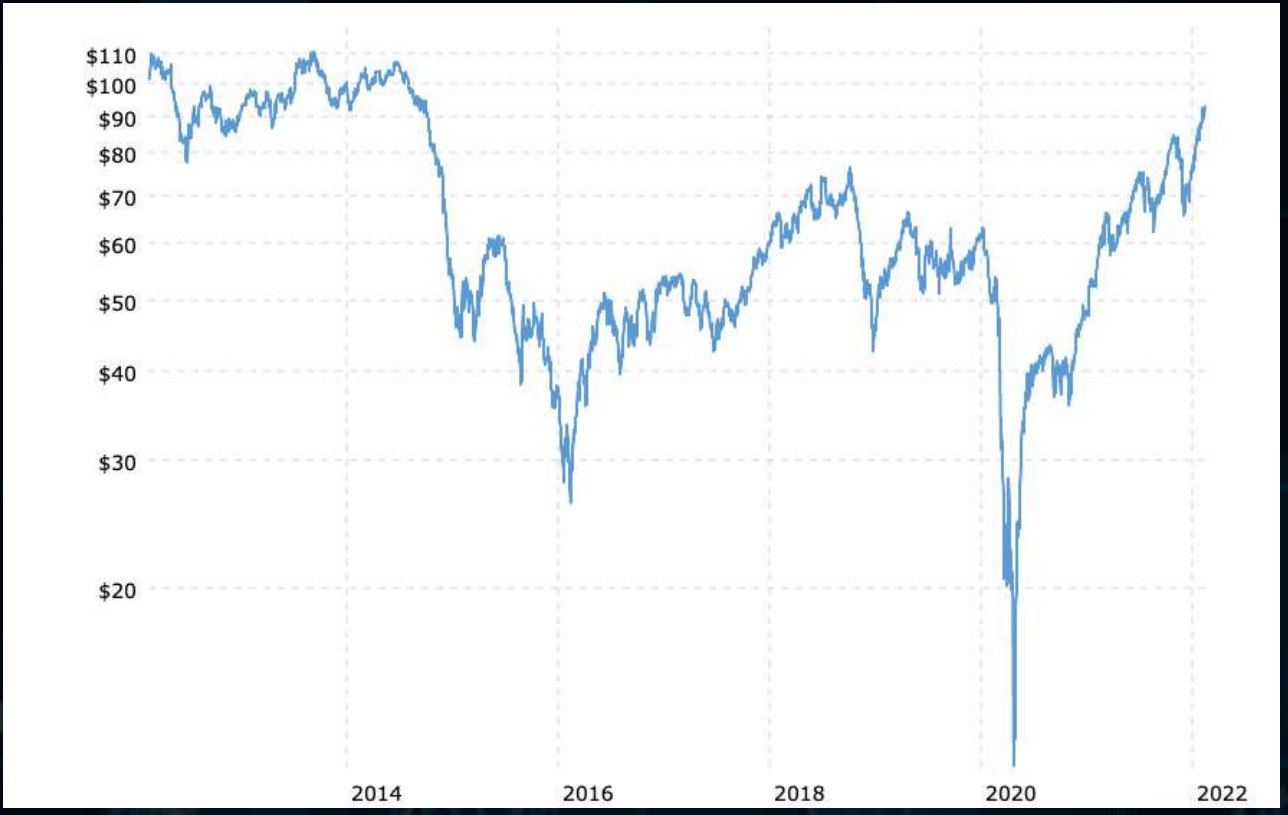


Fig 2: WTI Crude Oil daily price for 10 years

https://www.macrotrends.net/2516/wti-crude-oil-prices-10-year-daily-chart

Cyclic

A cyclical pattern happens when the data see rises and falls that are not fixed frequency. These oscillations extend for a longer time.

Irregular fluctuations

There are sudden changes in the time series, which are difficult to explain. These random variations are called residuals or random components.

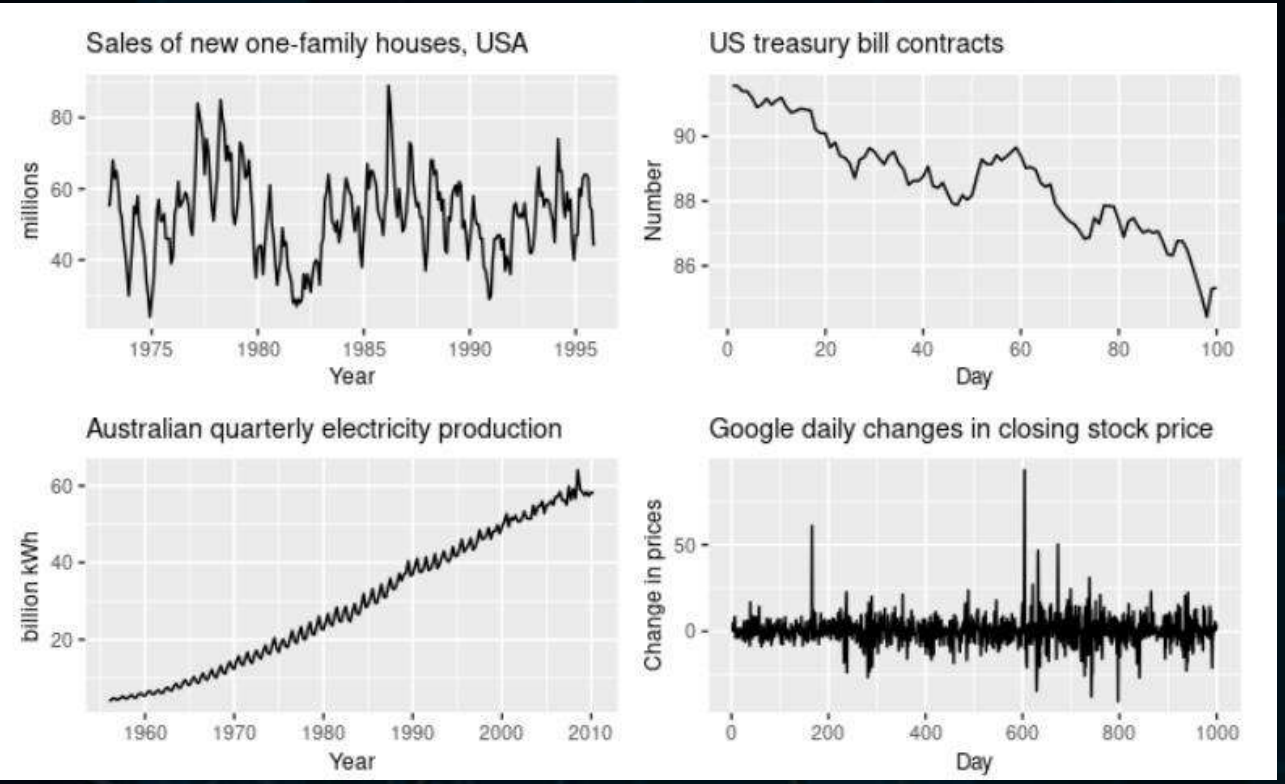


Fig 3: Different Examples of data exhibiting different patterns [13]

Figure 3 shows different components together in the time series data. The top-left figure is the monthly housing sales in the USA. There is no apparent trend in the data, but there is seasonality within each year and some cyclic pattern with about six to ten years. The top-right image shows the number of US treasury bill contracts over 100 days in a certain year. The data shows a clear downward trend. The low bottom-left image shows the Australian quarterly electricity production has a solid increasing trend pattern with seasonality. And the bottom-right image shows the Google daily changes in closing stock price over 1000 days that shows irregular random fluctuations.

Column of ALUMNUS

Autocorrelation

It measures the linear relationship between lagged values of a time series. Several autocorrelation coefficients are depending on the considered lag.

$$r_k : autocorrelation\ between\ y_t\ and\ y_{t-k}$$

If a time series data has a trend, the autocorrelations for the smaller lags tend to be large and positive, since the observations nearby in time are also nearby in size. If the time series is seasonal, the autocorrelation will be larger for seasonal lags (at multiples of seasonal frequency) than for other lags.

White noise

The time series that shows no autocorrelation is known as white noise. For such cases, autocorrelations for each lagged value are close to zero.

ACF Plot

The autocorrelation coefficients plotted to show the autocorrelation function is known as the ACF plot or correlogram. Figure 5 shows the ACF plot of monthly Australian electricity demand (Fig 4) corresponding to different lags. We can see a clear upward trend and seasonality present in the data from Figure 4 and that's why the ACF plot shows a slow decrease in the autocorrelations as lag increases with the scalloped shape present due to seasonality.

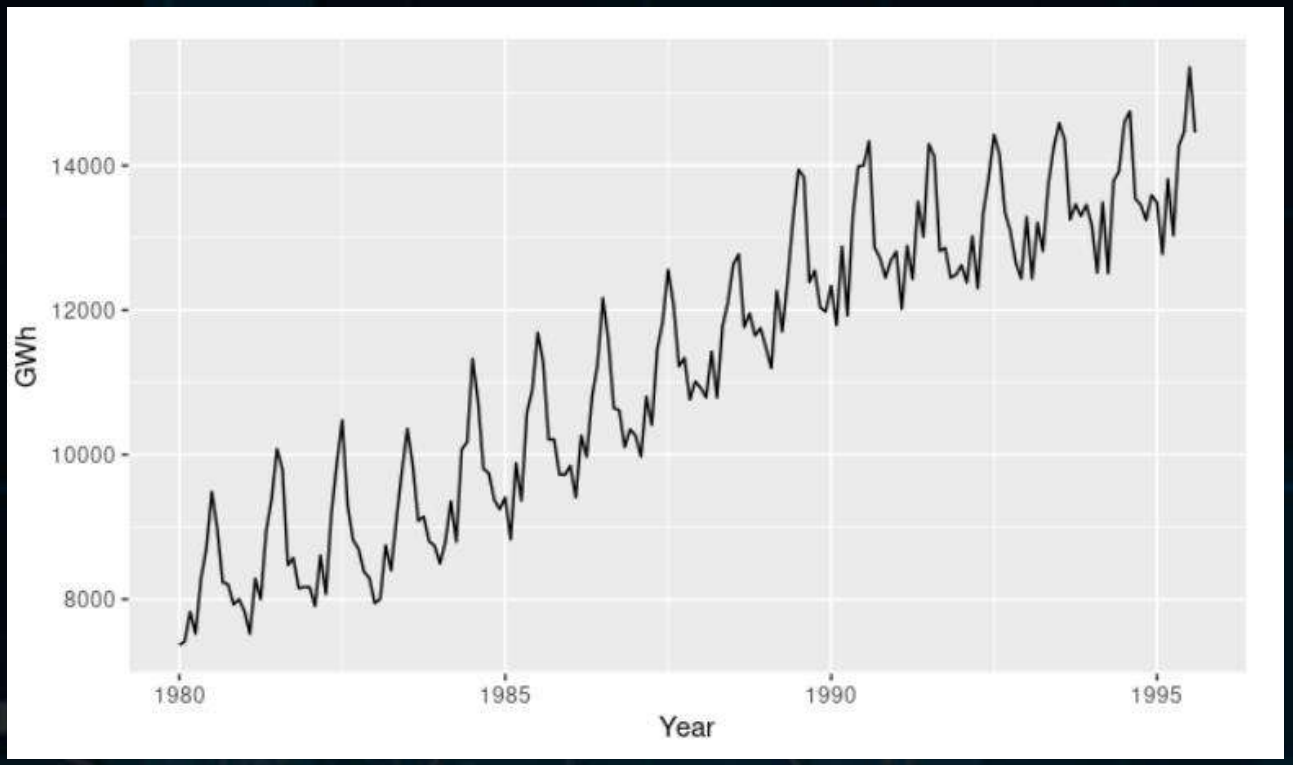


Fig 4: Monthly Australian Electricity Demand over fifteen years

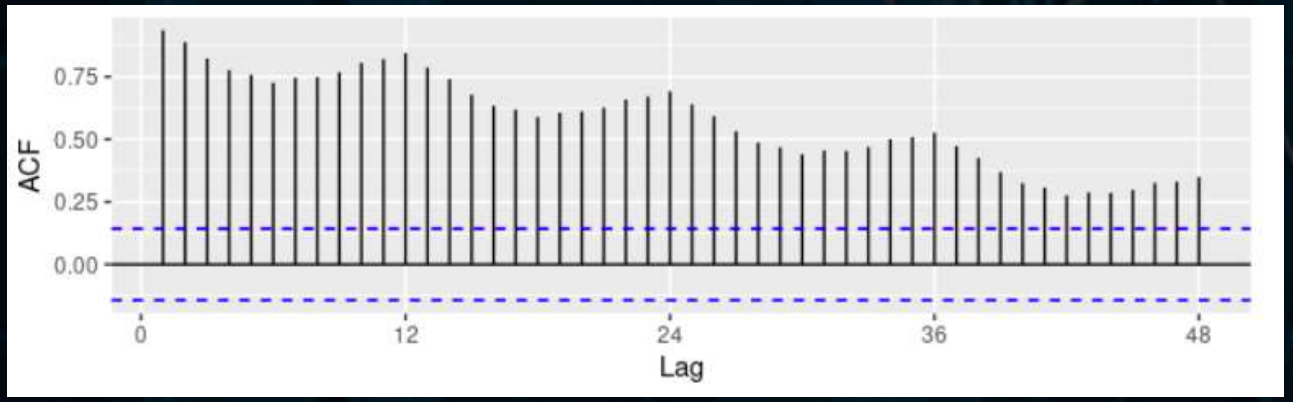


Fig 5: ACF plot for Fig 4. [13]

Column of ALUMNUS

Stationarity

A time series is stationary if its properties do not depend on the time at which the series is observed. So, time series with trend and seasonality patterns present in the data is not stationary. Such data has no predictable patterns in the long term. The plots will exhibit more or less horizontal patterns with constant variance and constant autocorrelation over time.

$If\ \{y_t\}\ is\ stationary, then\ \forall s, the\ distribution\ of\ (y_t, \dots, y_{t+s})\ does\ not\ depend\ on\ t$

Differencing

If the time series is non-stationary, we can compute the difference between consecutive observations to make it stationary. It helps stabilise the average of the time series by reducing the effects of trend and seasonality.

Unit root tests

These are statistical hypothesis tests to check if the time series is stationary or not. One such test is the Augmented Dicky Fuller Test, where the null and the alternate hypothesis are as below:

$$H_0: the\ time\ series\ is\ non-stationary$$

$$H_1: the\ time\ series\ is\ stationary$$

If the p-value obtained from the test is less than some significance level (0.05), then we can reject the null hypothesis concluding that the time series is stationary.

Column of ALUMNUS

Training and Test Data

The accuracy of the forecasting model can be found out by applying the same model on a new data set that was not used for the model fitting step. Before going to forecast error (or accuracy) metrics, let us first talk about the training and validation set. It is common practice to keep away a portion of the data (around 20%) while modelling which is known as validation set or test set (also referred to as hold-out-sample or out-of-sample data) and the remaining chunk is called the train set (also known as in-sample data). We build the model on the train set and apply the same on the test set to check the accuracy of the model. Since the test set is not getting used for model building purposes, it should show how well the model performs to forecast new data.

Forecast horizon

It is the prediction length that the model can make forecasts for. For example, the store manager wants to know the sales forecast for the yogurt category for the next three weeks or a tourism company wants to evaluate their strategies based on the monthly tourist incoming at a particular location next year.

Column of ALUMNUS

Forecast error

It is the difference between an observed value and the forecasted value, which means the unexplained part of the observation.

$$e_t = y_t - \acute{y}_t$$

The most commonly used error measures are:

$$Mean\ Absolute\ Error\ (MAE) = mean(|e_t|)$$

$$Root\ Mean\ Squared\ Error\ (RMSE) = \sqrt{mean(e_t^2)}$$

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = mean(\left|\frac{e_t}{y_t}\right|)$$

Using MAPE will lead to infinite or undefined value when y_t is zero. There are alternative measures that can be used in such cases like symmetric MAPE etc.

Prediction Intervals

It gives an interval within y_t is expected to lie with a certain probability. A 95% (assuming normality) prediction interval for an h-step forecast is as below,

$$\widehat{y_{T+h|T}} \pm 1.96\ \widehat{\sigma_h}$$

$\widehat{\sigma_h}$ is the standard-deviation of the h-step forecast distribution.

Column of ALUMNUS

ARIMA Models (Auto-Regressive Integrated Moving Average)

In an autoregressive model, we forecast the time series using its past observations of the variable (regressing using the linear combination of the values of earlier time points of the same variable). Thus, an autoregressive model of order ρ can be described mathematically as,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \varepsilon_t \; is \; white \; noise$$

We refer to this as the AR(ρ) model, autoregression of order ρ.

In a moving average model, we use a linear combination of past forecast errors in a regression-like model.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \varepsilon_t \; is \; white \; noise$$

We call this an MA(q) model, moving average model of order q.

If we combine these two methods with differencing, we obtain an ARIMA(ρ,d,q) model where ρ is the order of the autoregressive part, q is the order of the moving average part and d is the degree of first differencing involved.

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \varepsilon_t \; is \; white \; noise$$

ARIMAX Models (Auto-Regressive Integrated Moving Average with Explanatory Variables)

This particular model is nothing but an ARIMA model with external independent numerical and/or categorical variables. Including the exogenous variables adds complexity to the model to capture the influence of external influences and the underlying pattern of the time series. [11]

Exponential Smoothing Models

This method gives the forecasts produced by weighted averages of past observations, with weights decaying exponentially as the observations come further from the past.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-1} + \ldots$$

Where, $0 \leq \alpha \leq 1$, is the smoothing parameter. If α is small, more weight is given to the observations from the distant past.

Neural Network-based Models

These kinds of models help us unfurl the complex non-linear relationships between the time series data and the predictors. Here, we use lagged values of time series as the inputs to the neural network, just as we use the lagged values in the autoregressive models.

Bayesian Structural Models

A more transparent model to forecast the future values than ARIMA is the Bayesian structural time series model because it does not rely on differencing, lags or moving averages. We write a Bayesian structural model as,

$$Y_t = \mu_t + \beta x_t + S_t + e_t, e_t \sim N(0, \sigma_e^2)$$
$$\mu_{t+1} = \mu_t + \upsilon_t, \upsilon_t \sim N(0, \sigma_\upsilon^2)$$

Here x_t is the set of regressors, S_t is seasonality and μ_t is local level term which represents the evolving latent stage (referred to as unobserved trend). Furthermore, these models have the ability to reduce a large set of correlated variables using spike-and-slab priors. The spike part takes care of the probability of choosing a variable (having a non-zero coefficient) and the slab part shrinks the non-zero coefficients towards zero. [12]

Column of ALUMNUS

How to formulate the model structure and gather insights?

We often get the problem statement from the business stakeholders. It is essential to clarify a few points before starting a project, like

- Are there any business constraints?
- Is there any hierarchy we need to consider before diving into the modelling?
- What do we want to forecast?
- How much accuracy is necessary from the process?
- How much historical data do we have with us?
- What forecast horizon should we take?

The most standard practise to compute forecasts is from a common origin based on past data. Given y_1, y_2,..., y_N, N being the number of past observations, we will use any one of the multiple forecasting models to predict y_{N+1}, y_{N+2}, ..., y_{N+h}, h being the forecast horizon. The benefit of using more than one model is that different techniques can capture the various underlying pattern of the time series. ARIMA or ARIMAX model can understand the linear pattern in the data, whereas the neural network-based method will look into the non-linearity of the time series. To choose the best out of all the applied models, one can compute the error measures like MAPE or MAE on the out-of-sample set and choose whichever is giving the lowest error.

We can also take an ensemble approach to combine the forecasts; a weighted average of the multiple forecasts, weights being the inverse of MAPE. Since MAPE is an error measure, whichever forecast has the lower MAPE will get higher weightage in this way. Often the peaks and troughs observed in the time series can get deviated from our assumption of normality. In such cases, we can structure the model framework in multiple stages. The initial stages can contain the application of various forecasting models. The later stages can include modelling the error or residual part obtained from the initial stages or separately trying to detect the spikes and dips in the data.

It will be an excellent time to see some relevant R or Python packages. I feel R is way superior when it comes to forecasting. Rob J. Hyndman's 'forecast' library has a plethora of functions to perform modelling exercises like ARIMA, neural network-based non-linear models, exponential smoothing etc. Packages like ‘prophet’ (for both python and R) and ‘Kats’ packages, developed by Facebook are now widely popular. While applying forecasting in the industry, one should consider the scale of the problem. The omnichannel marketplace has billions of different items sold to customers. For example, Walmart has over 4500 stores in the US, and Target has over 1500 stores. Indian retailers like Future Group and Reliance industries show a steady ascending pattern. Forecasting at an item level can take an enormous amount of time which is costly. To reduce the time of mode execution, we need to seek refuge in modern tech-stack like Apache Spark.

Column of ALUMNUS

Is the list of techniques mentioned in this article exhaustive? Absolutely no. There are many other use cases related to predicting the future, which we could not even mention in this article. However, the intention was to make the readers aware of the various business problems, which nudges the urge to look into the future and apply the forecasting algorithms to carry out the same. I listed down the information on the books, blogs, package information below in the reference for you to have a look at them.

References:

1. *https://www.kdnuggets.com/2019/05/6-industries-warming-up-predictive-analytics-forecasting.html*
2. *https://www.michiganstateuniversityonline.com/resources/supply-chain/why-forecasting-is-essential-in-supply-chain-management/*
3. *An Intelligent approach to demand forecasting, Enterprise Information Management, Phillips Lighting, Bangalore, Proceedings of the 2nd International Conference on Inventive Computation Technologies (ICICT 2017)*
4. *https://corporate.walmart.com/newsroom/2019/01/14/high-tech-consolidation-center-set-to-open-in-july-adding-efficiency-to-walmarts-supply-chain*
5. *https://houston.ascm.org/blog/id/3*
6. *https://www.wizardingworld.com/news/500-million-harry-potter-books-have-now-been-sold-worldwide*
7. *https://diginomica.com/digital-growth-slows-target-stays-target-ongoing-omni-channel-transformation*

References:

8. *Economic Forecasting for Banking- Course, Massimiliano Marcellino, Florence School of Banking and Finance*
9. *Evaluation of forecasting methods from selected stock market returns, M. Mallikarjuna and R. Prabhakar Rao, Financial Innovation 5, Article Number 40, 2019*
10. *https://www.shopify.in/retail/demand-forecasting*
11. *Building ARIMA and ARIMAX Models for predicting long-term disability benefit application rates in the public/private sectors, Bruce H. Andrews et. Al., 2013 Society of Actuaries*
12. *https://multithreaded.stitchfix.com/blog/2016/04/21/forget-arima/*
13. *Another look at measures of forecast accuracy, Rob J Hyndman, Anne B Kohelar*
14. *Another look at forecast-accuracy metrics for intermittent demands, Rob J Hyndman, June 2006 Issue 4 Foresight*
15. *Forecasting Principles and Practice, Rob J Hyndman and George Athanasopolous*
16. *https://kourentzes.com/forecasting/*
17. *https://pkg.robjhyndman.com/forecast/*
18. *https://engineering.fb.com/2021/06/21/open-source/kats/*
19. *https://facebook.github.io/prophet/*

Student ARTICLES

‘Women and children must be saved first’- a statistical analysis of the order that was passed on the sinking of Titanic ship

- Vidhi Shah (3rd Year)

The ill-fated Titanic that sank on April 15, 1912 was one of the largest and most technologically advanced ships in the world. The ship took 2 hours and 40 minutes to sink after hitting the iceberg during which it is said that the crew passed the orders that women and children must be saved first.

Since then, extensive studies have been carried on the Titanic. One such study was conducted and presented in- Dawson, Robert J. MacG. (1995), The ‘Unusual Episode’ Data Revisited. Journal of Statistics Education, 3. doi: 10.1080/10691898.1995.11910499.

The findings of this report can be accessed in the dataset called ‘Titanic’ in the R software.

Student ARTICLES

The data on survival of passengers classified according to their age is again presented in the 2x2 contingency table below-

SURVIVAL / AGE	YES	NO	TOTAL
ADULT	654	1438	2092
CHILD	57	52	109
TOTAL	711	1490	2201

Again, any association between the two attributes is checked for. Using the Pearson's chi-square statistic, it is observed that there is a strong association between survival of a passenger and his/her age as $\chi^2 = 20.96$ for the above data.

The odds of survival of a child comes out to be 1.04 while the odds of survival of an adult was just 0.455.

This makes the odds ratio equal to 2.29 implying that the odds of survival for a child was 2.29 times the odds of survival for an adult.

Similarly, the data on survival of passengers classified according to their gender is presented in the 2x2 contingency table below-

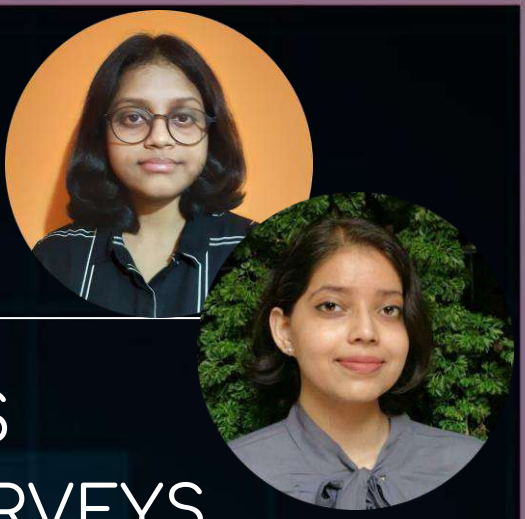
SURVIVAL / AGE	YES	NO	TOTAL
FEMALE	344	126	470
MALE	367	1364	1731
TOTAL	711	1490	2201

First, any association between the two attributes is checked. Using the Pearson's chi-square statistic, it is observed that there is a very high association between survival of a passenger and his/her gender as χ^2=455.33 for the above data. Moving further, the odds of survival of a female comes out to be 2.73 while the odds of survival of a male was merely 0.269. This makes the odds ratio equal to 10.14 implying that the odds of survival for a female was 10.14 times the odds of survival for a male.

CONCLUSION:

In the light of the given data, odds of survival of a child was higher than that of an adult and even amongst adults, the odds of survival of females was significantly higher than that of males. The responses to the orders are reflected in the above analysis. The notion of 'chivalry at sea' played a role as not everyone aimed to maximize their own well-being. The odds of survival would have been in favor of the fittest had people not acted upon the orders of the captain.

Student ARTICLES



HOW SAMPLING ERRORS PRESENT IN SAMPLE SURVEYS MAY MISLEAD VOTING RESULTS

- Shamie Dasgupta (3rd Year), Anuska Mukherjee (3rd Year)

When we survey a sample, our interest generally goes beyond just the people in the sample.

Moreover, we try to get information from the sample to project it onto a larger population. Hence utmost care is needed in both the design and the implementation of a sampling process to avoid biases in the end results. Simply selecting a sample where it is convenient may be logically easier, but it might not fairly represent the population of interest. Hence for this reason, it is very important for a surveyor to know the typical sampling errors so that one can avoid them while carrying out a survey.

In this article we discuss two such significant sampling errors that resulted in a hugely catastrophic mishap created by The Literary Digest in the 1936 presidential election between President Roosevelt and Landon.

The Literary Digest was a highly influential magazine in America. The weekly magazine primarily focused on opinion articles and analysis of current events.

Student ARTICLES

Literary Digest was renowned for its opinion polls, and it had correctly predicted the outcome of the previous five presidential elections. In 1936, the poll conducted by "Literary Digest" predicted that Landon would win the presidential election in a landslide with more than 57% of the popular vote.

However, the actual results of the election revealed that Roosevelt had actually won the election against Landon!

The enormity of the magazine's error destroyed its credibility and the magazine closed down within a few months of the election. Now, the only reason Literary Digest is remembered today is due to its unfortunate demise that resulted due to a wrongly executed survey for the 1936 election polls.

WHAT WENT WRONG WITH THE MAGAZINE'S POLL

There were two basic causes of the Literary Digest's downfall: selection bias and non-response bias.

- Selection bias

  First and foremost, the Poll method had an incorrect sampling frame. The researchers of the magazine surveyed over 2 million people, but they were not randomly chosen, which resulted in a huge selection bias.

- The entire sample selection process was non- probabilistic since the magazine had chosen its sample from a mailing list based on every telephone directory in the country, list of magazine subscribers, rosters of clubs and associations. The list had around 10 million names. Each person on the list was mailed a ballot and asked to mail back the marked ballot to the magazine. Out of the ten million names, only 2 million mailed back, and only their responses were accepted in the sample. In 1936, telephone was considered more of a luxury than a necessity. Most of the people having telephone lines or club memberships were in middle class or upper class. Hence, the list unknowingly excluded the voters belonging to the lower income group. The effect of a biased target population blew in quite badly in the polling results.

- Non-response bias

  - Another problem with the magazine's poll was the non-response of the survey. Only 2.4 million out of 10 million responded to the survey, which is almost 1/4th of the intended population.

  - Before selecting a sample, the population frame needs to be divided into parts that are called sampling units. Suppose that unit i, provided it is included in a sample s, responds with probability q_i, q_i not depending on s or $\underset{\sim}{Y} = (Y_1, \ldots, Y_N)$.

  - Suppose n units are drawn by SRSWOR (Simple Random Sampling Without Replacement) and define:

M i = 1 if unit i is sampled and responds

= 0 otherwise

Consider the arithmetic mean $\bar{y} = \frac{\sum_1^N M_i Y_i}{\sum_1^N Y_i}$ of all observations as an estimator of $\bar{Y}$. Then $E(M_i) = \frac{n}{N} q_i$ and $E(\bar{y})$ is asymptotically equal to $\frac{\sum q_i Y_i}{\sum q_i}$ .The bias $\sum(\frac{q_i}{\sum q_i} - \frac{1}{N}) Y_i$ is negligible only if approximately $q_i = \frac{1}{N}\sum q_i$.
Even if the last equality holds for $i = 1, 2, \ldots, N$, the variance of y is inflated by the reduced size of the sample of respondents. So, it behoves us to pay attention to the problem of non-response in sample surveys.

ADJUSTMENTS MADE BY GALLUP TO PREDICT THE CORRECT POLLING RESULT

It was Dr. George Gallup who revolutionized the faulty 'straw polls'. After countless research and experiments, Gallup invented the Gallup Poll which was a more scientific and systematic polling technique, that used a lot of sample theory.

His sampling procedures included the following stages:

Student ARTICLES

STEP1: SELECTION OF PRIMARY SAMPLING UNITS (PSUs)

The first stage of sampling is the identification of PSUs, consisting of clusters of households. Gallup stratified the U.S. electorate into precise number of demographic groups, which he then proportionally allocated in his sample group of 50,000 people.

STEP2: SELECTING HOUSEHOLDS

Gallup used random-route procedures to select sampled households. Unless an outright refusal occurs, interviewers made up to three attempts to survey the sampled household.

STEP3: SELECTING RESPONDENTS

Gallup implemented quality control procedures to validate the selection of correct samples and that the interviewer selected the correct person in each household.

By making these adjustments, even if Gallup could not make his results completely error free, his poll did predict the correct winner of the election and that too with a smaller sample size than the magazine.

Student ARTICLES

References:

1. *Qualtrics // October 12. (2020, October 7). The 1936 election - a polling catastrophe // Qualtrics. Qualtrics. https://www.qualtrics.com/blog/the-1936-election-a-polling-catastrophe/.*

2. *(2021). Coursehero.com. https://www.coursehero.com/file/25352049/2014-slides3*

3. *1936 election: Case Study I. (n.d.). Www2.Math.upenn.edu. https://www2.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html*

4. *Landon in a Landslide: The Poll That Changed Polling. (n.d.). Historymatters.gmu.edu. Retrieved May 29, 2021, from http://historymatters.gmu.edu/d/5168/*

5. *The Literary Digest. (2020, December 31). Wikipedia. https://en.wikipedia.org/wiki/The_Literary_Digest*

6. *How One Man Used Opinion Polling to Change American Politics. (n.d.). Time. https://time.com/4568359/george-gallup-polling-history/*

Student ARTICLES

A Dive into Categorical Data Analysis with a Real-Life Study

- Saptarshi Chowdhury (3rd Year)

Introduction:

Categorical Data Analysis is one of the most important topics in Applied Statistics, and it has been widely used since ages in numerous domains to draw conclusions or even infer about a bigger issue. Be it, data related to gender discrimination or even in the covid-19 pandemic, this topic has helped Statisticians numerous times to deal with corresponding pertinent data. So, in this article, a brief introduction of CDA is given, alongside a renowned real-life problem. Before going into the problem, here are a few extremely important terms that need to be known compulsorily:

- Categorical Data: Data, where information is classified into multiple categories or classes, is called Categorical Data. For example, Political Ideology has several categories such as Liberal, Moderate, Conservative: hence making it a Categorical Variable.[1]

Student ARTICLES

- Response and Explanatory Variable: The focus of a particular question in any statistical study is called Response Variable or Dependent Variable, denoted conventionally by Y. As the name suggests, Explanatory Variable is one that explains or demonstrates the change in the Response Variable. It is also called Independent Variable and denoted by X. For example, a model for Political Ideology can have certain explanatory variables such as Income Status, Education Status, Gender, Race, Age.[1]

- Nominal and Ordinal Data: Categorical Variables that cannot be ordered in any objective way are referred to as Nominal Data. For example, Gender consists of Binary and Non-binary, and they cannot be ordered in any way. On the other hand, Categorical Variables that can be ordered are called Ordinal Data. For example, the Severity of a Disease can be categorized into Mild, Moderate, Extreme: which can be ordered in ascending order. [1]

- 2*2 Contingency Table: Sometimes, data can be given to us in a way that can be rather difficult to interpret if not arranged properly. Hence, the renowned Statistician Karl Pearson introduced a 2-way or 2*2 Contingency Table, which is basically a 2D Array containing the joint distribution of two Categorical Random Variables.[2]

- Prospective and Retrospective Study: The study that generally involves studying a cohort of subjects for a certain period forward in time is called Prospective Study, while the study that involves looking backwards in time for a similar purpose is called Retrospective Study.[2]

- Experimental and Observational Study: As the name suggests, the study where the investigator has control over which subjects enter each group is called Experimental Study, while study where it is observed which subject chooses which group and who has the outcome of interest is called Observational Study. Clinical Trial is an Experimental Study, whereas Case-control, Cohort and Cross-sectional are all Observational Studies.[2]

There are more such terms that need to be known, but these few are the pillars of what is going to be discussed later. Now, let us directly dive into a very interesting real-world problem.

Source: 'The Work of Ritzel' in the paper "The Significance of the Evidence about Ascorbic Acid and the Common Cold" published by Dr. Linus Pauling, Department of Chemistry, Stanford University on August 9, 1971. [3][4]

Description: In 1961, G. Ritzel, a physician with the medical service of the School District of the City of Basel, Switzerland carried out a study on whether 1g of Ascorbic Acid per day helps in reducing mortality. It was conducted in a ski resort with 279 skiers during two periods of 5-7 days.

Student ARTICLES

The conditions were such that the incidence of colds during these short periods was large enough (about 20%) to permit results with statistical significance to be obtained. The subjects were roughly of the same age and had similar nutrition during the period of study, and the investigation was double-blind. Out of 140 subjects who were given placebo, 31 suffered cold, and out of 139 subjects who were given ascorbic-acid, 17 suffered cold as reported over the course of the study.

Here is a 2*2 Contingency Table based on the above information:

Common Cold Status(Y) / Dose (1 mg/day) (X)	Yes (Y=1)	No (Y=0)	Total
Ascorbic Acid (X=1)	17	122	139
Placebo(X=0)	31	109	140
TOTAL	48	231	279

Student ARTICLES

Comment on the type of study:

Here is what we can comment about the type of study:

1. It is an Experimental Study since an intervention has been introduced by the researchers, and further effects have been studied. In other words, it is a case of Clinical Trial as the effect of Ascorbic Acid (Vitamin C) in reducing mortality is to be checked.

2. Furthermore, it is Prospective in nature because the skiers were first administered Ascorbic Acid and Placebo accordingly, and then observed for a period of time to check mortality, thus being forward in time.

Statistical Explanation:

Here, let Y denote the Common Cold Status. It is a binary variable, such that, it takes the value 1 when the subject has contracted cold, 0 otherwise. Also, let X denote the Treatment Status. It is a binary variable as well, such that, it takes the value 1 when the subject has been administered Ascorbic Acid, 0 otherwise.

Now, we can further use a very important measure in order to draw a major conclusion from the data given. For that, we need to compute two probabilities from it.

Let p_1 denote the probability that the subject has contracted common cold given that Ascorbic Acid was administered to him/her, and,
Let p_2 denote the probability that the subject has contracted common cold given that Placebo was administered to him/her.

From the data collected by the researcher/s, we have,

$$p_1 = PrPr\,(Subject\ has\ contracted\ common\ cold \mid Subject\ was\ administered\ Ascorbic\ Acid)$$

$$=> p_1 = PrPr\,(Y = 1 |\ X = 1) = \frac{PrPr\,(\widehat{Y = 1,\ \ X = 1})}{PrPr\,\widehat{(X = 1)}}$$

$$[By\ the\ def^n\ of\ Conditional\ Probability]$$

$$=> p_1 = \frac{(\frac{17}{279})}{(\frac{139}{279})} = \frac{17}{139} \cong 0.1223$$

Also,

$$p_2 = PrPr\,(Subject\ has\ contracted\ common\ cold \mid Subject\ was\ administered\ Placebo)$$

$$=> p_2 = PrPr\,(Y = 1 |\ X = 0) = \frac{PrPr\,(\widehat{Y = 1,\ \ X = 0})}{PrPr\,\widehat{(X = 0)}}$$

$$[By\ the\ def^n\ of\ Conditional\ Probability]$$

$$=> p_2 = \frac{\left(\frac{31}{279}\right)}{\left(\frac{140}{279}\right)} = \frac{31}{140} \cong 0.2214$$

Here, instead of taking difference, we can take the ratio of proportion which is called Relative Risk.

$$Relative\ Risk = R.R. = \frac{p_2}{p_1} \cong \frac{0.2214}{0.1223} \cong 2$$

Interpretation:

Here, the proportion of subjects contracting common cold when administered Placebo is almost twice that of the proportion of subjects contracting common cold when administered Ascorbic Acid. So, in the light of the given data, we can say that a subject is less likely to contract common cold if he/she had been administered Ascorbic Acid. So, Ascorbic Acid seems to reduce the risk of contracting common cold.

Conclusion:

This was a rather brief explanation of Categorical Data Analysis and how it can be used to interpret data using very simple measures from Mathematics point of view. However, one thing to be kept in mind is misinterpretation. Numerous sources can use incorrect proportions from the same data and try to draw conclusions that can be misleading. Hence, as I have always said, we need to keep our eyes open!

References:

1. *Agresti, A. (2018). An introduction to categorical data analysis. John Wiley & Sons.*
2. *Stat 8620, categorical data ... - university of Georgia. (n.d.). Retrieved December 23, 2021, from https://faculty.franklin.uga.edu/dhall/sites/faculty.franklin.uga.edu.dhall/files/lec1(1)_0.pdf*
3. *Ritzel, G., Helv. Med. Acta, 28, 63 (1961).*
4. *Pauling, L., Proc. Nat. Acad. Sci. USA Vol. 68, No. 11, pp. 2678-2681, November 1971*

Student ARTICLES

EQUALITY OF PARAMETER ESTIMATES OBTAINED BY TWO METHODS: A THEORETICAL DISCUSSION

- Utsyo Chakraborty (3rd Year)

There are several ways by which one can estimate unknown parameters in a population. Two of the most widely used methods are those of Maximum Likelihood and Method of Moments. The former aims at maximizing a "likelihood function" (so that under the given model structure the observed data is most probable); thus, the maximum likelihood point estimate of the parameter is that value at which this maximization occurs. The latter on the other hand, tries equating sample moments with their population counterparts, using the fundamental assumption that the sample is an adequate representation of the larger unknown population at hand.

Although the method of moments bears historical significance, it has been observed that maximum likelihood estimates are usually more reliable in practice. However, one may ask, "can the two methods ever yield equal estimates?" They technically can, thinking intuitively. We will explore here a special situation in which if the population is said to follow a certain type of probability distribution, the two estimates will be exactly identical.

Distributional Setup and Assumptions:

Let us consider a discrete random variable X. X is said to follow a power series (related to g) distribution if its probability mass function is given as:

$f_\theta(x) = \frac{a_x \theta^x}{g(\theta)}$; $x \in \mathbb{W}, \theta \in [0, r)$ [where 'r' may be chosen as per our convenience]

We can easily see that the general structure of the distribution comes from that of the power series $g(\theta) = \sum_{n=0}^{\infty} a_n \theta^n$.

It can be shown that some of the popularly used discrete distributions viz. Binomial, Poisson, Geometric, Negative Binomial etc. are special forms of the power series distributions.

The Problem

Let a random sample of size n be sampled from a general power series distribution with the form of the p.m.f. as given above. It should be noted that a_x=0 may be a possibility for some values of x.

Thus $(X_1, X_2, \dots, X_n) \sim (iid) f_\theta$ where $f_\theta(x) = \frac{a_x \theta^x}{g(\theta)}$; $x \in \mathbb{W}, \theta \in [0, r)$.

We can now construct the likelihood function which is given by:

$$L = \prod_{i=1}^{n} f_\theta(x_i)$$

Taking logarithm on both sides, we get the log-likelihood as:

$$\log L = \sum_{i=1}^{n} \log f_\theta(x_i)$$

$$= \sum_i \log a_x + \log(\theta) \sum_i (x_i) - n \log g(\theta)$$

Maximizing the log likelihood, we get:

$$\frac{\delta \log L}{\delta \theta} = \frac{\sum_i (x_i)}{\theta} - \frac{n\, g'(\theta)}{g(\theta)} = 0$$

$$\Rightarrow \frac{\sum_i (x_i)}{n} = \frac{\theta\, g'(\theta)}{g(\theta)} \quad \ldots (1)$$

The method of moments estimate of θ can be computed from the sample by equating population mean to the sample mean (given by $\frac{\sum_i (x_i)}{n}$).

So, $E(X) = \sum_{i=0}^{n} x\, f_\theta(x) = \sum_i x\, \frac{a_x \theta^x}{g(\theta)}$

By the elementary properties of a probability mass function:

$$\sum_{i=0}^{n} f_\theta(x) = 1 \Rightarrow \sum_{i=0}^{n} \frac{a_x \theta^x}{g(\theta)} = 1$$

$$\Rightarrow \sum_{i=0}^{n} a_x \theta^x = g(\theta) \ldots (*)$$

Differentiating $(*)$ with respect to θ, we get:

$$\sum_{i=0}^{n} a_x x \theta^{x-1} = g'(\theta)$$

This can be rewritten as (using some simple transpositions and manipulations) as:

$$\sum_i x\, \frac{a_x \theta^x}{g(\theta)} = \frac{\theta\, g'(\theta)}{g(\theta)} = \frac{\sum_i (x_i)}{n} \left(using\ (i)\right)$$

Thus, we can see that the maximum likelihood estimate and method of moments estimate are identical and are equal to the sample mean $\frac{\sum_i (x_i)}{n}$.

This special property of equality of maximum likelihood and method of moments estimates can be used to our advantage a great deal: especially for the sake of problem solving.

References:

(This problem was originally framed as an examination question set by Delhi University, B.Sc. Statistics honours, 1989)

- *Statistical Inference, by Casella and Berger*
- *Fundamentals of Mathematical Statistics, by Gupta and Kapoor*
- *An Outline of Statistical Theory (Vol. II), by Goon, Gupta and Dasgupta.*

Student ARTICLES

Exploring the performance of Sample Median as an Estimator of Location Parameter for Cauchy Distribution.

- Shrayan Roy (3rd Year), Adrija Saha (3rd Year)

Abstract:

Estimation of location parameter of Cauchy distribution is of immense importance in statistical inference. But the estimation is difficult. Because, Cauchy distribution has thick tails, i.e., extreme observations. For this reason, Sample mean becomes a terrible choice for estimating the location parameter. An intuitive choice may be Sample median. But is it really a good choice of estimator? This article, using theoretical derivations, simulations and graphical understanding, tries to explore different properties of Sample median, its asymptotic distribution and compares it with other well-known classes of estimators (like – Trimmed Mean, MLE). Also, a process to get an efficient estimator (One step estimators) is discussed.

Student ARTICLES

Introduction:

Statistics is the art and science of extracting stories from data. Any given data has certain properties. Statistics is a way to extract those properties. In the process of summarization, the first thing we look into the data, is its central tendency. Central tendency is basically the representation of the data via a single value, around which the datapoints are clustered.

The commonly used measures of central tendency are –

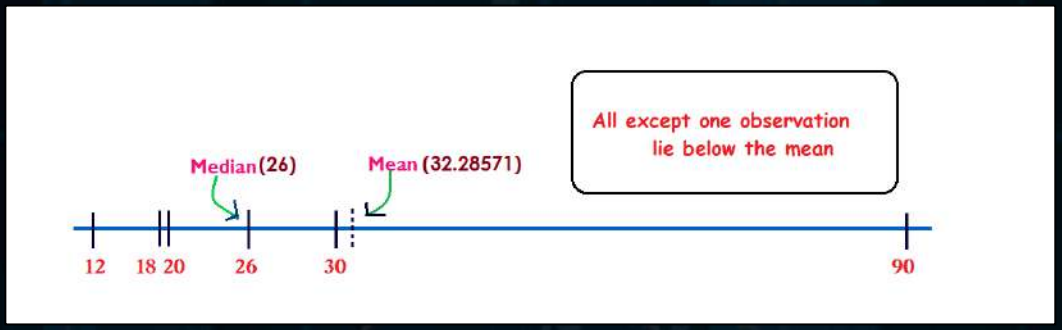
1. Mean
2. Median
3. Mode

Mean and Median are generally used for quantitative data, while Mode is generally used for qualitative data. Here, we will discuss about mean and median only.

Mean is the most commonly used measure of central tendency, but it is not a robust measure. It is readily affected by the presence of outliers. While median is a robust measure and is not affected by outliers. We illustrate this using an example –

Consider, the datapoints – 20,12,26,18,30,33,30. The mean of the data is 24.14286 and median is 26. Both are giving satisfactory result in this case.

But suppose we replace one observation (say, 33) by a larger value (say, 90). Then, the mean is 32.28571 and median remains 26. Clearly, mean is not a good measure of central tendency in this case. Since all except one observation lie below the mean, while median is still a good measure and is not affected by outlier.



So, when data contains extreme observations, it is better to use median.

Now we will explore these situations in case of Probability Distribution. Normal distribution is an ideal case. It possesses beautiful properties that Cauchy Distribution doesn't have.

Comparison between Normal Distribution and Cauchy Distribution:

A Random variable X is said to have a Normal Distribution with parameters μ and σ^2 if its pdf is given by -

$$(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \ -\infty < x < \infty$$

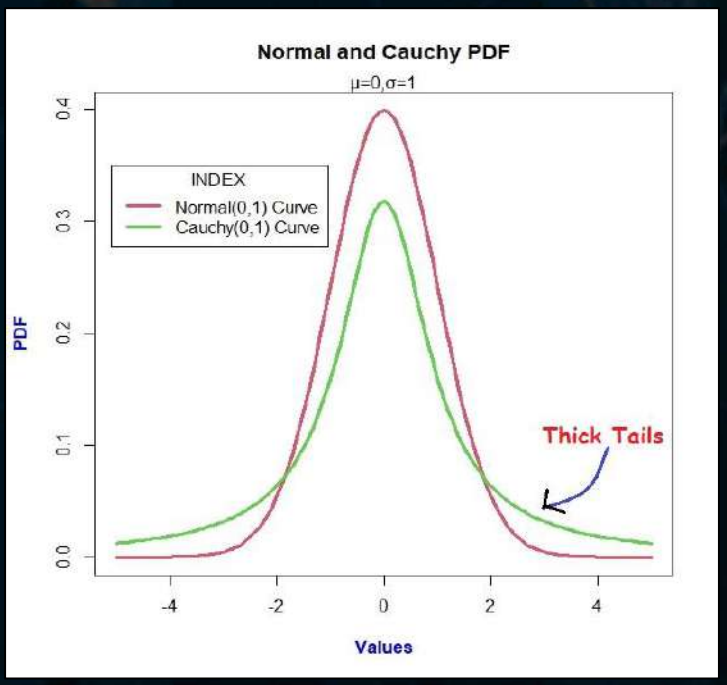
$\mu(\in \mathbb{R})$ is the mean and σ^2 $(\sigma > 0)$) is the variance of the distribution. Notationally, $X \sim Normal(\mu, \sigma^2)$

Again, a Random Variable Y is said to have a Cauchy Distribution with parameters μ and σ, if its pdf is given by –

$$f(y) = \frac{\sigma}{\pi} \frac{1}{\sigma^2 + (y-\mu)^2}; -\infty < y < \infty$$

$\mu(\in \mathbb{R})$ is the median and $\sigma(> 0)$ is the quartile deviation of the distribution. Notationally, Y ~ Cauchy(μ,σ) .

We Plot the pdf of Normal and Cauchy Distribution on the same graph.



The main difference between Normal distribution and Cauchy distribution is that Cauchy Distribution has extreme observations, which is reflected by the thick tail of Cauchy pdf. The presence of extreme observations leads to non-existence of moments of Cauchy distribution.

That is why quantiles are used for characterizing Cauchy distribution. The parameters of Cauchy distribution are based on quantiles, i.e., Median (μ) and Quartile deviation (σ).

Estimation of parameters is an important task in inference. So, here also we are interested in estimating the parameters of Normal and Cauchy distribution based on random sample from it. It is quite intuitive to use sample mean for estimating the location parameter.

Any good estimator has certain properties. Consistency is one of them.

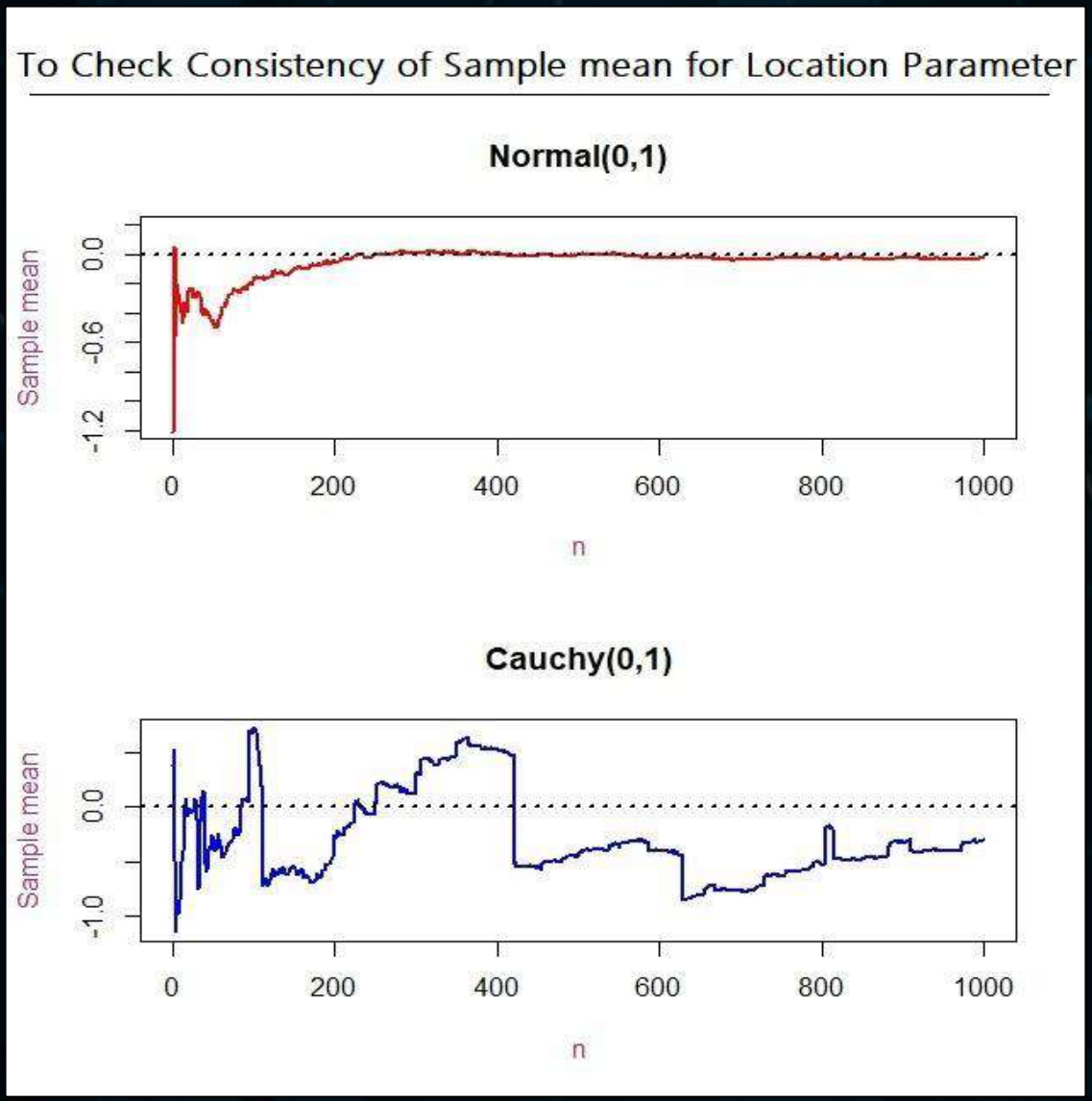
A sequence $\{T_n\}$ of estimators is said to be consistent for a parameter θ iff for every ε > 0,

$$P_r(|T_n - \theta| < \varepsilon) \to 1 \text{ as } n \to \infty.$$

Notationally, $T \xrightarrow{P} \theta$ .i.e., T converges in probability to θ. We will check through simulation whether sample mean is consistent for Location parameter of Normal Distribution and Cauchy Distribution.

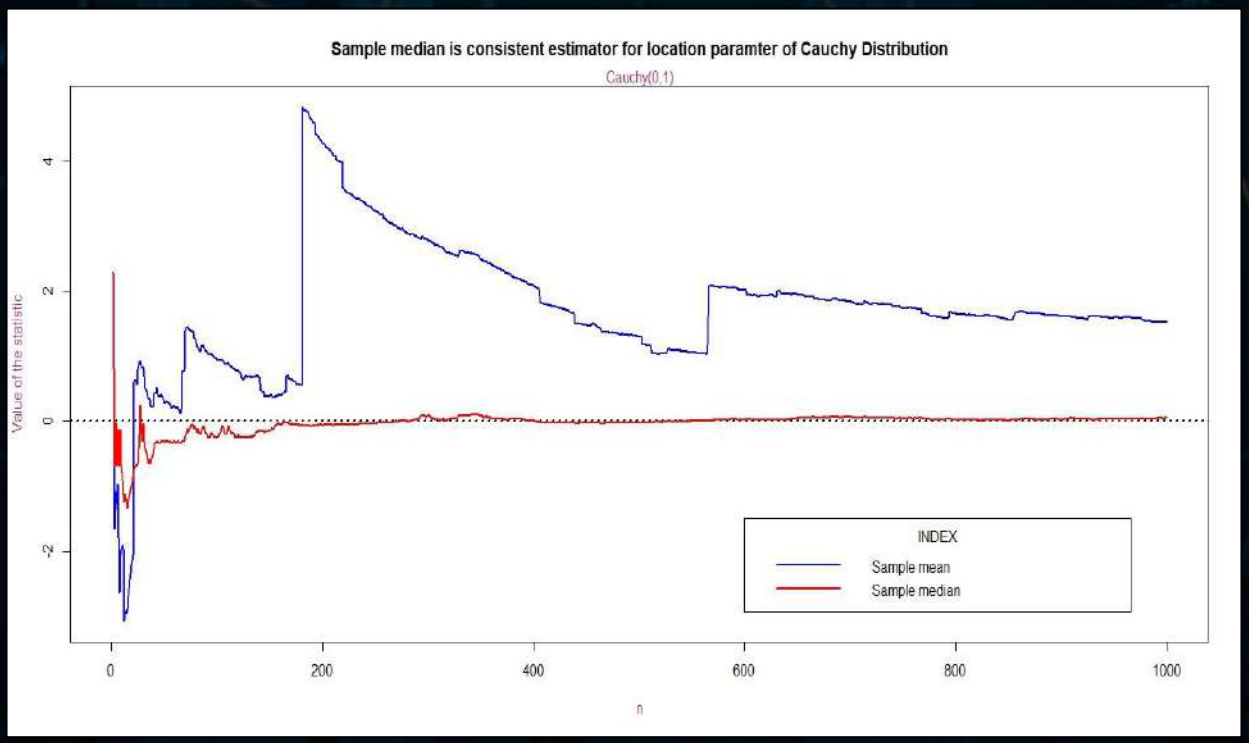
Student ARTICLES

First, we will draw a random sample $(x_1, x_2, \ldots, x_n)$ of size n (we take n = 1000) from Normal(0,1) distribution. Then, we calculate sequentially the following statistics $T_1 = x_1$, $T_2 = \frac{x_1+x_2}{2}$, $T_3 = \frac{x_1+x_2+x_3}{3}$, ..., $T_n = \frac{x_1+x_2+\ldots+x_n}{n}$. We then plot them against sample sizes i.e., 1, 2..., n. Similarly, we have done for Cauchy(0,1) also.



From the above graph, it is evident that Sample mean is no way consistent for location parameter of Cauchy Distribution, while it seems to be consistent for location parameter of Normal distribution. Let's try to find out the reason. A reason may be – Cauchy distribution has a wonderful property. If, X ~ Cauchy(μ,σ), then sample mean has same distribution Cauchy(μ,σ), whatever sample size may be. That is why the variability of the Cauchy sample mean is not reduced with increasing sample size.

Intuitively, we can think of sample median as a better choice for estimating the location parameter of Cauchy distribution. As, sample median may be a consistent estimator for the population median i.e., location parameter of Cauchy Distribution. Is it so? let's check through simulation.



Following the same steps as mentioned before we have sequentially calculated sample mean along with sample median and plot them. From the above graph it seems that Sample median is consistent for location parameter of Cauchy distribution. So, Sample median satisfies the criteria of consistency.

Now, we will find the probability distribution of Cauchy Sample median. It is not very difficult to find. But the expression of the pdf is a bit complicated.

Probability Distribution of Cauchy Sample median:

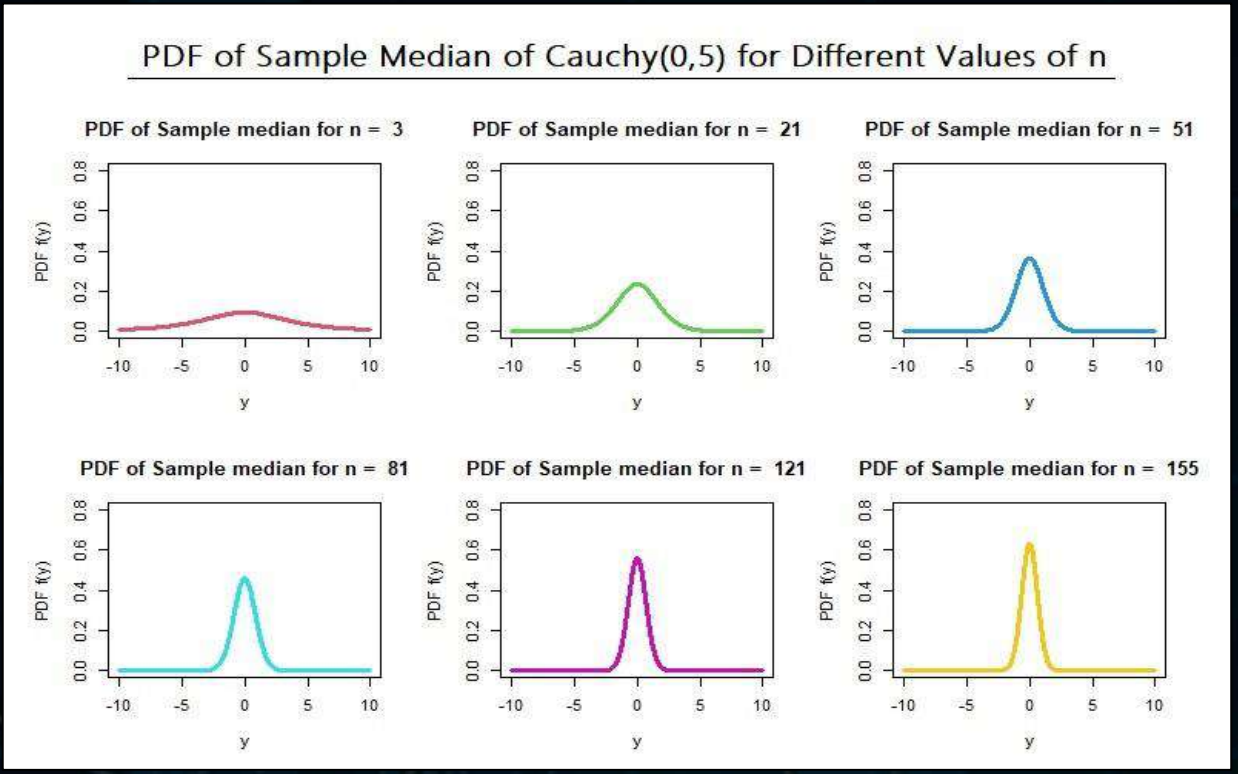
Now, we will derive the probability distribution of sample median of Cauchy distribution.

Let, $X \sim Cauchy(\mu, \sigma)$.

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from the distribution of X. Let n is odd (n = 2k+1,$k \in \mathbb{Z}^+$) as sample median is well defined for odd number of observations. Let,Y_n = Median($X_1, X_2, \ldots\ldots\ldots, X_n$). Then, the pdf of Y_n is given by –

$$f_{Y_n}(y) = \frac{(2k+1)!}{k!\ k!}[F_X(y)]^k f_X(y)[1 - F_X(y)]^k \ ; -\infty < y < \infty$$

Here, $f_X(.)$ & $F_X(.)$ are pdf and CDF of X respectively. We plot the pdf for six different values of n.



Note that the pdfs are symmetric about location parameter 0 and the pdfs are concentrating around 0, as n increases. This is because as n increases, the variability decreases. Now we will discuss different properties of distribution of sample median which are important in inference.

Properties of the Distribution:

1. Since Cauchy distribution is itself symmetric and we know the distribution of sample median (with sample size odd) is also symmetric about the point of symmetry of parent distribution. Hence, the distribution of sample median of Cauchy distribution is also symmetric about the population median (μ).

Here, $F_X(\mu + y) + F_X(\mu - y) = 1 \forall y \in \mathbb{R}$
So, $f_Y(\mu + y) = f_Y(\mu - y) \forall y \in \mathbb{R}$.

Hence, the distribution of Sample median of Cauchy is symmetric about μ.

2. Since, pdf of Cauchy sample median is symmetric about μ, so Expectation, if exists, must be equal to μ . It can be shown that Expectation exists and is equal to μ.

3. The expression for variance of Cauchy Sample median is complicated. However, for large sample the variance can be written as –

$$Var(Y_n) \approx \frac{1}{4n}(f_X(\mu))^{-2} = \frac{\pi^2\sigma^2}{4n}$$

Although we have obtained using numerical integration, the value of variance for different sample size using R software. We tabulated the values of variance for different sample sizes and σ.

Table 01: Numerically obtained Variance of Sample Median for different values of σ and n.

Sigma(σ) / Sample Size(n)	1	5	10	15	20	25	30
5	1.2213	30.53	122.1	274.78	488.5	763.3	1099
9	0.4087	10.22	40.87	91.948	163.5	255.4	367.8
13	0.2456	6.141	24.56	55.265	98.25	153.5	221.1
17	0.1756	4.39	17.56	39.514	70.25	109.8	158.1
21	0.1367	3.417	13.67	30.753	54.67	85.42	123
25	0.1119	2.797	11.19	25.172	44.75	69.92	100.7
29	0.0947	2.367	9.47	21.307	37.88	59.19	85.23
33	0.0821	2.052	8.209	18.47	32.84	51.31	73.88
37	0.0725	1.811	7.245	16.301	28.98	45.28	65.2
41	0.0648	1.621	6.483	14.587	25.93	40.52	58.35
45	0.0587	1.467	5.866	13.2	23.47	36.67	52.8
49	0.0536	1.339	5.357	12.053	21.43	33.48	48.21
53	0.0493	1.232	4.929	11.09	19.72	30.81	44.36
57	0.0456	1.141	4.564	10.269	18.26	28.53	41.08
61	0.0425	1.062	4.25	9.562	17	26.56	38.25
65	0.0398	0.994	3.976	8.945	15.9	24.85	35.78

4. For Large n, $\frac{Y_n - \mu}{\sqrt{\frac{\pi^2\sigma^2}{4n}}} \xrightarrow{L} Z \sim N(0,1)$

   Hence, In Large sample (for known σ) we can perform certain tests of location parameters of Cauchy Distribution and equality of Location parameters of two independent Cauchy distribution using sample median. This normal approximation is valid as well in small sample sizes also. One can perform simulation study on this.

5. Under certain regularity conditions, the sample quantiles are consistent for population quantiles. Hence, sample median of Cauchy Distribution is consistent for Population median μ.

$$Y_n \xrightarrow{P} \mu$$

6. We can form *confidence interval* of Population Location parameter using Sample Median.

Asymptotic Efficiency of Sample Median:

So, we have found that, though sample mean is a terrible estimator for location parameter of Cauchy distribution, sample median is reasonably a better one. Now, we will look into another property of estimation, which is efficiency of estimator. But what is Efficiency?

Suppose the regularity conditions of Cramer – Rao Inequality hold. Then according to Cramer– Efficiency of an Unbiased Estimator T of g(θ) is defined as the ratio of the Cramer-Rao lower bound (CRLB) to the variance of T. It is denoted by –

$$eff(T) = \frac{CRLB}{V_\theta(T)}$$

If eff(T) = 1, then the estimator T is said to be efficient estimator of g(θ) and if for large n eff(T) → 1. Then, it is called asymptotically efficient.

Here, $\theta = \mu$ and $g(\theta) = g(\mu) = \mu$. Hence, $g'(\mu) = 1$.

Now, the *fisher Information* of μ in $(X_1, X_2, \ldots\ldots, X_n)$ is given by –

$$I(\mu) = \frac{n}{2\sigma^2}$$

Hence, CRLB is given by, $\frac{\{g'(\mu)\}^2}{I(\mu)} = \frac{2\sigma^2}{n}$. Hence, for large n,

$$eff(Y_n) \approx \frac{\frac{2\sigma^2}{n}}{\frac{\pi^2\sigma^2}{4n}} = \frac{8}{\pi^2} < 1$$

Hence, Sample median is *asymptotically inefficient*. So, we may search for some efficient estimator for estimating location parameter. *One Step estimation* may be helpful in this case.

Student ARTICLES

One Step Estimation: Improving Sample median as Estimator-

One Step Estimation is a process of producing an asymptotically efficient estimator for some parametric function $g(\theta)$. In this process, we start with a consistent, but inefficient estimator. Then, we produce an Asymptotically Efficient estimator using this inefficient estimator. The asymptotic variance of the estimator will be equal to inverse of Fisher Information $i.e., \{I(\theta)\}^{-1}$.

The one step efficient estimators based on a consistent estimator $\hat{\theta}_n$ are as follows –

$$\hat{\theta}^{(1)} = \hat{\theta}_n - (l_n\ddot{(}\hat{\theta}_n))^{-1} l_n(\dot{\hat{\theta}}_n) \quad \text{..... (1)}$$

$$\hat{\theta}^{*} = \hat{\theta}_n + (I(\hat{\theta}_n))^{-1} \frac{1}{n} l_n\dot{(}\hat{\theta}_n) \quad \text{..... (2)}$$

Where, $l_n(\dot{\hat{\theta}}_n)$ is value of the single derivative with respect to θ of log likelihood at $\theta = \hat{\theta}_n$ and $l_n(\ddot{\hat{\theta}}_n)$ is the double derivative of log likelihood at $\theta = \hat{\theta}_n$. Also, $I(\hat{\theta}_n)$ is the value of Fisher Information at $\theta = \hat{\theta}_n$.

The first estimator is obtained by using *Newton Raphson Method* and second estimator is obtained using *Method of Scoring*. The one-step estimators are asymptotically equivalent to the efficient estimator.

Here, our consistent estimator is Sample median Y_n. Hence, for a random sample $(X_1, X_2, \ldots\ldots, X_n)$ of size n from *Cauchy*(μ, σ),σ is known. The log likelihood is given by –

$$l_n(\mu) = nlog(\frac{\sigma}{\pi}) - \sum_{i=1}^{n} log\,(\sigma^2 + (x_i - \mu)^2)$$

Then, $l_n\dot{(\mu)} = \sum_{i=1}^{n} \frac{2(x_i - \mu)}{\sigma^2 + (x_i - \mu)^2}$, $l_n\ddot{(\mu)} = -2\sum_{i=1}^{n} \frac{\{\sigma^2 - (x_i - \mu)^2\}}{\{\sigma^2 + (x_i - \mu)^2\}^2}$ and $I(\mu) = \frac{n}{2\sigma^2}$

So, therefore the one step estimators are given by –

$$\hat{\mu}^{(1)} = Y_n - (l_n\ddot{(Y_n)})^{-1} l_n(\dot{Y_n})$$

$$\hat{\mu}^{*} = Y_n + (I(Y_n))^{-1} \frac{1}{n} l_n(\dot{Y_n})$$

Hence, we get one step estimators which are efficient. The asymptotic variance will be equal to $\frac{2\sigma^2}{n}$.

Another efficient estimator is *Trimmed mean*. We will compare efficiency of Sample Median and Trimmed Mean.

Student ARTICLES

Trimmed Mean as a more efficient estimator than Sample Median:

Trimmed Mean is basically the average of sample order statistics after discarding some extreme observations. Clearly, Sample mean and Sample median are both extreme cases of Trimmed mean. Suppose, we have drawn a random sample $(X_1, X_2, \ldots\ldots\ldots, X_n)$ of size n. For our case let n is odd $(2k + 1, k \in \mathbb{Z}^+)$. Consider the order Statistics $X_{(-k)} \leq X_{(-\overline{k-1})} \leq \ldots\ldots\ldots \leq X_{(k)}$. Then, trimmed mean is defined as,

$$H(a) = \frac{1}{2a+1} \sum_{i=-a}^{a} X_{(i)}$$

clearly, $0 \leq a \leq k$. H(a) is mean of 2a+1 central sample values. This is a *class of estimators*. If a = 0, then H(0) is Sample Median and when a = k, then H(k) is sample mean. We can show that H(a) is unbiased for a < k. We can find the asymptotic variance of H(a), the expression is a bit complicated. Defining m = a/k and for σ = 1, we have –

$$V(m) = \frac{1-m}{m^2} tan(\frac{\pi m}{2})^2 + \frac{2}{\pi m^2} tan(\frac{\pi m}{2}) - \frac{1}{m}$$

We can find this expression for any σ. The tabulated values of V(m) are given by –

Table 02:*Values of V(m) for some specific values of m*

m	V(m)
0	2.467
0.1	2.34
0.2	2.283
0.24	2.278
0.3	2.29
0.4	2.37
0.5	2.546
0.6	2.872
0.7	3.48
0.8	4.772
0.9	8.773
1	∞

From the above table it is clear that for m = 0.1,0.2,0.24,0.3,0.4; the estimator H(a) is more efficient as compared to sample median(H (0)). Even, among those estimators m = 0.24 is the most efficient.

Comparison Between Trimmed Mean (m = 0.24) and Sample Median In terms of MSE:

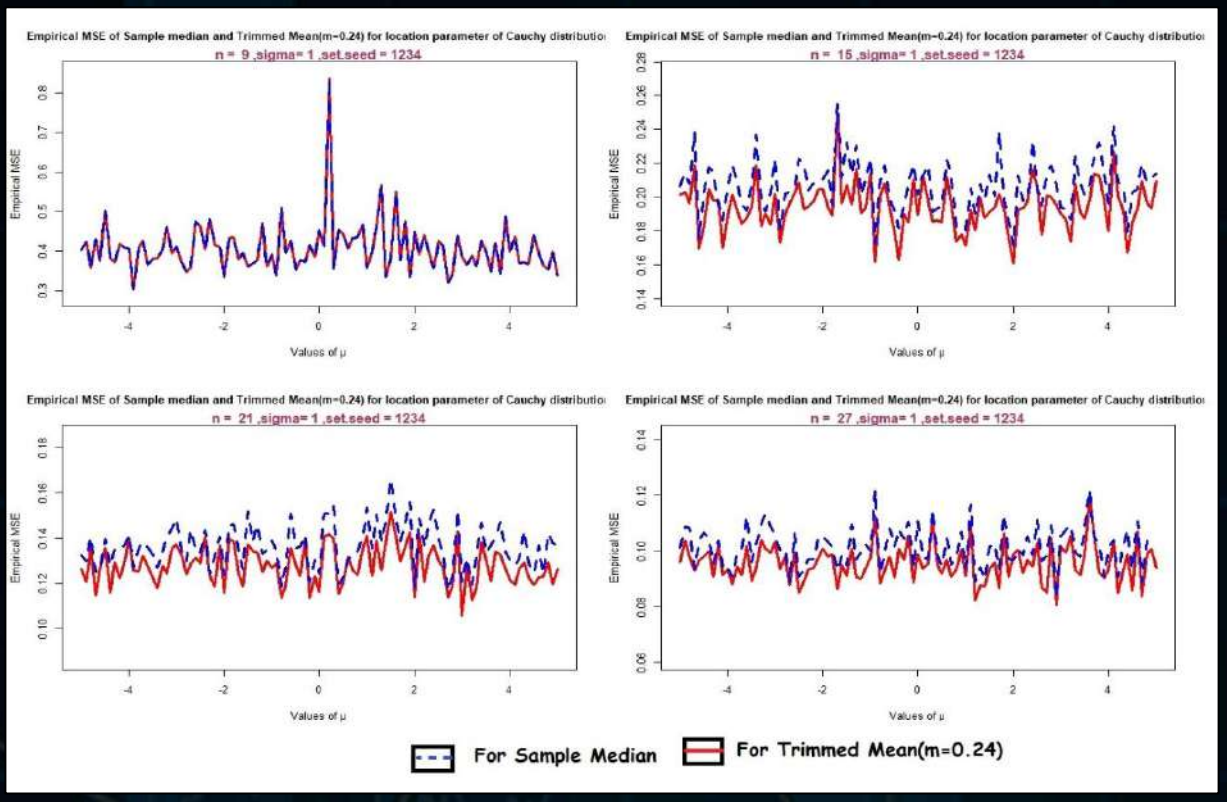
We can compare the Goodness of two estimators in terms of having less *mean-squared-error*(MSE) for estimating the parameter. Since, trimmed mean with m = 0.24 is the most efficient estimator in the class. We compare the suggested estimator trimmed mean (m = 0.24) with sample median in terms of MSE for estimating μ.

The MSE of an Estimator T in estimating $g(\theta)$ is given by –

$$MSE_\theta(T) = {E_\theta}^2(T - g(\theta)) \; ; \; \theta \in \Omega \; (Parameter \; Space)$$

So, we may perform a simulation study to compare the Empirical MSEs. Simulation is a good idea to compare the MSEs (Why?).

For different sample sizes (odd) and for different values of μ, we run the simulation R = 500 times and store the value of sample median and trimmed mean (m = 0.24) in each case. Then we note the difference and squared difference between the true parameter value and estimated value. Ultimately, we compute the average of these quantities to find empirical bias and empirical MSE's. We represent the simulation results of empirical MSEs using graph.



From the above diagram, (*Notice the y axis scales are different*) we can clearly see for every chosen value of μ and sample size, trimmed mean (with m = 0.24) has uniformly less *Empirical MSE* than that of sample median, except the case n = 9 where both estimators become identical (Why?). Though the difference is not very large. Also, as sample size increases MSE decreases.

So, in terms of MSE also trimmed mean (m = 0.24) is superior to sample median.

Student ARTICLES

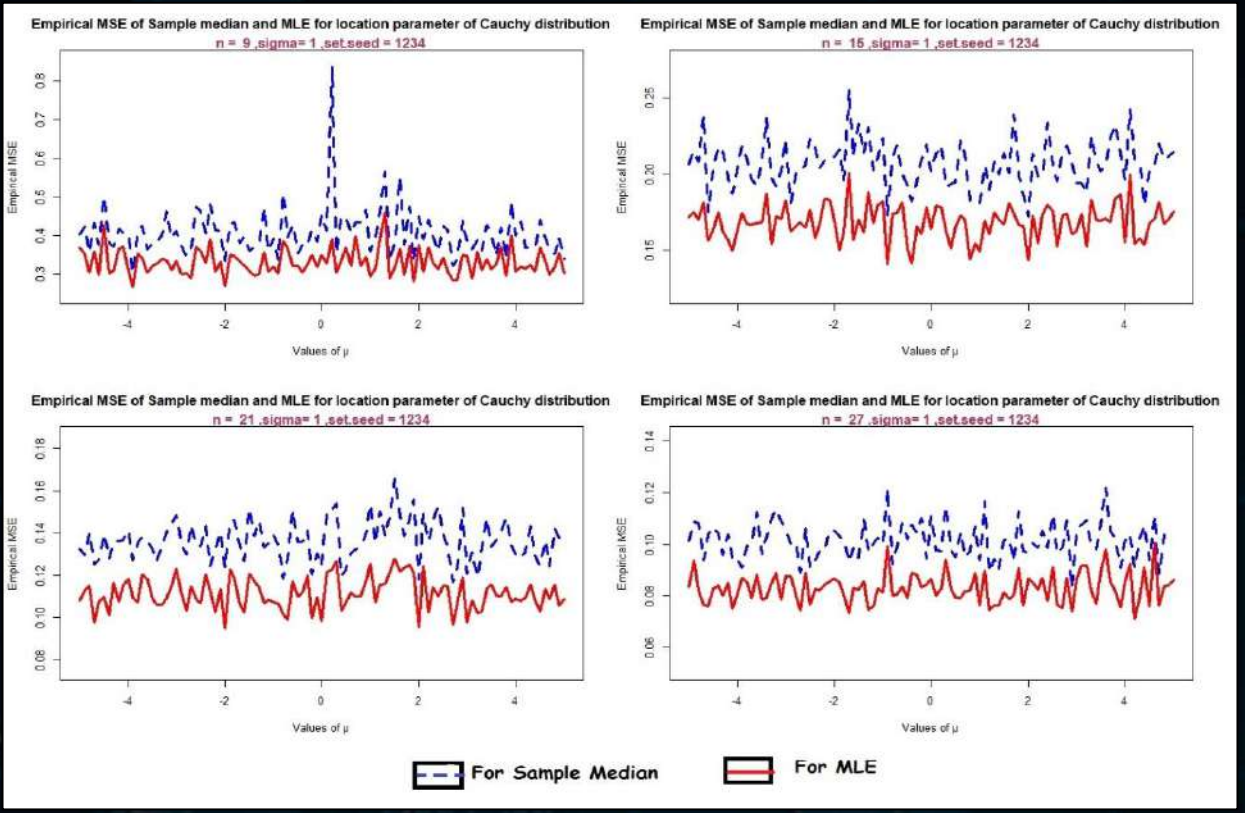
Comparison Between MLE of Location parameter and Sample Median:

We know an elegant method of estimation is *Maximum Likelihood Estimation*. But, unfortunately in case of Cauchy distribution MLE of its location parameter has no closed form. i.e., we can find it only by numerical methods. So, we cannot have the idea of its distribution or other properties. But we know MLE has some nice large sample properties. In fact, MLE is the most asymptotically efficient estimator. i.e., its asymptotic variance is equal to inverse of fisher information i.e., $\{I(\mu)\}^{-1}$.

So, we may wish to compare sample median and MLE of μ in terms of having less Mean Squared Error for estimating μ.

At this juncture we face the same problem for not having any closed form of MLE. So, we may perform a simulation study to compare the Empirical MSEs in a similar way as we have done in case of comparing trimmed mean (m = 0.24) with sample median.

We represent the simulation results using a graph.



It is very clear from the above diagram (*Notice the y axis scales are different*)that for every chosen value of μ and sample size, the MLE of μ has uniformly less Empirical MSE than that of sample median as it was in the case of trimmed mean (m = 0.24). Though the difference is not very large. Also, as sample size increases MSE decreases.

So, in terms of MSE also MLE of μ is superior to sample median.

Conclusion:

From the above discussion, we have seen though some estimators are there (Trimmed Mean, MLE) which are somewhat better than Sample Median in terms of *Efficiency* or *MSE*, they cannot be always readily tackled. In one case no closed expression is available (MLE) and in other case computations are tedious (Trimmed Mean). But, Sample median (when sample size is odd) is easy to calculate, unbiased and consistent with not so large MSE. Hence, sample median is found to be a reasonably good estimator for estimating the location parameter of Cauchy distribution.

Simulation Code:

You can find our simulation study code here-

https://codeboard.io/projects/296580

References:

1. *'A NOTE ON ESTIMATION FROM A CAUCHY SAMPLE' - THOMAS J. ROTHENBERG and FRANKLIN M. FISHER and C. B. TILANUS*
2. *Rand R. Wilcox, 'Applying Contemporary Statistical Techniques'*
3. *'An Introduction to Probability and Statistics' by A. K. Md. Ehsanes Saleh and V. K. Rohatgi*

Student ARTICLES

Proving the Famous Wallis Product Formula by the Concept of 'Probability Distribution'

- Rishiraj Sutar (3rd Year)

There are many beautiful formulae for π, like the famous Basel's formula proposed by Euler, the Wallis formula, the Gregory Leibnitz formula, and many more.

The Wallis product formula for π is given by

$$\prod_{k=1}^{\infty} \frac{2k.2k}{(2k-1)(2k+1)} = \frac{\pi}{2}$$

and it was discovered by the English mathematician, John Wallis in the year 1656.

There are several methods to prove the Wallis formula for π, and one such method is by using the concept of *Probability Distribution*.

Steven Joel Miller, a mathematician who is currently a professor at Williams College, introduced an interesting idea of using the Student's t-Distribution for deriving the Wallis Product formula for π.

Student ARTICLES

The Student's t-Distribution

The ρ.d.f of Student's t-distribution (with 'n' degrees of freedom) is given by:

$f(t) = \left(\frac{\Gamma(\frac{n+1}{2})}{(\sqrt{\pi n})\Gamma(\frac{n}{2})}\right).\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$,$-\infty < t < \infty$, $n > 0$

$= D_n\left(1 + \frac{t^2}{n}\right)^{\frac{-(n+1)}{2}}$ where, $D_n = \frac{\Gamma(\frac{n+1}{2})}{(\sqrt{\pi n})\Gamma(\frac{n}{2})}$

The Proof

We know, $\int_{-\infty}^{\infty} f(t)\, dt = 1$

Or, $\int_{-\infty}^{\infty} D_n\left(1 + \frac{t^2}{n}\right)^{\frac{-(n+1)}{2}} dt = 1$

Or, $\int_{-\infty}^{\infty}\left(1 + \frac{t^2}{n}\right)^{\frac{-(n+1)}{2}} dt = \frac{1}{D_n}$

Taking the limit $n \to \infty$, we get,

$Lim_{n\to\infty} \int_{-\infty}^{\infty}\left(1 + \frac{t^2}{n}\right)^{\frac{-(n+1)}{2}} dt = Lim_{n\to\infty}\frac{1}{D_n}$

Interchanging the integral and limit, we get,

$Lim_{n\to\infty}\frac{1}{D_n} = \int_{-\infty}^{\infty} Lim_{n\to\infty}\left(1 + \frac{t^2}{n}\right)^{\frac{-(n+1)}{2}} dt$

$= \int_{-\infty}^{\infty} e^{-Lim_{n\to\infty}(t^2/n).(\frac{n+1}{2})}\, dt$

$= \int_{-\infty}^{\infty} e^{-t^2/2}\, dt$

$= \sqrt{2\pi}$

Therefore,

$Lim_{n\to\infty} D_n = \frac{1}{\sqrt{2\pi}}$

We defined earlier that, $D_n = \frac{\Gamma(\frac{n+1}{2})}{(\sqrt{\pi n})\Gamma(\frac{n}{2})}$

Therefore, $D_{2n} = \frac{\Gamma(\frac{2n+1}{2})}{(\sqrt{2\pi n})\Gamma(n)}$

We will use the following results

- $\Gamma(\frac{2n+1}{2}) = \frac{(2n-1)\,(2n-3)\,\ldots\ldots\ldots 5.3.1.\sqrt{\pi}}{2^n}$
- $\Gamma(n)$ = (n-1)!
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Now, D_{2n} can be expanded as:

$D_{2n} = \frac{1.3.5\ldots\ldots(2n-3).(2n-1)}{1.2.3\ldots\ldots(n-1).(\sqrt{2\pi n})}\frac{\sqrt{\pi}}{2^n}$

$= \frac{1.3.5\ldots\ldots(2n-3).(2n-1)}{2.4.6\ldots\ldots(2n-2).(2n)}\frac{\sqrt{n}}{\sqrt{2}}$ ---------- (i)

Squaring both sides of (i) and multiplying the RHS by (2n+1)/(2n+1) we get,

$D^2_{2n} = \{(\frac{1\,.\,3}{2\,.\,2})\,(\frac{3\,.\,5}{4\,.\,4})\,\ldots\ldots\ldots(\frac{(2n-1).(2n+1)}{2n\,.\,2n})\,\frac{1}{(2n+1)}\}.\frac{n}{2}$ ----------(ii)

Now,

$$\prod_{k=1}^{n} \frac{2k.2k}{(2k-1)(2k+1)} = \frac{2\,.\,2}{1\,.\,3} \cdot \frac{4\,.\,4}{3\,.\,5} \cdots\cdots\cdots \frac{2n\,.\,2n}{(2n-1).(2n+1)}$$

$$= \frac{n}{(4n+2)} \frac{1}{D_{2n}^2}$$

As, $Lim_{n\to\infty} D_{2n}^2 = \frac{1}{2\pi}$, implies that,

$$Lim_{n\to\infty} \frac{n}{(4n+2)} \cdot \frac{1}{D_{2n}^2} = \frac{\pi}{2}$$

Thus,

$$\prod_{k=1}^{\infty} \frac{2k.2k}{(2k-1)(2k+1)} = \frac{\pi}{2}$$

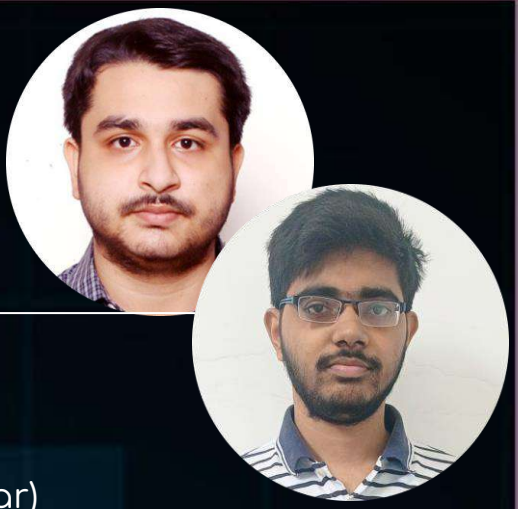
The Conclusion

Thus, we have obtained the Wallis product formula by using the concept of *probability distribution*. The proof above shows the beauty of the 'probability distribution' approach and how it can be used to provide solution for such famous formulae.

References:

1. *https://www.researchgate.net/publication/1764934_A_Probabilistic_Proof_of_Wallis's_Formula_for_p*
2. *J. F. Scott, Mathematical Works of John Wallis.*
3. *An Introduction to Probability Theory and Its Applications by W. Feller.*

Student ARTICLES



SIMPSON'S PARADOX

- Debarghya Baul (3rd Year), Arnab Roy (3rd Year)

Simpson's Paradox is a statistical phenomenon which occurs when sets of data indicate a specific trend but after combining the data, the trend is reversed. It is very important to understand and identify the paradox to interpret data correctly.

History:

Although the paradox bears the name of British Statistician Edward Simpson, he did not discover this phenomenon. It was British Statistician George Udny Yule who first reported the existence of this paradox between categorical variables in 1903.

Example:

Let us consider an example to explain this. Suppose, a soft drink company undertakes a survey to find which flavour of the strawberry and peach soft drinks tastes better to the customers. They set up one sampling stall for each flavour and ask 1000 people at each stall to taste the respective flavours. Information on the gender of the people is also collected.

Student ARTICLES

Let X be an attribute denoting the flavour, namely, strawberry and peach; Y be an attribute denoting the preference towards the flavours, namely, yes or no; Z be an attribute denoting the gender of respondent, namely, male and female.

The data collected from this survey is summarized in the following table:

GENDER (Z)	LIKED FLAVOUR(Y)		
	FLAVOUR NAME (X)	YES	NO
MALE	STRAWBERRY	750	150
	PEACH	590	110
FEMALE	STRAWBERRY	50	50
	PEACH	160	140
TOTAL	STRAWBERRY	800	200
	PEACH	750	250

For the purpose of analysis, the above table is divided into three sub tables:

a) Partial Table For Males:

FLAVOUR NAME	LIKED FLAVOUR(Y)		
(X)	YES	NO	TOTAL
STRAWBERRY	750	150	900
PEACH	590	110	700

From this table, we observe that,

I. Proportion of male who liked strawberry flavour = $\frac{750}{900}$ = 0.833

II. Proportion of male who liked peach flavour = $\frac{590}{700}$ = 0.843

b) Partial Table For Females:

FLAVOUR NAME	LIKED FLAVOUR(Y)		
(X)	YES	NO	TOTAL
STRAWBERRY	50	50	100
PEACH	160	140	300

From this table, we observe that,

I. Proportion of female who liked strawberry flavour = $\frac{50}{100}$ = 0.5

II. Proportion of female who liked peach flavour = $\frac{160}{300}$ = 0.533

(c) Marginal Table combining the two genders:

FLAVOUR NAME	LIKED FLAVOUR(Y)		
(X)	YES	NO	TOTAL
STRAWBERRY	800	200	1000
PEACH	750	250	1000

From this table, we observe that,

I. Proportion of people who liked strawberry flavour = $\frac{800}{1000}$ = 0.80

II. Proportion of people who liked peach flavour = $\frac{750}{1000}$ = 0.75

Observing the results from the partial tables, it can be said that strawberry flavour is preferred by 83.3% men and 50% women while peach flavour is preferred by 84.3% men and 53.3% women. On the other hand, from the marginal table we observe that 80% people prefer strawberry flavour and 75% people prefer peach flavour. Thus, it is seen, by ignoring the gender, the preference towards strawberry flavour is higher than towards peach flavour. However, when the two genders are considered separately, there seems almost similar preferences towards the two flavours. This is what we call Simpson's Paradox.

The main reason of this paradox is explained below:

Lurking variables

Often, we cannot detect the effect of a lurking variable on the response though it may affect the interpretation of relationships between variables.

In the presence of lurking variables, the data breaks into multiple distinct distributions. That is why Simpson's Paradox appears. In most of the cases, it is difficult to detect these lurking variables.

In the example cited above, we observe that the gender of the respondent (Z) has an influence on their opinion.

We can explain this paradox with the help of probability theory by taking the lurking variable (i.e. gender) into account.

P(A person liked strawberry) = P(The person liked strawberry |The person is a man) ×P(The person is a man) + P(The person liked strawberry | The person is a woman) ×P(The person is a woman)

That is, $\frac{800}{1000} = \frac{750}{900} \times \frac{900}{1000} + \frac{50}{100} \times \frac{100}{1000}$

P(A person liked peach) = P(The person liked peach | The person is a man) ×P(The person is a man) + P(The person liked peach | The person is a woman) ×P(The person is a woman)

That is, $\frac{750}{1000} = \frac{590}{700} \times \frac{700}{1000} + \frac{160}{300} \times \frac{300}{1000}$

We now consider the marginal probabilities of gender (i.e. P(The person is a male) and P(The Person is a female)) as weights. Observe that in the case of strawberry flavour, the P(The person is a male) is 0.9 which is very high. For this reason, the total probability that a person liked strawberry is highly affected by the opinions of male. In the case of peach flavour, the P(The person is a male) is 0.7 which is not as high as strawberry flavour. So, here, the opinion of females highly affects the preference towards the peach flavour. For this reason, we observe a lower marginal probability for the whole population to prefer peach flavour though both men and women separately prefer peach.

Detecting Simpson's Paradox:

It is very difficult to detect the occurrence of Simpson's Paradox. To detect this, there must be effective randomisation in the allocation of treatment to the experimental unit.

Conclusion:

It is obvious that if we want to detect the effect of lurking variables through the analysis of our data, we have to break the data in hand into several subsets having different distributions with respect to the lurking variables. However, it is more important to first analyse the data thoroughly and decide if it is reasonable to break the data into subsets. Otherwise, we will keep the data in its usual form. Based on the circumstances, we will take actions appropriately.

Student ARTICLES

The Existence of Mathematics Its Various Paradoxes

- Abhay Ashok Kansal (2nd Year), Arushi Bajoria (2nd Year)

The existence of Mathematics has always been the topic of many debates and controversies. Nobody can deny the importance of Mathematics, not only in electronic and scientific theory, but also in the natural world. Mathematics exists when you are counting the trees of a jungle or when you are trying to see the growth of a rabbit population (this can be interpreted using the Fibonacci sequence). We all know one plus one is two, but how was that worked out? It is complicated. Whether Math is a discovery or an invention hasn't received a definite answer.

Let us dive into the basic difference between discovery and invention and how they relate to the existence of Math.

A discovery is recognizing and concluding the presence of something that has existed way before people and its existence was previously unknown to people. It is the recognition and description of natural law. For example, the discovery of America by Christopher Columbus or the discovery that planets rotate around the sun.

Student ARTICLES

An invention is the creation of something that didn't exist before, with one's original ideas and creativity. It is the making of something new and different. For example, the invention of the telescope by Galileo or the invention of cars. Invention seems to be highly correlated to discovery, given the fact that no invention is possible without using existing materials, items, and theories. It's the use of creativity and ideas that make inventions unique. The combinations of available materials and theories allow for unlimited chances to make unique and new products.

"How is it possible that Mathematics, a product of human thought that is independent of experience, fits so excellently the objects of reality?"

~Albert Einstein

Now we shall explore Mathematics by stating various reasons why different scientists believe what they believe.

Mathematicians believe that arithmetic and geometry form the basis of every mathematical work that is presented. Most of the things that surround us can be decoded mathematically. Every natural phenomenon can draw a relation with Mathematics. The accuracy of gravity, moment of stars and Earth, can be mathematically interpreted. This accuracy led people to ask,
"Does God exist, and if he does, is he a Mathematician?"

Student ARTICLES

Let's imagine you throw a stick, and your dog runs to catch it. That dog, with no knowledge of parabola and projectile motion, catches the stick. Does that mean Math has a presence in the natural world and humans have been discovering mathematical concepts? Around 1000BC, when most of mathematical discoveries weren't made, Egyptians and Romans had made many discoveries which requires mathematical theorems, like celestial bodies and their movements, the use of Pythagoras' theorem to make pyramids and many more.

Mathematics has a presence in the physical world and the facts and concepts we use, have a rooted connection with nature. mathematical functions like addition, subtraction, multiplication, and divisions can be used and proved using natural objects. Using 10 trees, you can prove that 7+3=3+7 or one can use a pineapple to study the Fibonacci series. Thus, a connection between physical and scientific world can be established easily. But with so much of work done in this field and so many new and old theorems that exist between us, there are scientists who even believe that there are mathematical calculations and theorems that have no presence in the natural world. These calculations may be mathematically true, but they cannot be connected with the real world.

Student ARTICLES

"There are tons and tons of mathematical structures that are of no use at all in studying the physical world."

~Philosopher Mark Balaguer

The argument by Dr. Mark may seem actual, but there have been instances where mathematical structures, which were at the beginning considered to be just useless, have been of considerable importance to future discoveries and works. Mathematician Bernhard Riemann discussed new types of geometrics, which were very useful when Einstein formulated his theory of General Relativity in 1915.

Nobel Laureate Roger Penrose, who shares the discovery of black hole formation with Stephen Hawking, said,

"Mathematics has an independent existence, a 'Platonic existence'... People find it confusing how Mathematics beautifully defines reality."

There have been popular arguments between Mathematicians and Philosophers about the existence of Math. The popular question remains unanswered:

Is Math an invention of the basic human mind? Or does Math exist in the abstract world with humans merely discovering its truths?

Student ARTICLES

Mathematics is indisputable in terms of its effectiveness. Math is the core of many scientific works. But the essence of Math, the physical or mental existence of Mathematics is still a foggy topic. Mathematics describes the physical world and its happenings with remarkable precision. Math is a tool that helps us define the movements of planets and stars, the growth and decay of things. But is the existence of Mathematics a bluff? Are humans trying to define activities surrounding them by forcefully relating Mathematics with it? Would Mathematics exist if humans didn't? It is difficult to conclude the essence of Math. While one can relate so much of Math with nature, one cannot ignore the calculations that seem to have no existence in the real world. Whether Math was a discovery or an invention may be an unending debate, but the existence of paradoxes in Math is undeniable.

A mathematical paradox is a statement that seems to contradict itself while simultaneously seeming completely logical. Paradoxes occur when we deduce from a known premise but end up deriving a conclusion that seems logically unreasonable or contrary to one's expectation.

To better understand paradoxes, let's examine the statement, "This statement is false". If the statement is indeed false, and it proclaims that it is false, that means the statement is stating the truth. The statement is being truthful about being false, and hence it is impossible to justify whether the statement is true or false. This results in a self-contradictory statement or the Liar's Paradox.

Student ARTICLES

Paradoxes exist in all fields of Mathematics and logic, from set theory to geometry. Data Science and Statistics is no exception to this rule. Some of the statistical results, which go the most against our intuition, are –

Simpson's Paradox

Simpson's paradox, or the Yule-Simpson effect, states that the trends or results shown by a population may reverse, disappear or completely new trends may emerge when the population is broken down into subpopulations.

In 1973, the University of California, Berkeley was scared of facing lawsuits for being gender-biased against women while admitting graduate applicants. The admission figures showed that men were a lot more likely to get admitted to UC Berkeley than women were.

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	**44%**	4,321	35%

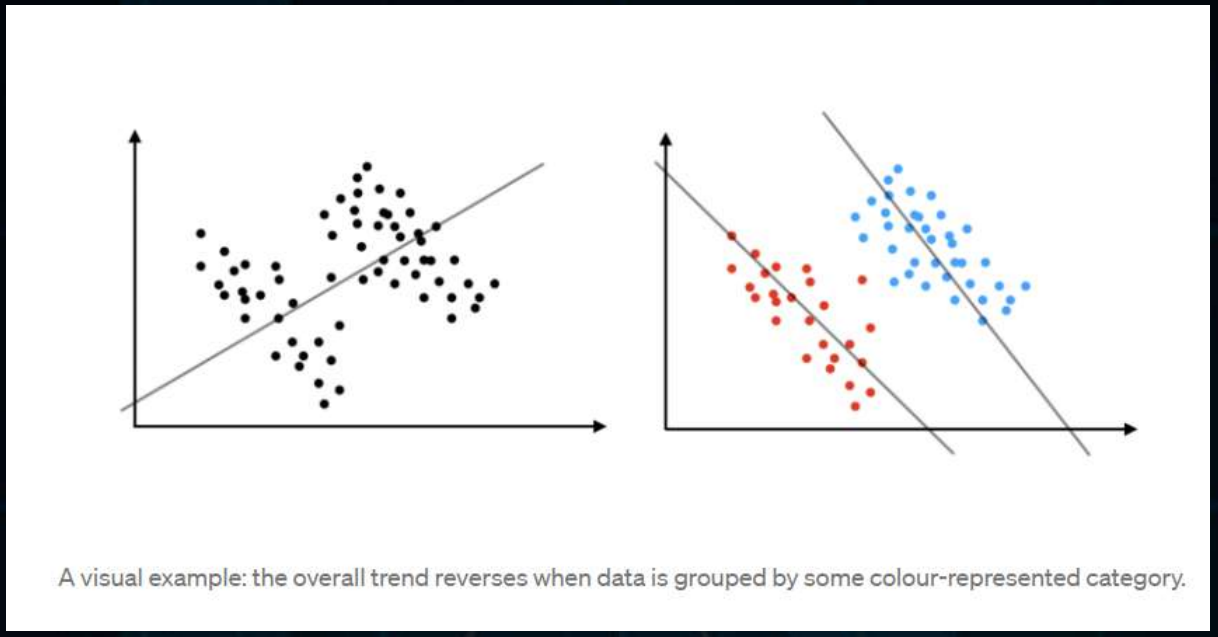
When statistician Peter Bickel looked more closely to the admission data of each of the 85 departments, he found that a few departments were biased in favour of women while admitting applicants, while other departments showed no significant gender bias. The data for 6 of the largest departments are given below.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	***825***	62%	108	**82%**
B	585	63%	***560***	63%	25	**68%**
C	918	35%	325	**37%**	***593***	34%
D	792	34%	417	33%	375	**35%**
E	584	25%	191	**28%**	***393***	24%
F	714	6%	373	6%	341	**7%**
Total	4526	39%	2691	45%	1835	30%

Peter Bickel concluded that the lower admission rates for women were because most women applied for more competitive departments (such as English), whereas men applied for departments with greater acceptance rates (such as Engineering).

The reasons for observing Simpson's paradox can be examined by looking beyond the data already provided. It often results because of the presence of a missing or lurking variable that hides the full story. In the UC Berkeley case, the confounding effect of having a strong association between departments and admissions (some departments admitted a much smaller percentage of applicants than others) and the strong association between departments and gender (females tended to apply to more selective departments) caused us to see a flip in gender bias when we studied the trends of each department in isolation.

Student ARTICLES



A visual example: the overall trend reverses when data is grouped by some colour-represented category.

Accuracy Paradox

Accuracy Paradox states that over-fitting a model to the data may cause an eventual decrease in accuracy. It explains that overspecialization in a particular feature can cause one to forget the big picture and produce incorrect results over time. For example, if a complex Machine Learning model is used, it may scrutinize each given (training) data point intensively while overlooking the overall trends. This would give excellent results in the training stage but would struggle with unseen test data.

For example, let us look at the regression model of how house prices vary with the increase in the number of rooms of a house in Italy.

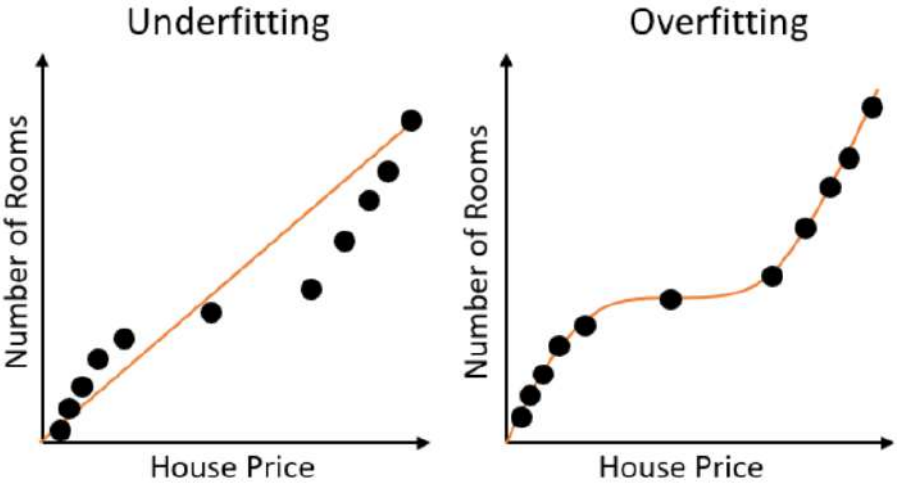


Figure 3: Underfitting vs Overfitting (Image by Author).

The more complicated model (over-fitting) would be extremely accurate while estimating the prices of houses in Italy (training data) but would be outperformed by the simple linear regression model (underfitting) when used to estimate the house prices in France (unseen test data).

A consequence of the Accuracy Paradox is that to train our model to produce better results over time, the accuracy of estimates during production would fall. To account for this paradox, a model complexity penalty (or regularization) can be used. This reduces weights associated with less important factors and focuses the model's attention on features that give the most information. Accuracy Paradox also illustrates why accuracy is not a good measure of performance of a predictive model.

Student ARTICLES

References:

1. *https://towardsdatascience.com/paradoxes-in-data-science-cab0869ef23d*
2. *https://brilliant.org/wiki/introduction-to-paradoxes/*
3. *https://pubmed.ncbi.nlm.nih.gov/29484824/*
4. *https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b*
5. *https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765*
6. *https://ed.ted.com/lessons/how-statistics-can-be-misleading-mark-liddell*

Student ARTICLES

Interpretation of Statistics in Astronomy

- Adrija Bhar (2nd Year), Tamasha Dutta (2nd Year), Manopriya Pal (2nd Year)

Introduction

In our article, we are going to discuss some applications of statistics in astronomy and their interpretations. Many problems in astronomy had led to the development of many statistical methods, from traditional and classical methods such as least squares estimation to modern methods like nested sampling.

However, with the development of technology, especially in the field of computer science, it's now possible for astronomers to collect a large amount of data from the universe. Is there any kind of extraterrestrial life out there? When will the universe come to an end and how? How many stars in each galaxy are similar to the Sun? These are the most common questions which can only be answered after collecting, analyzing and interpreting the astronomical data.

This is when the need arises for a new field of study that is known as Astro statistics. It is a prominent field of statistics which is used as a tool to analyze and interpret the big databases relating to the universe.

So, the challenge of Astro statistics is to build statistical tools to refine and interpret massive amount of data from space and with the help of that to generate relevant and useful information that answers the most intriguing and big questions of astronomy.

History of Statistics in Astronomy

Astronomy is the oldest observational field of science. Now, let us see how the application of Statistics, in Astronomy, gradually increased day by day.

- The Greek natural philosopher, Hipparchus, made one of the first applications in finding scatter of the length of a year in Babylonian measurements, defined as the time between solstices, he took the middle of the range, rather than the mean or median, to estimate the value. Today this method is known as the 'Midrange estimator of location'.
- In the sixteenth century, Tycho Brahe and Galileo Galilei promoted the utility of the mean of discrepant observations to increase precision. Galileo also gave his insights about the properties of errors in nonmathematical language in his 'Dialogue on the Two Great World Views, Ptolemaic and Copernican'. Later these properties were incorporated by Gauss into his quantitative theory of errors.

Student ARTICLES

In 18th century, some astronomers tried to tackle numerous inaccurate astronomical data, combine those observations and estimate the physical quantities through celestial mechanics more earnestly.

- In 1767, British astronomer John Michell applied a significance test, based on the uniform distribution, to show that the 'Pleiades (The Seven Sister Star Cluster)' is a physical grouping of stars. Though there were some technical errors in his techniques.
- Bernoulli and Lambert had laid the foundations of the concept of maximum likelihood which was later developed more thoroughly by Fisher in the early 20th century.
- Through a complex and difficult course of reasoning, Laplace proved that method of least squares was the most convenient method for finding parameters in orbital models from astronomical observations.
- In 1733, the mathematician Abraham De Moivre used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin. This approximation is known as 'Central Limit Theorem'. Later improvements were developed by Simeon Denis Poisson Friedrich Bessel.

Student ARTICLES

In the 20th century, we witnessed a gap between statistics and astronomy. None of the major works regarding Astronomy, like, discovery of the interstellar medium, analysis of the geometry of the Galaxy, discovery of extragalactic nebulae; did not involve statistical theory or application. The modern field of Astro statistics grew suddenly and rapidly starting in the late 1990s.

- In the 1920's, Fisher formulated the method of maximum likelihood estimation. They were instrumental in discovering galaxy streaming towards the "Great Attractor" (a gravitational anomaly in intergalactic space and the apparent central gravitational point of the "Laniakea Supercluster") and in computing the 'galaxy luminosity function' from flux-limited surveys.
- In the year 1970, the nonparametric Kolmogorov–Smirnov statistic was used for the first time for 'two-sample' and 'goodness-of-fit' tests.
- Since 1990's, Bayesian classifiers for discriminating stars and galaxies are used to construct large, automated sky survey catalogs.

Use of Regression Analysis in Astronomy

In Astronomy, regression analysis is one of the most frequently used statistical techniques. Most astronomical data analyses feature intrinsic scatter regarding the regression line.

In the year 2007, scientist Kelly described a Bayesian Method (MLINMIX) based on the likelihood function of the measured data (Later, in 2015, Scientist Mantz extended the MLINMIX algorithm to the case of multiple response variables). The method will account for measurement errors, intrinsic scatter, multiple independent variables, non-detections, and choice effects within the variable quantity.

Later, in the year 2014, scientist Maughan suggested a model to constrain simultaneously the form and evolution of the scaling relations. The method distinguishes between measured values, intrinsic scattered values and model values and can constrain the intrinsic scatter and its covariance.

Use of Linear Scaling of Bayesian Method in Astronomy:

Case 1: Case of Scattered Quantities.

Most of the scaling relations in astronomy are time evolving different power-laws. This straightforward schematics is supported by different types of observations, theoretical considerations, and numerical simulations.

The general form of the relation between two properties, e.g., the observable O and the mass M, is,

$$O \propto M^{\beta} F^{\gamma}{}_{z} \text{----------(1)}$$

Where, β is the slope and also the red shift evolution within the median scaling relation, which is accounted by the factor F_z. According to the context, the redshift factor F_z may be the factor (1 + z). In logarithmic variables, the scaling relation is linear, and the scatter is Gaussian,

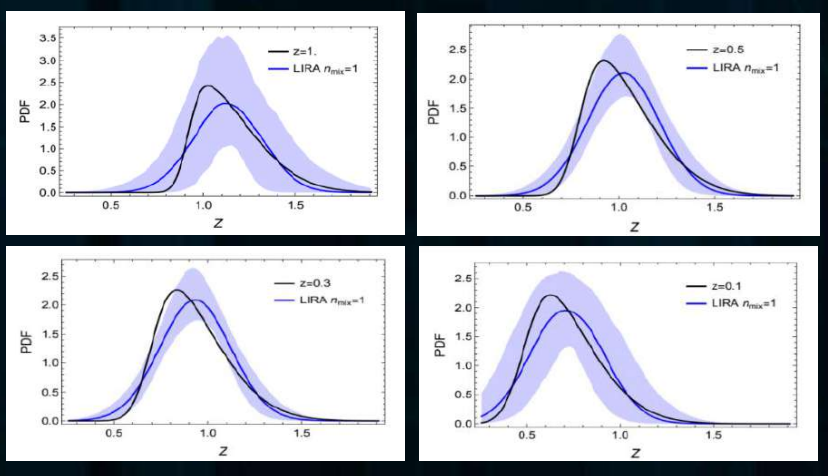
$$\log O = \alpha + \beta \log M + \gamma \log F_z \text{ ----------(2)}$$

In the usual framework, the time ($T = \log(F_z)$) evolution does not depend on the mass scale and only affects the normalization. However, the interplay between different physical processes that can be more or less effective at different times and can make the slope time dependent, β.

Assuming that the evolution of the slope with redshift is linear in T, Equation (2) may be generalized as,

$$Y = \alpha + \beta X + \gamma T + \delta X T \text{ ----------(3)}$$

Where, X = log M, and Y = log O. The time variable T is deterministic, not affected by measurement of errors and the variable X is random.



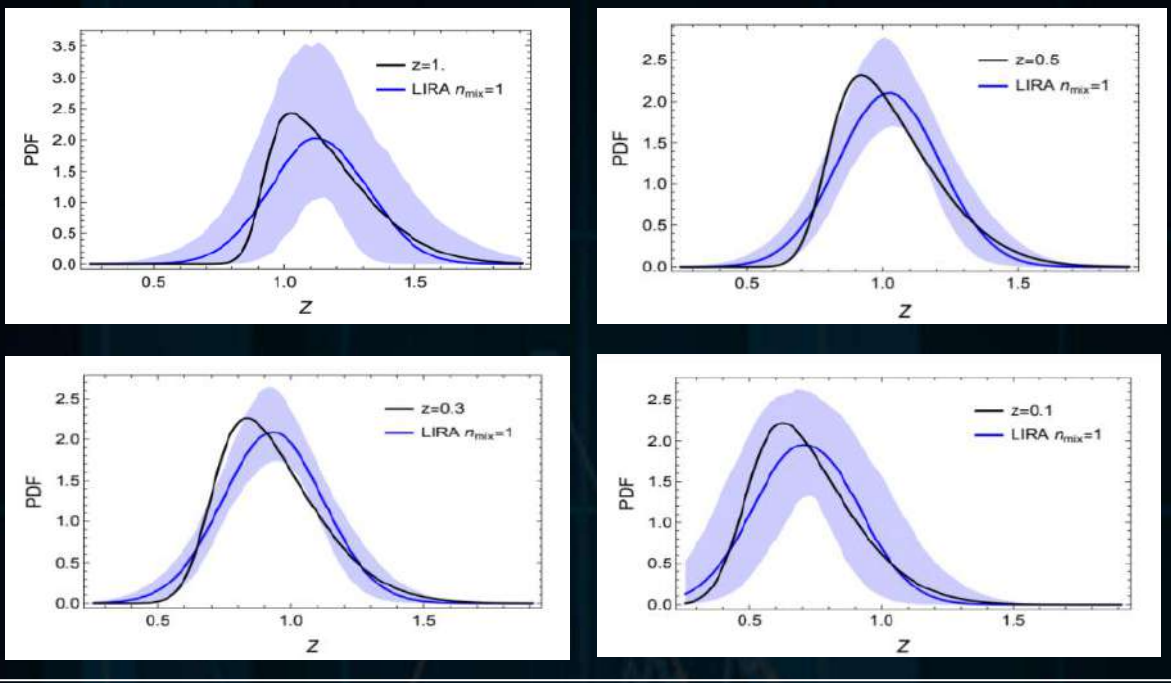


Figure 1:

The reconstructed intrinsic distribution of the independent variable Z at different redshifts, for, z = 1.0, 0.5, 0.3, 0.1. The black line is the input distribution, the blue line is the median reconstructed relation, the shadowed blue region encloses the 1-σ confidence region for each value of Z. For a total of n_{sample} = 100 data.

Case 2: Case of Unscattered Quantities.

Also, the linear relation between two Unscattered quantities can be expressed, as,

$$YZ = \alpha_{Y|Z} + \beta_{Y|Z} Z + \gamma_{Y|Z} T + \delta_{Y|Z} Z T \text{---------- (4)}$$

Where, α denotes the standardization, the slope β denotes the dependence with Z, the slope γ denotes the time-evolution of the standardization and δ quantifies the tilt of the slope with time.

Departure From Linearity:

Different physical processes are effective at different scales, which may cause deviation from linearity. Gravity is the actuation behind formation and evolution of galaxy clusters however at small scales baryonic physics will play a distinguished role. As a result, linearity can break. This can be shaped with a knee in this relation, such that before the breaking scale Z_{knee}, the scaling as follows

$$Y_Z = \alpha_{Y|\,Z,\,knee} + \beta_{Y|\,Z,\,knee}\, Z + \gamma_{Y|\,Z,\,knee}\, T + \delta_{Y|\,Z,\,knee}\, Z\, T \text{---------- (5)}$$

The standardization $\alpha_{Y|\,Z,\,knee}$ and the time evolution $\gamma_{Y|\,Z,\,knee}$ is determined by requiring equality at the transition Z_{knee},

$$\alpha_{Y|\,Z,\,knee} = \alpha_{Y|\,Z} + (\beta_{Y|\,Z} - \beta_{Y|\,Z,\,knee}) \text{---------- (6)}$$

$$\gamma_{Y|\,Z,\,knee} = \gamma_{Y|\,Z} \text{---------- (7)}$$

The transition between the two regimes is often modeled through a transition perform,

$$f_{knee} = 1 / 1 + exp\,[(Z - Z_{knee}) / l_{knee}] \text{---------- (8)}$$

where the size l_{knee} sets the transition length. The relation over the full range reads

$$Y_Z = \alpha_{Y|Z} + \beta_{Y|Z} Z + \gamma_{Y|Z} T + \delta_{Y|Z} Z T + (Z_{knee} - Z) f_{knee}(Z) \times (\beta_{Y|Z} - \beta_{Y|Z, knee}) + (\delta_{Y|Z} - \delta_{Y|Z, knee}) T \text{ ---------- (9)}$$

Similar physical processes will have an effect on the scatter too, which is modeled as

$$\sigma_{Y|Z}(Z, z_{ref}) = \sigma_{Y|Z,0} + (\sigma_{Y|Z,0,knee} - \sigma_{Y|Z,0}) f_{knee}(Z) \text{ ----------(10)}$$

Assuming that the redshift evolution of the scatter is not affected.

Remark:

Bayesian linear regression models have involved a large number of parameters. Since all relations in the model are expressed as conditional probabilities, thus, the posterior can be efficiently explored. All of these procedures have their own specifications and strengths that can make them preferable under some given circumstances. It also allows the consistent treatment of time-evolution, intrinsic scatter, and selection effects. Deviations from linearity relations with knees can be accounted for. Thus, the feature of a linear regression model, to stay simple and to add complexity if needed, is then important.

Student ARTICLES

Inference in Astronomy

Statistical inference is an art of forming an idea about an unknown population on the basis of a completely known sample. It is not possible to observe all the units in a population when the population size is very large. Astronomers often work with stars and planets. Now the problem is that there are countably infinite numbers of stars in a galaxy. In this situation we take the help of statistical inference to estimate the desired characteristics of the population. Astronomers usually take a sample from the population and measure the properties of the sample to know the properties of the vast underlying population of similar objects in the Universe.

It comes into the picture when the astronomer:

- Smooths over separate observations to know the underlying continuous development
- Seeks to quantify relationship between determined properties
- Tests to know an observation matches with an assumed astrophysical theory
- Subdivides a sample to compensate for flux limits and no detections
- Investigates the temporal behavior of variable sources

- Infers the evolution of cosmic bodies from studies of objects at totally different stages
- Characterizes and models the patterns in wavelength, pictures or space and lots of different things.

Now, in order to know some properties of the population i.e., to find the value of the parameter of interest first we need to find a good estimator of this parameter. The method of moments, least squares (LS) and maximum likelihood estimation (MLE) are vital and normally used procedures for constructing estimates of the parameters.

Suppose θ be an unknown parameter of the distribution of a variable X and T is an estimator for estimating θ on the basis of a random sample (X_1, X_2, ..., X_n). Now, T is said to be a good estimator if, for all $\xi>0$, $\eta>0$, however small, it is possible to find a n_0, depending on ξ, η, such that,

$$P[|T-\theta| \leq \xi] > 1-\eta, \text{ whenever } n \geq n_0.$$

An alternative method is to provide an interval within which the parameter may be supposed to lie. This is called interval estimation. The confidence interval of a parameter θ, a statistic derived from a dataset X, is outlined by the range of lower and higher values [l (X), u (X)] that depend upon the variable(s) X outlined such that

$$P[l(X) < \theta < u(X)] = 1 - \alpha \quad [\alpha=\text{level of significance}]$$

Where 0 <α< 1 is usually a small value like α = 0.05 or 0.01. That is, if θ is the true parameter, then the coverage probability that the interval [l(X), u(X)] contains θ is at least 1 − α. The quality of confidence intervals is judged using criteria including validity of the coverage probability, optimality (the smallest interval possible for the sample size) and invariance with respect to variable transformations.

Bayesian Parameter Estimation:

In the Bayesian approach, we can test our model in the light of our data and see how our degree of belief in its 'fairness' evolves, for any sample size, considering only the data that we did actually observe.

P (model| data, I) = k * P (data| model, I) * P (model| I)
[k=proportional constant]

Bayesian inference relates the probability of model parameters θ to experimental data d, and a hypothesis for the data H, via Bayes theorem P(θ| d, H) = π(θ| H) L(d| θ , H) Z(d| H) . Here, P(θ| d, H) is the posterior probability density of the parameters θ given d and H; L(d| θ, H) is the likelihood of d given θ and H; π(θ| H) is the prior probability of θ; and Z(d| H) is the evidence of d given H.

Student ARTICLES

Bayesian parameter estimation is the workhouse of gravitational-wave astronomy. For instance, determining the mass and spins of merging black holes, revealing the neutron star equation to state and unveiling the population properties of compact binaries. It is the tactic by which gravitational-wave data is employed to infer the sources' astrophysical properties.

There are mainly two scales that verify the overall computational cost of Bayesian inference. They are, (i) The price of evaluating parameterized models of the data, and (ii) the rate of convergence of the sampling algorithms. The typical wall time may be calculated firstly by considering the total CPU time. To the leading order, the CPU time, T_c, of Bayesian inference scales such as the average call-time of the data model $\langle T_m \rangle$, multiplied by the total number of calls to the likelihood function N of the stochastic sampling algorithm $T_c = N\langle T_m \rangle$. We have a tendency to treat N as an overall standardization which is typically N ~O (10^7). Once serial sampling algorithms are used, the CPU time T_c is equal to the wall time T_w. The average call-time $\langle T_m \rangle$ is powerfully passionate about the complexness of the GW signal models, and perhaps depends on the models for the noise.

Application:

For example, suppose a spacecraft is sent to a moon of Saturn and, employing a penetrating probe, detects a liquid ocean deep beneath the surface. Now if the thermometer encompasses a fault, then it can't felicitate to detect the actual temperature of the liquid. By Bayesian hypothesis testing we are able to determine the temperature of the liquid assuming it water and then assuming it fermentation alcohol.

Prior Distribution:

Prior distribution is a probability distribution of attainable values for an unknown population characteristic that is developed before one obtains any current data observations about the phenomenon of interest. A posterior probability is the probability that assigns observations to groups for the given data. Samples are accepted subject to the limitation that those drawn on subsequent iterations have a better likelihood than those on previous iterations. The algorithm rule is seeded by drawing a variety of K live points from the prior. These points are arranged from highest to lowest likelihood. The algorithm then proceeds by drawing samples θ_i from the prior on each iteration i. The aim is to exchange the live point with the lowest likelihood L_{min}, on every iteration.

Model Selection:

Model selection is the task of choosing a statistical model from a collection of candidate models, given the data. Within the simplest cases, a pre-existing set of data is taken into account. It may be a polynomial also.

A good model selection technique will help to balance goodness of fit with simplicity.

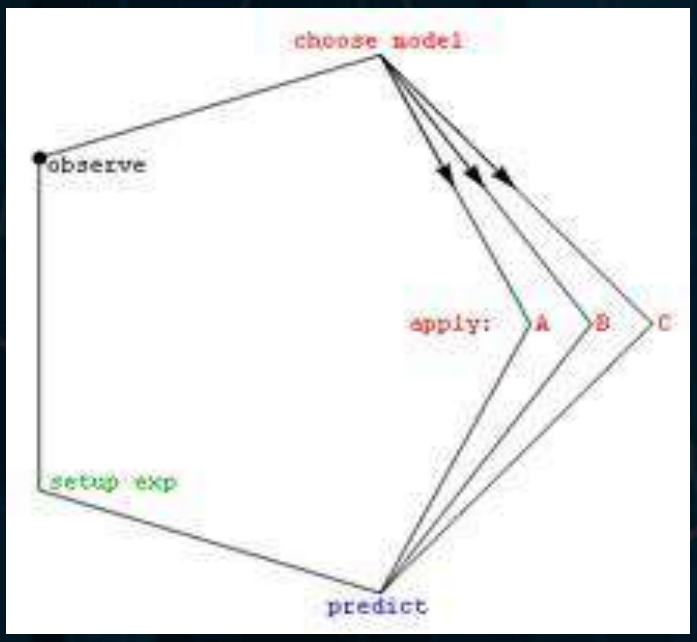


Figure 2:
The Scientific Observational Cycle.

Likelihood Based Model:

Let us assume that U denote the observed data and let M1, ..., Mk denote the models for U, under consideration.

For every models, Mj, let L(U| θj; Mj) and (θj) = ln f (U| θj; Mj) denote the likelihood and loglikelihood respectively, where θj is a ρj-dimensional parameter vector. Here L(U| θj; Mj) denotes the ρ.d.f or the ρ.m.f evaluated at the data U. For most of the cases, we have a comparison between two models, M1 and M2. The model M1 is known to be nested in M2 if some elements of the parameter vector θ1 are fixed (and possibly set to zero), i.e., θ2 = (α, γ) and θ1 = (α, γ_0), where γ_0 is some known fixed constant vector. Comparison between M1 and M2 can then be considered as a classical hypothesis testing problem where the null hypothesis, H_0: γ = γ_0. The nested models of this type appear frequently in different astronomical modeling. In the astrophysical modeling, stellar photometry might be modeled as a blackbody (M1) with absorption (M2), the structure of a dwarf elliptical galaxy might be modeled as an isothermal sphere (M1) with a tidal cutoff (M2), or hot plasma might be modeled as an isothermal gas (M1) with nonpolar elemental abundances (M2).

Let, X follows N (μ, σ^2) ($M1$) and Y follows N (0, σ^2) ($M2$), where μ and σ^2 both unknown.

To test, H_0: μ = 0 against H_1: μ ≠ 0

Suppose T is the test statistic and c is the critical point

We reject $H_{0,}$ if ($|T|>c|H_0$)

Model selection and goodness-of-fit is very much dependent on the sample size, if sample is large, very small discrepancies lead to rejection of the null hypothesis, while for small samples, even large discrepancies might not result in rejection.

Nonparametric Statistics

We have limited knowledge about the planets, stars, galaxies or accretion phenomena which provides very little information about the underlying conditions. In some situations, the mathematical assumptions related to the statistical procedures are often not well explained. As a result, some interesting characteristics or facts about the data may be obfuscated by using a simplistic model. So, astronomers need nonparametric methods because in those cases no assumptions regarding the underlying probability distribution are needed. An example of such a nonparametric method is the Kolmogorov–Smirnov (KS) two-sample test. Modern nonparametric statistical tools may best uncover the underlying mass distribution of galaxies which have complicated tri-axial structures dominated by dark matter without the simplification and physically unrealistic assumptions of the analytical formulae.

Thus, nonparametric approaches to data analysis are very much needed in Astro statistics. Nonparametric methods can provide us considerable capability to analysis and interpret the astronomical data.

An application in Astrostatistics:

Using nonparametric methods, two astrophysicist Robert Nichol, Chris Miller and two statisticians Larry Wasserman, Christopher Genovese have established the three-peak structure of the cosmic microwave background fluctuation spectrum, without using the high precision of parametric modeling. Here, Figure 3 shows the three-peak structure of the cosmic microwave background fluctuation spectrum using nonparametric methods.

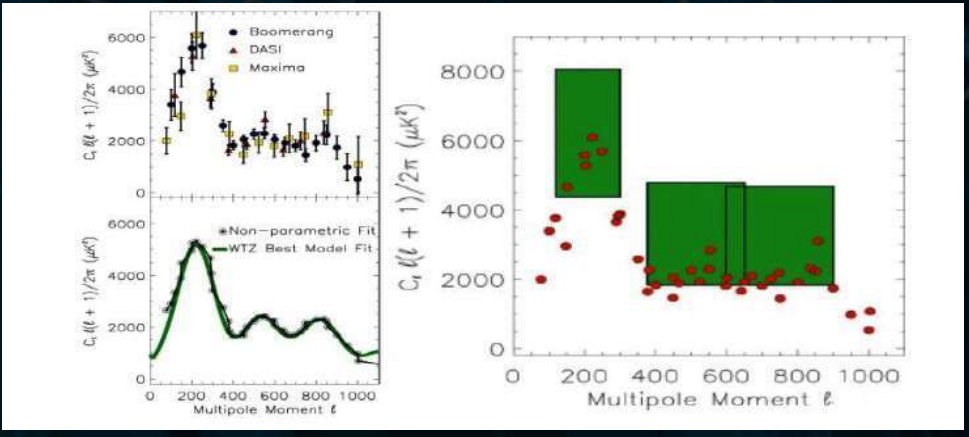


Figure 3:

The three-peak structure of the cosmic microwave background fluctuation spectrum using nonparametric methods.

Examples of nonparametric methods:

Some examples of nonparametric methods are:

1. Nonparametric Density Estimation
2. Tests of hypotheses without parameters
3. Nonparametric Goodness-of-Fit tests
4. One-sample Kolmogorov-Smirnov Test
5. Two-sample Kolmogorov-Smirnov Test
6. K-sample tests, Correlation coefficients
7. Nonparametric Regression:
    - Kernel Estimation
    - K-Nearest Neighbor Estimation
    - LOESS Estimation

Student ARTICLES

TYPES OF DATA USED IN ASTRONOMY

To make an Astronomical study, with the help of Statistical Methods, we need some data on the basis of particular field, in Astronomy. For this Astronomical study, we have some special Datum. Mainly three types of Data are used for this study. These are: (A) Image Data (B) Spectral Data and (C) Time series and Functional Data.

Image Data, in Astronomy:

For astronomical Study, Image data is one of the most essential data types. The images taken by different telescopes is used as the Astronomical Data. One of the most useful telescopes is the Dark Energy Camera (DECam). It takes the images, as the part of the Dark Energy Survey (DES), with a photometric filter which blocks certain light wavelengths. Here, the Figure 4 shows one night sky image, taken by DECam.

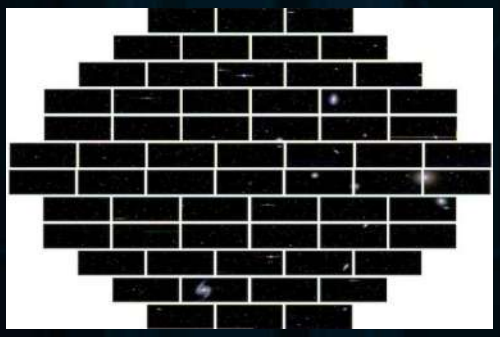


Figure 4:
The night sky image, taken by the DECam. The white lines are gaps between the CCDs.

Figure 5:
Dark Energy Camera (DECam).

Spectral Data, in Astronomy:

Spectral Data is another important tool for analyzing Astronomy. Basically, spectrum represents the intensity of light in different wavelengths, providing considerably more information than can be directly inferred from image data. Figure 6, shows the structure of the Spectrums. In Figure 7, it shows the spectrum of the galaxy Messier 77, a barred spiral in the Cetus constellation.

Spectrums basically, carry information about some of the most important properties like, temperature, chemical composition, etc.

With the development of large spectrographic surveys, a large amount of spectral data can be easily obtained through various data mining methods, to assist astronomers' spectral analysis.

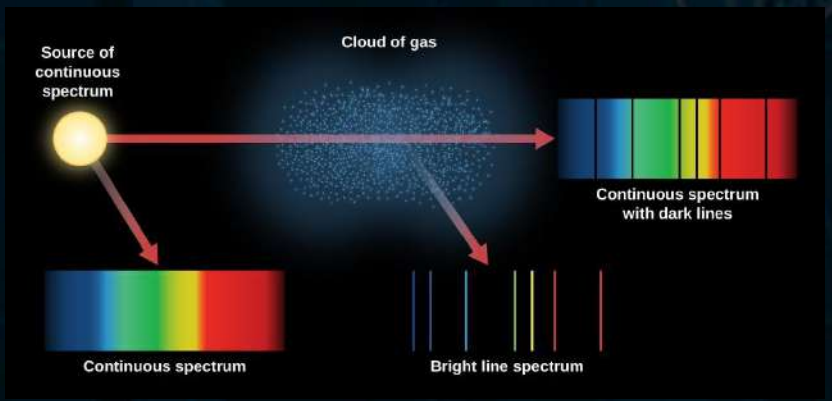


Figure 6:
Structure of the Spectrums.

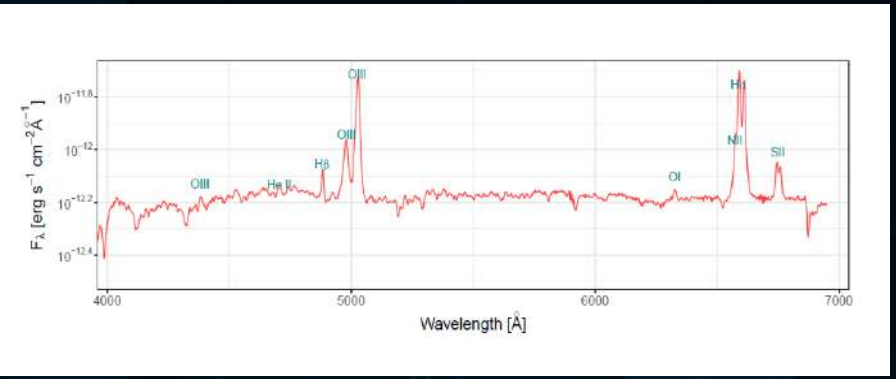


Figure 7:
Example of a galaxy spectra from Messier 77.

Time Series and Functional Data:

Time series data serves an important role in Astronomical analysis. There are various kind of variable objects exist in this universe, including lots of stars, with their predictable behavior, various kind of objects with their behavior, which are inherently unpredictable, and objects with both predictable and irregular variability, in their patterns. To understand the nature of variability, of different objects in the universe, the Astronomers often use the time series data to predict their behaviors.

In Figure 8, it shows a time series data, for a certain star, observed by the Optical Gravitational Lensing Experiment (OGLE). The data is represented in two types of structures, one is represented by orange crosses and another one with blue circles, over the course of approximately 10 years.

The time spacing between each of the observations, is irregular, which is a typical feature in astronomical data. In this experiment, approximately 400,000 of these light curves are collected. Also, in Figure 9, it shows the time series data plot on 3 types of random stars (Produced from Rebbapragada et al. [Rebb09]).

In astronomy, time series analysis of flux measurements is very common. As a consequence of many decades of observations in stars, a large variety of flux variations have been detected by many astronomical objects, including periodic variations, such as, pulsating stars, pulsars, rotators, eclipsing binaries, planetary transits; quasi-periodic variations, such as, star spots, active galactic nuclei, neutron star oscillations, etc. Such astrophysical phenomena are wavelength-specific cases or were discovered as a result of wavelength-specific flux variations, like soft gamma ray repeaters, X-ray binaries and gravitational waves.

In addition to flux-based time series analysis, astronomical data also include motion-based time series data. These include the trajectories of different planets, comets, and asteroids in the Solar System, the motions of stars around the massive black hole from the center of the Milky Way galaxy.

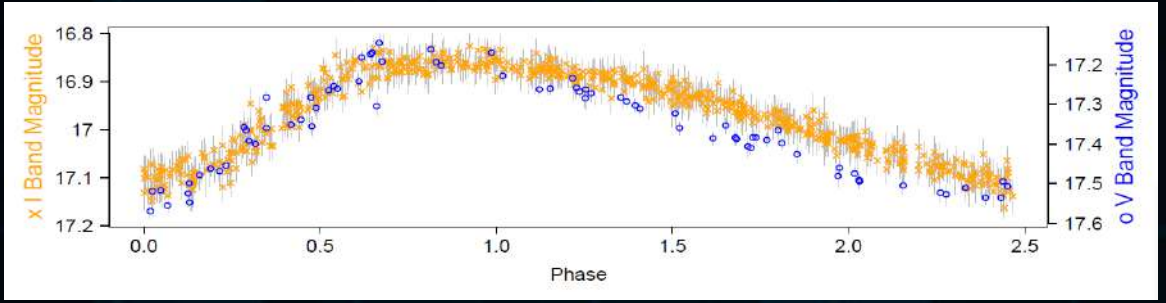


Figure 8:
Light curve of a variable star observed by OGLE. Models from the time series and functional data analysis literature are often used for studying these objects.

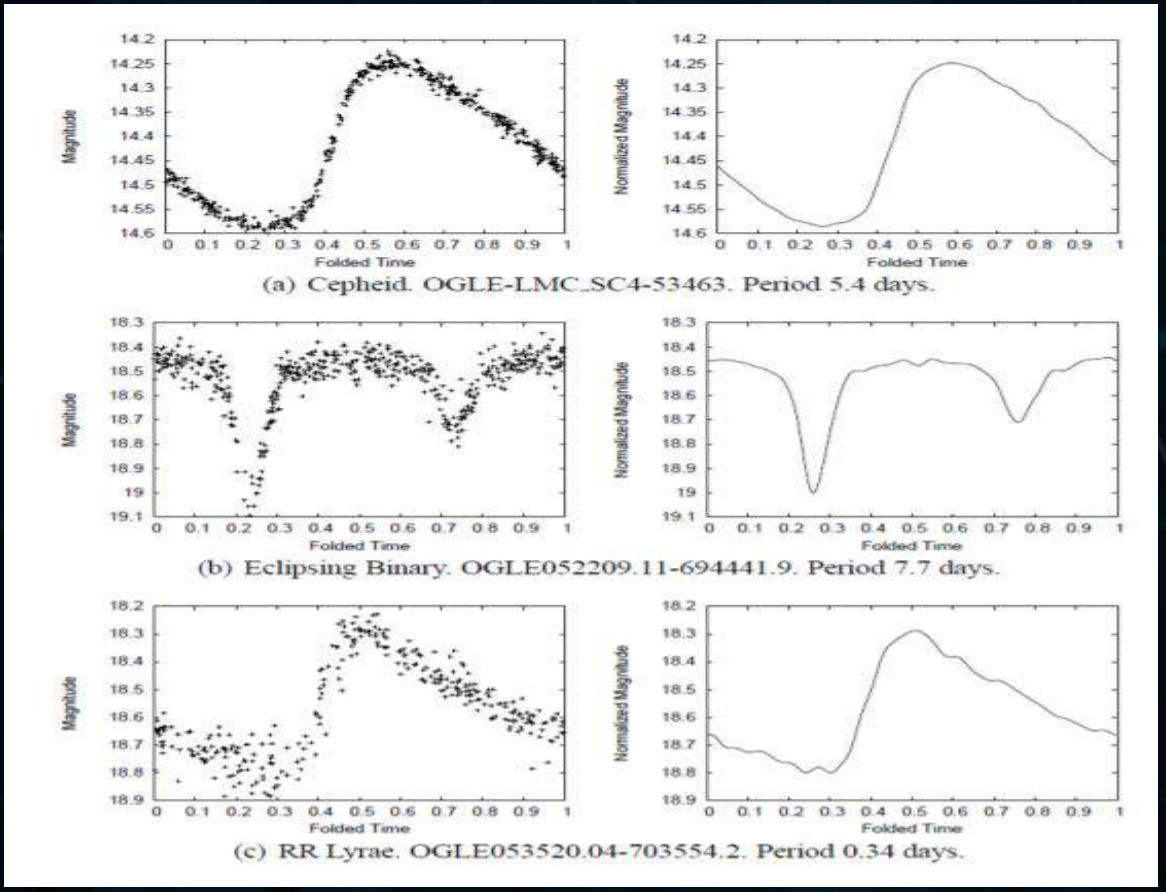


Figure 9:
Examples of time series data for 3 different types of variable stars (reproduced from Rebbapragada et al. [Rebb09]).

Conclusion

In the past, astronomers did everything one by one, from the conception of a project to assortment of information and their analysis. As the instruments become complicated, teams had to be come upon and that they more and more enclosed folks specialized in technology. Now it's not possible to run a project at the forefront of astronomical analysis while not facilitate of those technologies. Thus, we ought to work on larger samples if we wish to require advantage and fully use the information contained all told these data, globally and one by one. A technique to figure with efficiency of massive samples is to use and if necessary to develop associate degree adequate statistical methodology.

Now, we will nearly perceive why statistics is so much important in astronomy. Astronomers are absolutely obsessed with it, without any application of statistics it's impossible to move on and answer such questions: Can all the eight planets ever line up to an equivalent aspect of sun? Is it possible to see the ring of Saturn from the equator or the poles of it? Is there a center of the universe? What came before the Big Bang? How to measure a galaxy's distance? What percentages of stars are there within the Milky Way?

In the application of distance scale, more specifically, analysis of errors and calculation of regression coefficients are two important things in measuring the speed of enlargement of the universe (one of the foremost necessary parameters in cosmology), estimating the age of the universe and uncovering massive scale phenomena like super clustering and galaxy streaming.

Multivariate methods such as Principal Components Analysis (PCA) and Cluster Analysis are the most common methods that are used in astronomy. Often clustering method is applied to choose anomalous or peculiar objects. These techniques will also work in multidimensional parametric space. In stellar astronomy, one must study the interface between photometry and spectrographic analysis, particularly within the framework of stellar classification. In star-galaxy separation process also we seek to look for the assistance of regression and inference.

So, it's clear that in order to figure with astronomical data, we will be benefitted if we have a tendency to take help of statistics. In order to manage all kinds of statistical data, its method and mindsets are very crucial during this developing society toward a better and sustainable future.

References:

1. Sources of the different Figures:

1. Figure 1: https://www.google.com/url?sa=t&source=web&rct=j&url=https://academic.oup.com/mnras/article/455/2/2149/1111686&ved=2ahUKEwiX0YGjxfnzAhUEzzgGHfn9CtQQFnoECBwQAQ&usg=AOvVaw1sJyNI60vBpjPVlfeAs__O (Page 8)
2. Figure 2: https://images.app.goo.gl/8hGwjWeBAvt5vsSQ6
3. Figure 3: Nonparametric and robust statistic by Eric Feigelson (3rd INPE Advanced School in Astrophysics: Astro statistics 2009) (Page 5)
4. Figure 4: https://www.google.com/url?sa=t&source=web&rct=j&url=https://arxiv.org/pdf/1707.05834&ved=2ahUKEwjo1PrsxPnzAhUkzDgGHUYaBA4QFnoECAQQBg&usg=AOvVaw1fNRPIXR_19HHnxtHBPaQ6 (Page 2)
5. Figure 5: https://www.darkenergysurvey.org/the-des-project/instrument/

6. Figure 6: https://courses.lumenlearning.com/towson-astronomy-2/chapter/formation-of-spectral-lines/

7. Figure 7: 834&ved=2ahUKEwjo1PrsxPnzAhUkzDgGHUYaBA4QFnoECAQQBg&usg=AOvVaw 1fNRPIXR_19HHnxtHBPaQ6 (Page 2)

8. Figure 8: 834&ved=2ahUKEwjo1PrsxPnzAhUkzDgGHUYaBA4QFnoECAQQBg&usg=AOvVaw 1fNRPIXR_19HHnxtHBPaQ6 (Page 3)

9. Figure 9: https://r.search.yahoo.com/_ylt=Awrxzw8jfdRhcTwAbAC7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1641344419/RO=10/RU=https%3a%2f%2fcs.gmu.edu%2f~jessica%2fpublications%2fastronomy11.pdf/RK=2/RS=L3uJLuHMdGQxwYG1EFZlPF2FQF4- (Page 3)

2. Sources of the Equations in Regression Analysis Section:

- https://www.google.com/url?sa=t&source=web&rct=j&url=https://academic.oup.com/mnras/article/455/2/2149/1111686&ved=2ahUKEwiX0YGjxfnzAhUEzzgGHfn9CtQQFnoECBwQAQ&usg=AOvVaw1sJyNI60vBpjPVlfeAs__O (Page 3 and Page 5)

Student ARTICLES

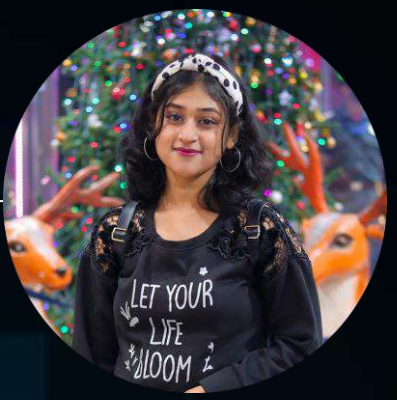
3. Sources of the Equations in Likelihood Based Model Selection:

https://www.iiap.res.in/astrostat/School10/LecFiles/Karandikar_Babu_ModelSelGOF_notes.pdf (Page 5)

For Further Readings:

- Modern Statistical Methods for Astronomy-by Eric D. Feigelson and G. Jogesh Babu
- Statistical methods in astronomy by James P. Long and Rafael S. De Souza https://www.google.com/url?sa=t&source=web&rct=j&url=https://academic.oup.com/mnras/article/455/2/2149/1111686&ved=2ahUKEwiX0YGjxfnzAhUEzzgGHfn9CtQQFnoE CBwQAQ&usg=AOvVaw1sJyNI60vBpjPVlfeAs__O
- https://www.google.com/url?sa=t&source=web&rct=j&url=https://core.ac.uk/download/pdf/25195154.pdf&ved=2ahUKEwiX0YGjxfnzAhUEzzgGHfn9CtQQFnoECBsQAQ&u sg=AOvVaw3jxIG7Jj1Dyn8k4MVl7Ub0
- https://www.google.com/url?sa=t&source=web&rct=j&url=https://arxiv.org/pdf/1707.05834&ved=2ahUKEwjo1PrsxPnzAhUkzDgGHUYaBA4QFnoECAQQBg&usg=AOvVaw 1fNRPIXR_19HHnxtHBPaQ6
- Modern Statistical Methods for Astronomy – NASA/IPAC Extragalactic Database
- Nonparametric and robust statistic by Eric Feigelson (3rd INPE Advanced School in Astrophysics: Astro statistics 2009)
- https://science.psu.edu/science-journal/winter-2020/astronomy-is-better-with-betterstatistics

Student ARTICLES



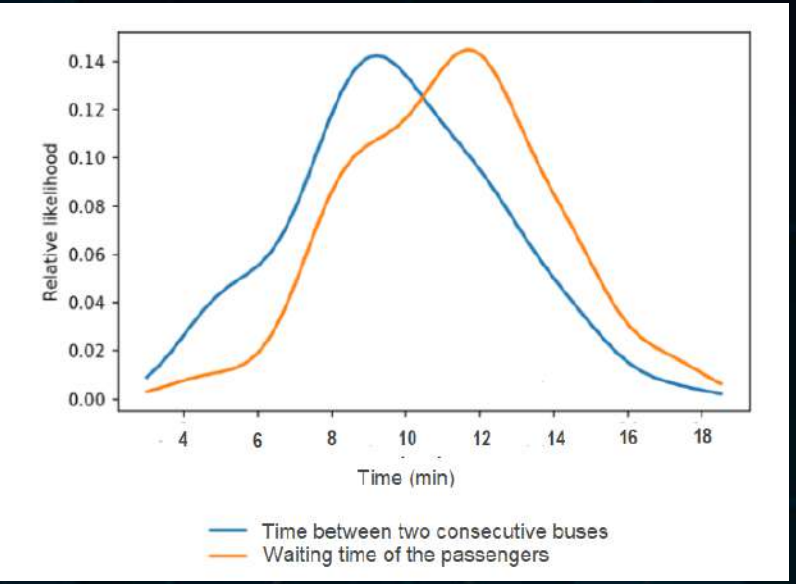
INSPECTION PARADOX

- Atreyee Roy (2nd Year)

Does it seem that every time you wait for a bus at a bus stop, the time you spend waiting is much more than what you expect? As you stand waiting for your bus you think that every vehicle that passes by must be your bus. An ambulance passes by but it's NOT YOUR BUS! So many cars pass by but still it's NOT YOUR BUS! The waiting time does not seem to end.

So, one day, while waiting at the bus stop, you decide to inquire some of your fellow passengers standing there, for how long they have to wait for the bus. You observe that you are not the only one who has been unlucky. Rather, many others, including yourself, have to wait for the bus for SO LONG. Since you need to pass the time somehow, you decide to calculate the mean of all the waiting times of your fellow passengers.

Now, one may expect that if the average time interval between the arrival of two consecutive buses is 10 mins, then mean of the waiting times will be near to $\frac{10}{2}$ = 5mins. However, to your utter surprise, the mean turns out to be 10.8 mins. In other words, the average waiting time of the passengers is greater than the average time between the consecutive bus arrivals!



Are your calculations wrong? Possibly not. Rather, you are just a victim of an actual phenomenon known as INSPECTION PARADOX.

Now, let us look into the matter. In real world, time intervals between consecutive buses arrivals fluctuate (they are not deterministic); some intervals are longer, and some are shorter. Passengers are more likely to arrive during the longer time intervals. Hence longer waiting times are given more weightage, which makes the mean more than the average time between the consecutive bus arrivals.

The inspection paradox suggests that data may be very useful in decision making. However, it also requires that it absolutely considers the observer. Who is the observer then? Is it the user or the service provider? It suggests that data may be confusing, if it is incorrectly presented.

Student ARTICLES

Box Model

Let's look at the phenomenon from a mathematical standpoint now.

Consider a hypothetical machine which repeatedly places a random number of objects in boxes. We assume the no. of objects in consecutive boxes is independent of each other. We want to determine the mean no. of items in the boxes. There are two methods to do so.

METHOD I: Choose a random box. Let N be a Random variable denoting the number of objects present in a randomly chosen box. Its probability mass function is known as: $f(x)_{x>=1}$

Hence the expectation of N is given by:

$$E(N)= \sum_{x>=1} x\, f(x)$$

METHOD II: Choose an item at random and then determine the no. of objects present in the box which contains that item (including the selected item). Let S be the Random Variable denoting the no. of objects in the box in which the chosen item belongs. Our aim is to find the expectation of S. Clearly here choosing an item at random is an indirect way of choosing a box. Hence, the probability of selecting a box is proportional to the no. of items it contains.

Consider a case where we have only two boxes such that

$$f(1) = \frac{1}{2} \quad \text{and } f(2) = \frac{1}{2}$$

Let B_i be the no. of objects present in i^{th} box.

- $P(S=1) = P(S=1 \mid B_1=1, B_2=1) \times P(B_1=1, B_2=1)$
  $+P(S=1 \mid B_1=2, B_2=1) \times P(B_1=2, B_2=1)$
  $+P(S=1 \mid B_1=1, B2=2) \times P(B_1=1, B_2=2)$
  $+P(S=1 \mid B_1=2, B_2=2) \times P(B_1=2, B_2=2)$

$$= 1\times \frac{1}{2}\times \frac{1}{2} + \frac{1}{3}\times \frac{1}{2}\times \frac{1}{2} + \frac{1}{3}\times \frac{1}{2}\times \frac{1}{2} + 0\times \frac{1}{2}\times \frac{1}{2}$$

$$= \frac{5}{12} \qquad [B1 \text{ and } B_2 \text{ are independent}]$$

- $P(S=2) = 1 - P(S=1) = \frac{7}{12}$

Thus, we see that, if we adopt method 1 to determine the mean no. of items in the boxes, each box is equally likely, i.e., P(N=1) = P(N=2). However, if we adopt method 2, it is more likely to choose the box with 2 items, since P(S=1)<P(S=2).

$$E(N) = 1\times \frac{1}{2} + 2\times \frac{1}{2} = \frac{3}{2}$$
$$E(S) = 1\times \frac{5}{12} + 2\times \frac{7}{12} = \frac{19}{12}$$

$$E(N) < E(S)$$

Hence using this box model, we see method 2 will give us a higher mean than method 1.

We can extend this model to various cases. For example, children in families, passengers in a car, residents in a country, students in a class etc. If we apply method 2, the likelihood of choosing a box with more items is proportional to the number of items it contains in each of these situations.

It is more likely to choose Texas than Idaho (with respect to the no. of people living there).

If we adopt method 2 as our sampling process, we are unintentionally inviting some kind of sampling bias in our sample. This is known as length bias sampling, where the probability of a unit to be in the sample is actually proportional to some kind of size, length, duration in time etc. Thus, it leads to Inspection Paradox, where subtly different sampling processes yield surprising results.

Inspection Paradox has found its application in many fields. Given below are some of the examples.

Determining Average Class size

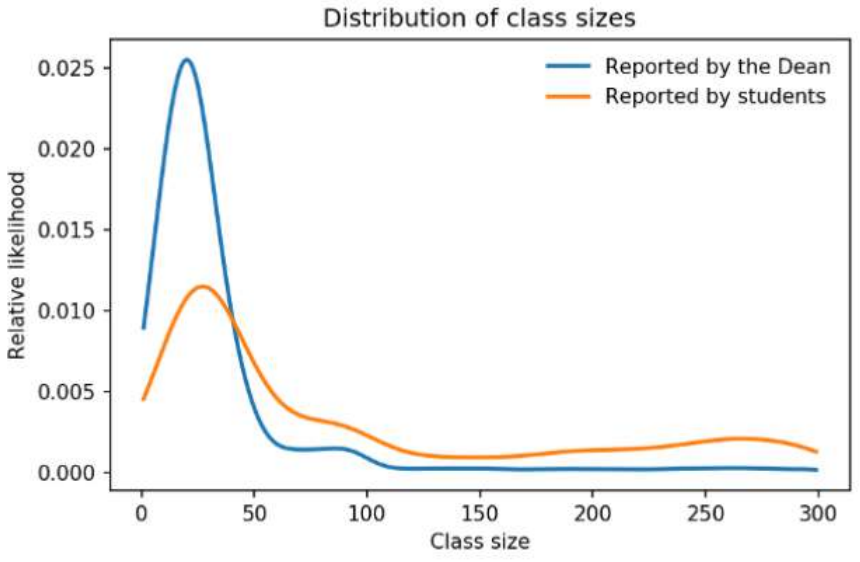
Assume we're asked to calculate a university's average class size. So, we adopt a sampling method where we will walk around the campus and ask the college students present there about how big their classes are. We will be recording their responses and calculating the mean. Surprisingly, the average we calculated is far higher than the real average of all class sizes at that university. How is this possible?

Student ARTICLES

The same experiment was conducted in Purdue University to determine the average class size for undergraduate classes in the 2013–14 academic year. The real average class size, as reported by the professors was 31, although the students reported it to be 56.The given class distribution is this:

No. of classes	Class strength
138	1 student
33	100+ students

Let's compare the case to our box-model to see why there's a difference. (Discussed in the previous section).

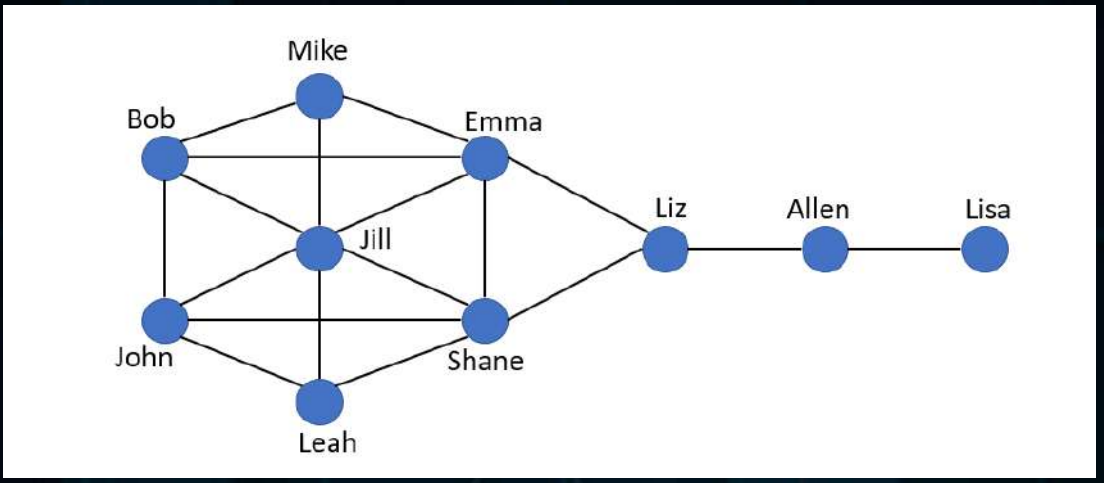


Undergraduate class sizes at Purdue University, 2013–14 academic year: Unbiased distribution as reported by the University and biased view as seen by students.

Consider each class of that University as a box and the students of that class as the contents of the box. Now, according to the sampling method we have adopted, if we go around the campus and randomly select students to ask, there will be only 138 students present there who will report the class size as 1. Whereas, there will be $33300+$ students whose answer will be more than 100. So, a student from bigger class size is more likely to be chosen. Bigger class sizes will have larger representation in our sample and hence more weight will be given to them. This sampling method adopted is similar to the method II as there is a sampling bias. Hence, this explains the reason behind this inspection paradox.

Friendship Paradox

Some people often get down on themselves. They believe they are not well-liked, or that their buddies have far more friends than they do. It's possible that they're more self-critical or introverted. However, it turns out that they are also the victims of the Inspection paradox. This particular case is known as friendship paradox. It was first described by the sociologist Scott Feld back in 1991 paper entitled "Why your friends have more friends than You". According to that paper, there are 80% chances that your friend has more friends than you do.



The above picture shows a network of people where each node is a person and each edge denotes the friendship between two nodes. First, let's try to figure out how many friends an individual has on average.

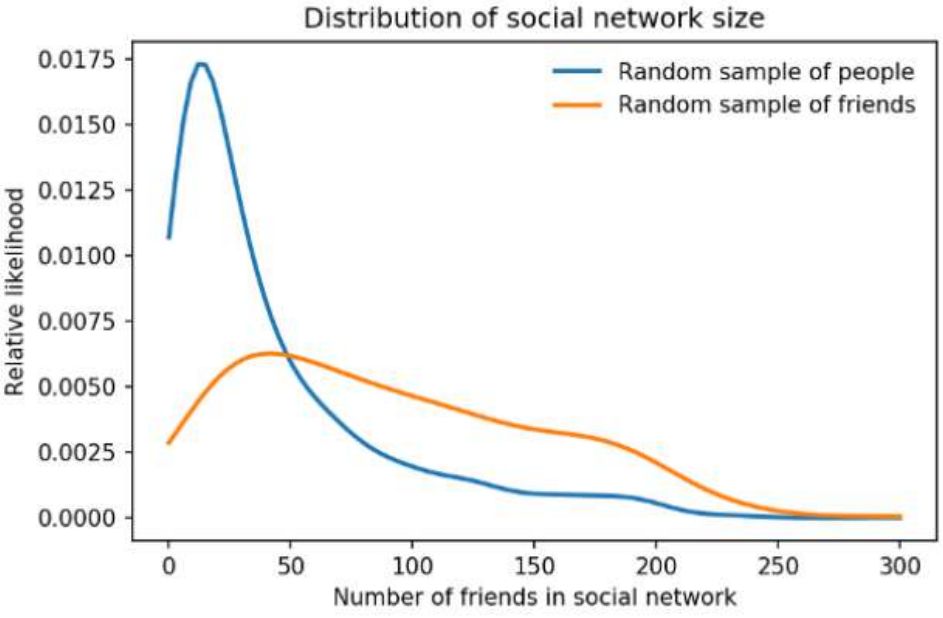
- Choose a random node (or a person) and ask how many friends he has and keep track of the numbers.
- Continue the process and find out the mean of those numbers.
- That's the number of friends a person has on an average (say M**1**).

The next step is to determine the average number of friends that a person's friend has. In order to do this, we will adopt a special method.

- Choose a random node and then inquire about the number of friends that node (or the person) has.
- Choose a random edge from that node which will lead us to another node and ask the same question.

- Continue this process and keep noting down the numbers to find the mean.
- This will be the average number of friends a person's friend has (say M2).

Now if we compare the two, M2 will be greater than M1. In the first case, when we are choosing a random node, every user is equally-likely. On the other hand, when we are choosing a random edge, it is more likely to choose someone who has more no. of friends. In other words, the probability of landing on a node is proportional to the number of edges which leads to that node. Hence, people with greater number of friends will get "oversampled" and thus will shift the mean towards the higher side.



Number of online friends for Facebook users: actual distribution and biased distribution seen by sampling friends.

If we grab the data from Stanford Large Network Dataset Collection, which includes a sample of about 4000 Facebook users (in 1998) and compute the no. of friends each user has and the number of friends their friends have, we would get the following graph.

It is evident from the plot, that in the second method of random sampling (which is, pick a node and then pick an edge) people with more no. of friends show up more no. of times than those who have lesser friends.

Length bias sampling is everywhere. It can be a subtle source of error. However, is it always a nuisance?

Inspection Paradox at rescue

You've definitely heard of the effective reproduction number, R, which is the average number of persons infected by each affected person during the epidemic. 'R' is significant since it determines the epidemic's large-scale path. The number of cases will rise exponentially as long as R is bigger than 1; if we can reduce R below 1, the number of cases will fall to zero.

In order to do this, we have two possible strategies for contact tracing.

The goal of "forward tracing" is to discover persons who the patient may have infected, while "backward tracing" attempts to find the person who infected the patient.

Student ARTICLES

According to a study conducted in Hong Kong, 19 percent of cases of COVID-19 were responsible for 80% of transmission, and 69 percent of cases did not transmit the virus to anyone. In other words, a limited number of super-spreaders are responsible for the majority of infections.

Now as a public health officer, which one do you find more effective? Is it forward tracing or backward tracing?

Backward tracking, according to the *inspection paradox*, is more likely to uncover a super-spreader and the others they have infected. Any new case is more likely to have arisen from a cluster of illnesses rather than from a single person, therefore it's worth going backwards to find out who else was linked to that cluster.

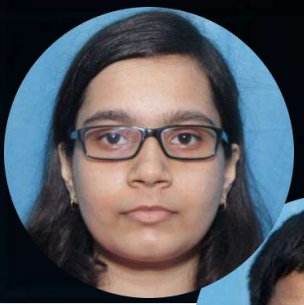
To quantify this effect, assume that 70% of infected patients do not infect anyone else and the remaining 30% infect 1 to 15 other people in a uniformly distributed manner. Let us say we find an infected patient, follow the trail, and find someone who was infected by the patient. We estimate that this person will infect 2.4 other people on average (which is a reasonable value of R). However, if we go backward and look for the person who infected the patient, we are *more likely* to find someone who has infected a large number of individuals, rather than someone who has just infected a few. In reality, the likelihood of finding a specific spreader is proportionate to the number of persons afflicted.

In conclusion, the inspection paradox manifests itself in a variety of fields, sometimes in subtle ways. It can produce statistical errors and erroneous inferences if you aren't aware of it. However, in many circumstances, it can be avoided or even employed as part of an experimental design on purpose.

Reference

1. *https://towardsdatascience.com/the-inspection-paradox-is-everywhere-2ef1c2e9d709*
2. *https://www.allendowney.com/blog/2021/08/19/covid-19-and-the-inspection-paradox/*
3. *William E. Stein and Ronald Dattero. Sampling Bias and the inspection paradox. Mathematical Association of America: Mar., 1985, pp. 96-99*

Student ARTICLES



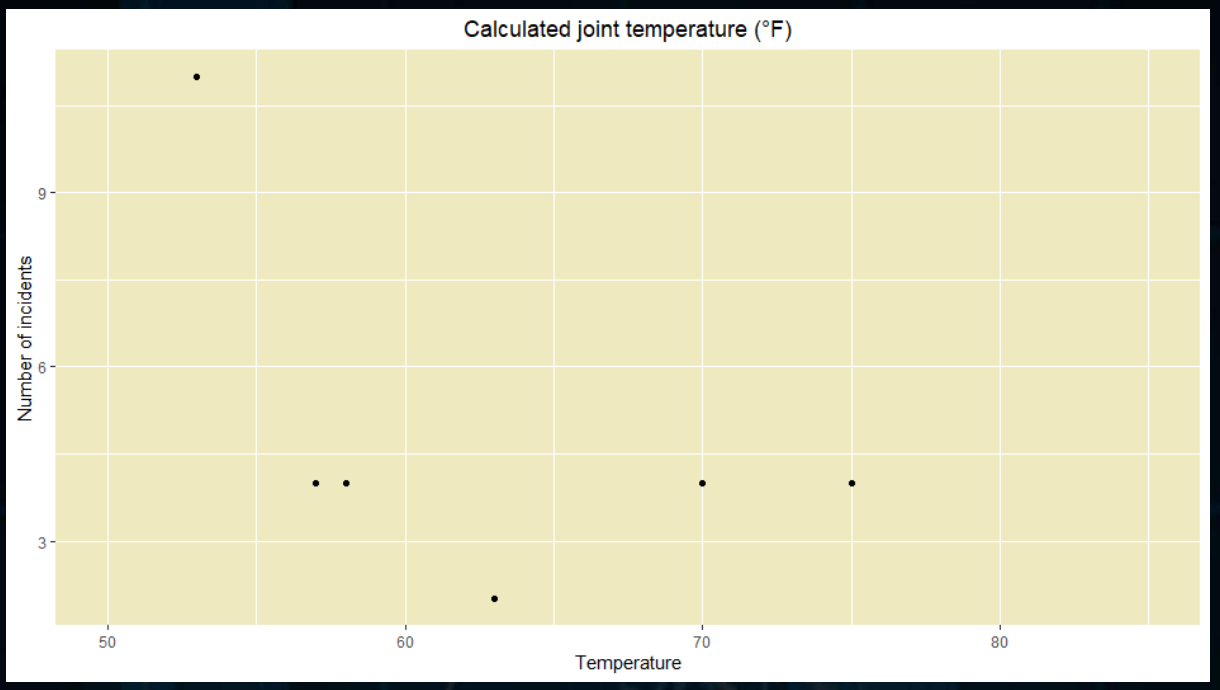
DARK DATA: THE DATA BEYOND OUR REACH

- Saikat Datta (3rd Year), Saheli Datta (3rd Year)

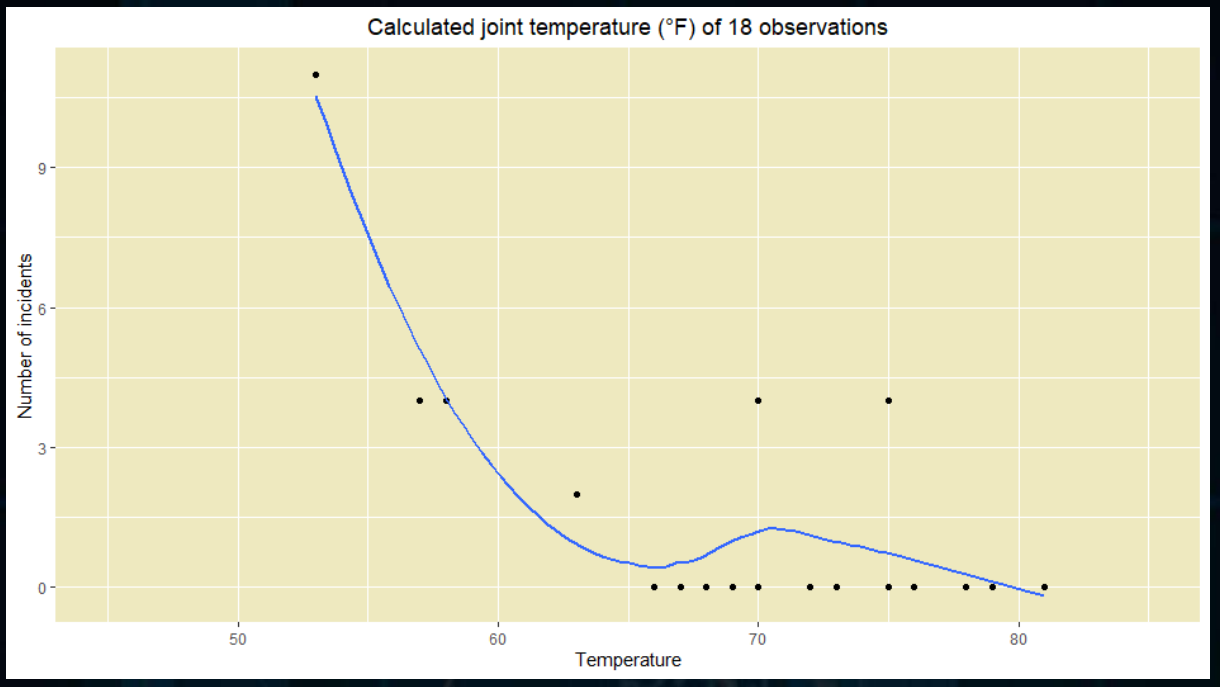
"It makes sense to predict likely future performance from past performance. Indeed, we often don't have much else to go on. Unfortunately, however, the past can be an uncertain guide to the future."

- David J. Hand [3.i]

It was a Tuesday morning, January 26, 1986. Daily Records writes the front-page news with the title "SHUTTLE EXPLODES: ALL 7 CREWMEMBERS ARE PRESUMED DEAD." The statistical story behind the infamous Challenger Space Shuttle disaster involves a statistical analysis based on only seven pre-observed points. The graph shows the relationship between air temperature at previous seven shuttle launches and whether the seals on the joints between segments of the rocket boosters are distressed.



The above graph showed no relationship beyond random variability. However, all the seven data points were points for all those previous launches which had problems. The graph did not include the points which did not have a problem. If included, the conclusion could be far different from what it appeared to be, as the latter case study reveals that higher air temperatures at launch are associated with fewer problems. The mistaken inference of no relationship led to the death of seven astronauts.



Let us consider another example. In the early times of COVID-19 outbreak, many people thought they were 'invincible' and took their health for granted. So, when they were advised to wear masks or to maintain social distance as a primary safety measure to prevent an infection they had never heard of, many of them took it with a pinch of salt and now we know the catastrophic consequences of not maintaining the safety protocols. Although various health organisations had stated their insights regarding the future based on the data they had at that time, people failed to comprehend the severity of a situation they were unfamiliar with.

Student ARTICLES

THE UNKNOWN:

Dark data is the data we do not have, or the data we may have but we are unaware of its presence, or the data we do not have access to. Isaac Sacolik describes it as 'data that is kept 'just in case' but hasn't (so far) found a proper usage'. The missing data in the Challenger plot and the unnoticed COVID data are the examples of dark data. Each dataset or even a single data point carries minimum information. So, dark data may contain valuable information. In statistics, our conclusions are purely based on the data in hand. We do not know what our conclusions would have been if we had the missing data. As a result, we completely miss the opportunity to see the world through dark data which implies we are at a risk of misunderstanding, of drawing incorrect conclusions about the way the world works.

DO WE HAVE ALL THE DATA?

We do not have data we never had until and unless we collect them. That is where dark data arises most often. A familiar example of occurrence of dark data is in the fields of survey sampling. 'Non response' in a survey is a major problem[3.ii]. Even when statisticians have several tools to tackle the non-response issue, a minimum effect is cast by the dark data.

Irrelevant data is also one kind of dark data. For example, suppose an online business firm keeps track of all the details regarding the purchase interest of a customer and makes offers based on it. If the data is not processed immediately, it may be irrelevant in future. According to IBM, about 60% of data loses its value immediately[4].
Even if we are correct in data selection and processing, well, dark data can still be there. No measure is completely accurate. So, we round up the number, which is in reality a greater or lesser value of the true value. For example, a thermometer will not record a value less than the freezing point of mercury. But as always, the results can still be severe. The failure of American Patriot Missile battery at Dharan in Saudi Arabia to intercept an Iraqi Scud missile which killed 28 and injured 100 was attributed to rounding to 24 bits of the unending binary expansion of 1/10 (of a second) when this error was aggregated over the 100 hours while the system was operating[5].

DIFFICULTIES IN ANALYSIS AND STORAGE:

It is difficult to read, categorise and analyse the unstructured nature of dark data by computers. Also, it requires a large number of resources to be analysed, which requires great expense that is often not financially affordable for companies. Cost is a big factor while considering the utilisation of data. According to Datamation, "the storage environments of EMEA organisations consists of 54% dark data" and to store such data by 2020, management costs would require $891 billion which could be avoided otherwise[6].

Also, due to the fact that dark data is the resource of unutilised valuable information, storage of dark data can put organisations at risk of data piracy or identity theft. Therefore, storing such data requires strong encryption and security. Hence, a better choice can be to discard these sensitive data in such a way that it is irretrievable.

THE FUTURE:

"Data and analytics will be the foundation of the modern industrial revolution"

> *Denis Paul, Teradata, Factories of The Future: The Value of Dark Data.*

The more the advancement in technology and advanced computing, the higher is the value of dark data. Such data can be used to draw insights that one never could have imagined; they can be used to bring maximum productivity to meet customer demands. Moreover, understanding the utilisation of data is important. Dark data teaches us how important it is to look into situations thoroughly so that we do not miss a single opportunity to make use of the information which is within our reach.

Student ARTICLES

References:

1. *The Challenger Disaster (https://bookdown.org/egarpor/PM-UC3M/glm-challenger.html)*
2. *Data for graph: R package: DAAG, data(orings)*
3. *Why What You Don't Know Matters - David J. Hand*
4. *Quote*
5. *Nothing Happened, So We Ignored It (page -21)*
6. *http://siliconangle.com/blog/2015/10/30/ibm-is-at-the-forefront-of-insight-economy-ibminsight/*
7. *https://www.iro.umontreal.ca/~mignotte/IFT2425/Disasters.html*
8. *Hernandez, Pedro (October 30, 2015). "Enterprises are Hoarding 'Dark' Data: Veritas", Datamation.*

Student ARTICLES

Data: Football's New Signing

- Shaun Chirag Lakra (3rd Year)

You may have seen your grandparents, parents or other elders chat or debate about some of the greatest players of their times like Diego Maradona, Pelé, Johan Cruyff, etc. While the game itself has not changed since then, whatever happened behind the scenes has transformed completely. Advancements in science and technology has found its way into Football and its implementations are extremely vital today.

BRENTFORD'S RECRUITMENT MODEL

Traditionally, football clubs would have their respective scouts observing young players at various clubs around the world and come to conclusions if they are worthy enough to get signed to their club. Football clubs also rely on their respective academies in order to shape and nurture young talent. While such a method of recruitment observing raw skill and talent is still relevant, data driven analysis of such players have also shown immense pertinence in today's world of football. One of the finest examples of those is the "Moneyball Approach" implemented by the newly promoted Premier League club, Brentford. Moneyball, as first defined by Billy Beane (former Baseball General Manager), means overlooking traditional wisdom around player scouting and using data to, crucially, find what the market undervalues.

Student ARTICLES

"For David to beat Goliath, he needed to use a different weapon. If David had used the same weapon, he would have lost the battle. You've got to find your weapons. That's what Brentford is about.", says Rasmus Ankersen, Brentford's co-director of Football.

Matthew Benham, a professional sports gambler, took over the control of shareholding of the club in 2012 and stated that he had a special plan of action and asked the fans to come on a great journey with him. His statement was backed by a statistics heavy approach that had already panned out at the Danish club FC Midtjylland, Brentford's sister club. Midtjylland, since 2014 have become three times Danish champions and also defeated football giants, Manchester United at the first leg of the last 32 of the UEFA Europa League in 2015.

For Brentford, the situation was quite unconventional. They did not have an appreciable academy functioning which led them into completely abandoning that project. The club also operates on a pretty tight budget, thus abolishing the academy saved them around £1.5 million each year. The owner, since his arrival at Brentford, hired mathematics graduates that analysed data behind players from around the world that were being unnoticed.

Student ARTICLES

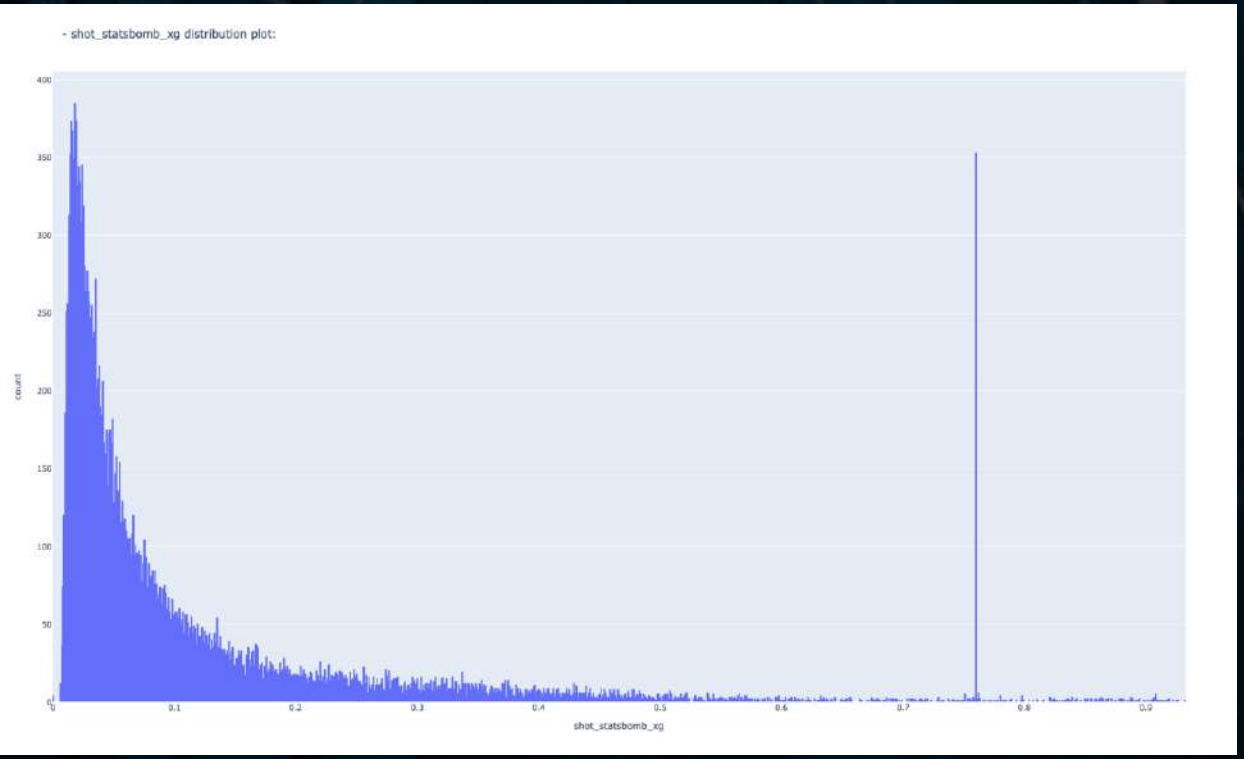
Let us take a look at an example of how Brentford recruits its players cost efficiently. Let's assume there are two players A and B who play as forwards, that is, their main objective is to find the back of the net with the ball. Both these players are alike in almost every way possible, for example, they are of the same age, play in the same division, have played the same number of matches (say, 20) and have taken the same number of attempted shots (say, 50). However, A has scored 15 goals and B has scored 5. Now, on paper it will seem that A is a much better forward and hence has a higher money value than B. The concept of expected goals (denoted by "xG") now comes into account that may contradict our assumption. Let's say that A has scored 15 goals from an xG of 10 goals and B has scored his 5 goals from an xG of 12. This information itself can be interpreted in a plethora of ways. Firstly, it can be said that A may just have gotten lucky or faced mediocre goalkeepers and there is no guarantee that A would continue to perform at that level in the future. On the other hand, B is likely to score more goals if he performs at his expected level. Thus, in "Moneyball" terms, buying B would be a recommended recourse as he would be a relative bargain and might also end up being a better player than A. Following an approach close to this, Brentford were able to acquire talents such as Neal Maupay, Saïd Benrahma and Ollie Watkins on whom they made profits of £20.2 million, £21.4 million and £26.8 million respectively by selling them to big Premier League clubs.

Student ARTICLES

A BRIEF STATISTICAL EVALUATION OF A FOOTBALLER'S SKILL

Before we dive into the evaluation of a player's skill, let us first define, "skill". In a very mainstream sense, skill can be defined as, 'the ability to perform a certain task'. However, in this particular discussion we shall define it as the 'likelihood to complete a certain task given a specific difficulty', thus making it a probability distribution.

As mentioned earlier, we shall use the concept of xG (expected goals) to measure the skill distribution of shooting. In this probability model, we look at the expected outcome of each shot rather than its actual outcome. We observe the distribution plot (figure 1) of the count of scoring from a shooting event (higher the probability of scoring, lower the shot difficulty will be). (Reference no. 5)



Student ARTICLES

Figure 1:

In figure 1, Y axis denotes the count of the shots corresponding to the expected goals value (xG) in the X axis. We can observe that the distribution is highly positively skewed except for the xG from penalties (between 0.75 and 0.8).

"Lift" is a measure which compares a metric with respect to a baseline. Thus, it's a ratio that indicates how much better the calculated value is over the baseline. We shall compute the xG lift AUC (Area under the Curve) for different footballers (Figure 2). We take baseline or expected AUC as 0.5.

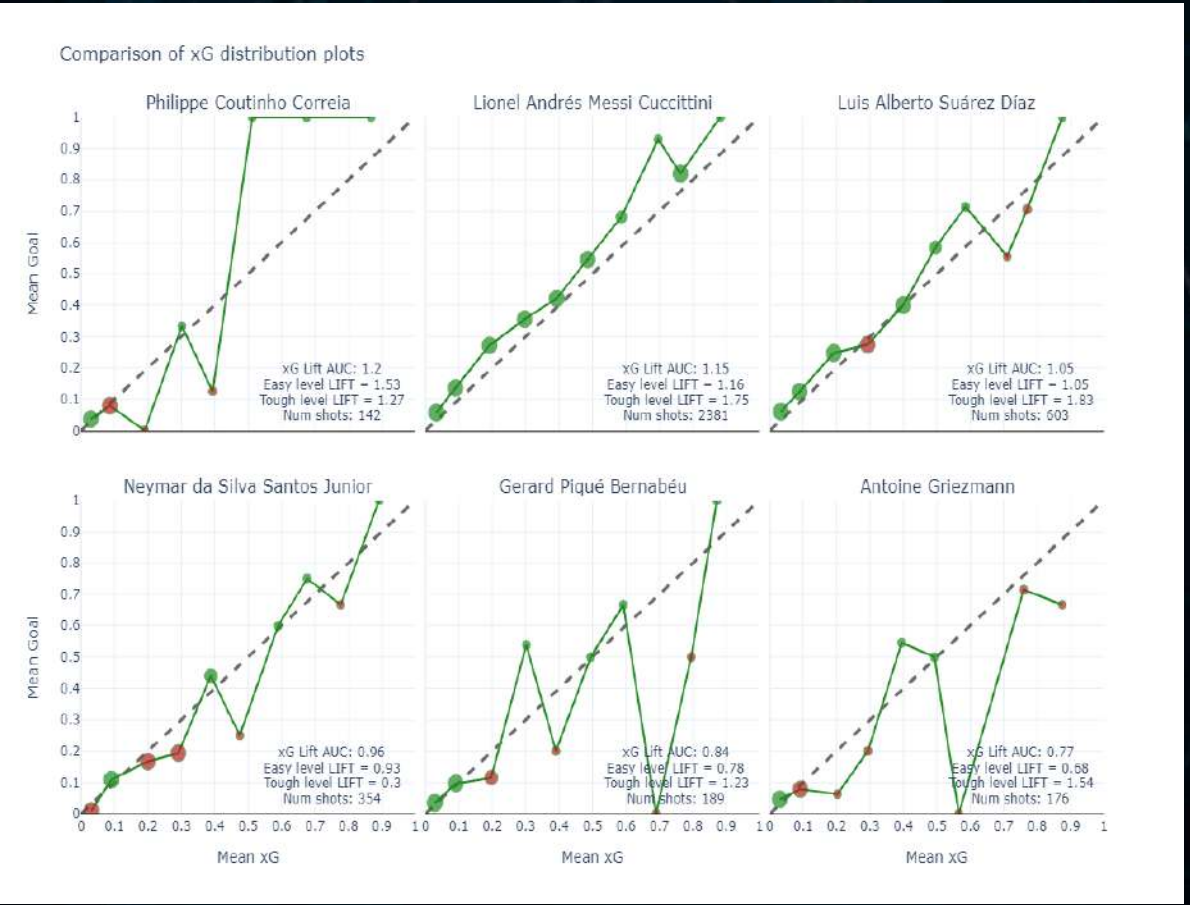


Figure 2

In figure 2 (reference no. 5), the dotted line indicates the expected performance from each player. From row 1, column 2 (distribution plot of Lionel Messi) we can see that he beats the standard for any given difficulty.

Let us try and visualise the “skills radar chart” (figure 3) of arguably one of the greatest footballers of all time, Lionel Andrés Messi.

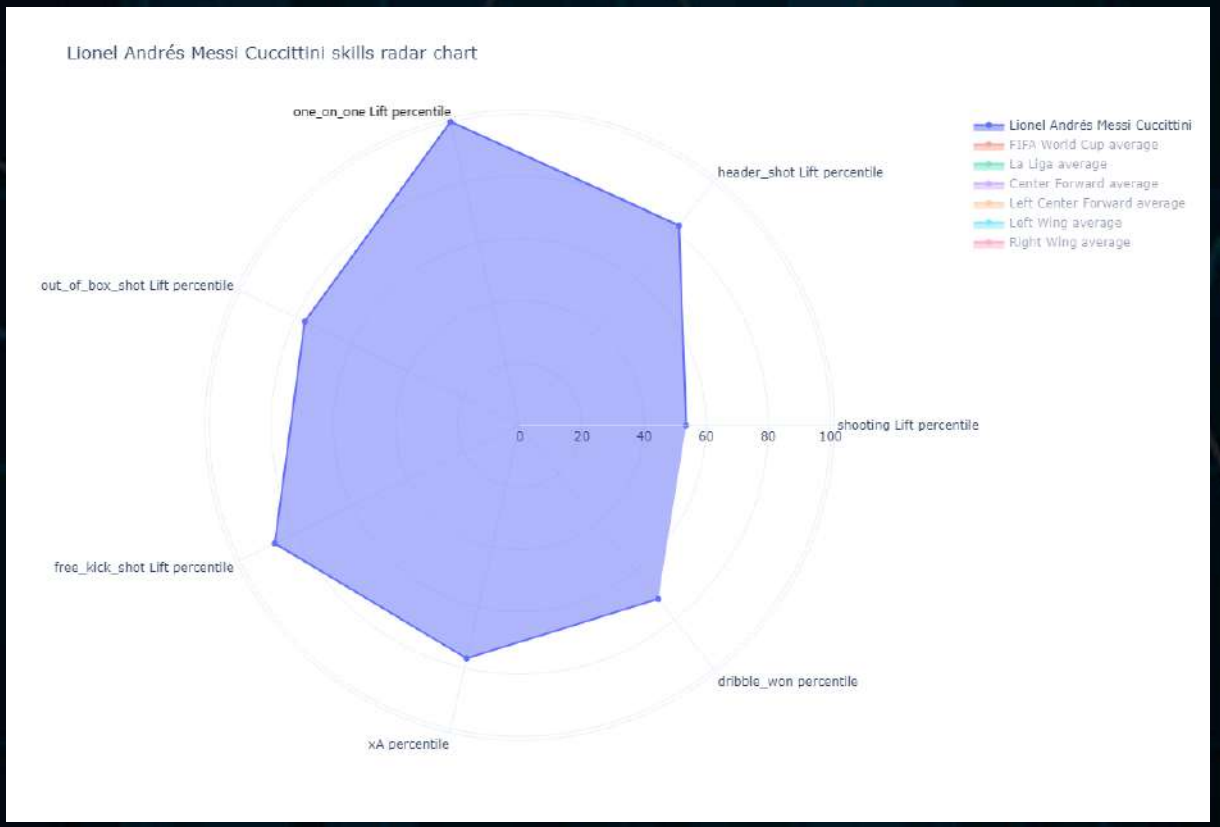


Figure 3

Greater area of the polygon indicates higher performance in a certain attribute. We can observe that his one-on-one finishing is the best (100th percentile).

In order to get a better idea to how he compares to the rest of the world, we compare his radar chart with that of an average centre-forward player. (Figure 4)

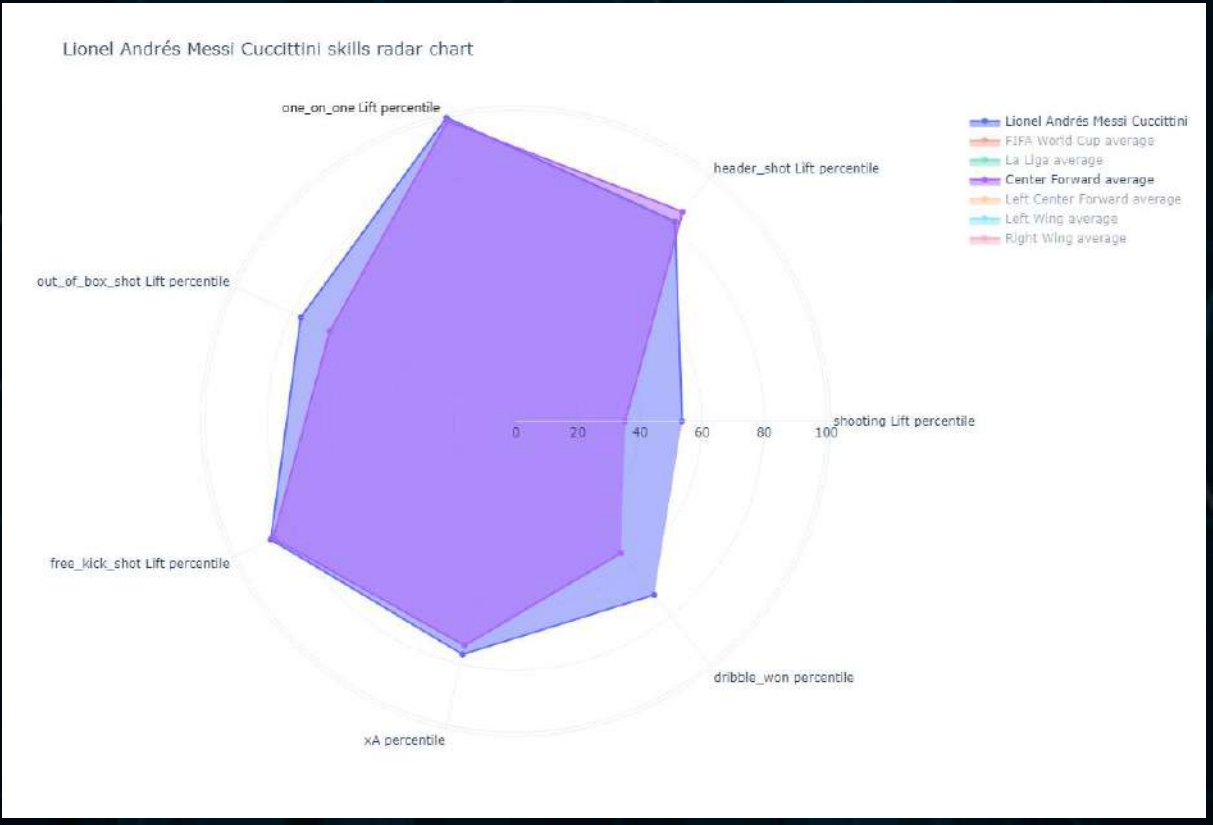


Figure 4

We observe that Lionel Messi is better than the average centre-forward player in almost every field except that of header shots, which is expected as he is not as tall as the average centre-forward.

REFERENCES:

1. *https://www.youtube.com/watch?v=mrwwVoD1RQQ*
2. *https://towardsdatascience.com/data-driven-evaluation-of-football-players-skills-c1df36d61a4e*
3. *https://bleacherreport.com/articles/2718752-brentfords-moneyball-way-to-beat-football-teams-with-huge-budgets*
4. *https://www.bloomberg.com/news/articles/2021-05-28/big-data-model-takes-local-london-team-to-soccer-s-richest-game*
5. *All datasets and graphs obtained from https://statsbomb.com*

Student ARTICLES

How important is “home performance” and “winning the toss” in order to win the IPL?

- Abhiroop Basu (3rd Year), Shameek Bhowmick (3rd Year)

Introduction:

The grandest franchise cricket tournament in the world is the Indian Premier League. The quality of cricket here is at the top level. And for the fans, witnessing their favourite team win the IPL is the ultimate dream. Some fans have tasted the feeling of triumph multiple times, whereas for others it’s still a dream. Here we will look into the performance of the champion teams from each season and try to analyze whether showing consistency at home games and/or winning the toss has played a crucial role in their triumph or at least helped them reaching the playoffs.

The reason behind choosing these two factors are:

I. There are 14 league stage games for a team in an 8-team IPL tournament, 7 home and 7 Away. A team gets 2 points if they win a game. 16 points in total can guarantee a team a place in the playoffs, even 14 can be enough. Now in a home game in the sport of cricket, the home team always gets an advantage over their opponents as they get to choose the pitch type, have a better knowledge of the pitch conditions, ground dimensions (length of the boundaries), and get the backing of home crowd. So, the odds in favour of a home team winning the game is usually higher and thus, inconsistent performance at home can be fatal for a team in their chase for a playoff spot. Here we will see whether even the champions have been dependent on their home performance in order to win the IPL, or they have been equally good in both home and away.

II. We will also look into the toss factor and see whether winning the toss has played an important role even for the champion teams or is it only a myth and a frequently used excuse for the underperforming teams (Being really harsh here; tosses do matter in some matches depending on the conditions).

III. There maybe a few more factors, but none of them is as considerable as are the two mentioned above. The strength of the teams is also not taken into consideration as every team gets nearly an equal opportunity to make a squad as strong as any, in the auctions.

Student ARTICLES

Data Collection:

In order to study the home performance properly, data from seasons which were partially or completely played outside India were not included. That means the seasons played in 2009, 2014, 2020 and 2021 were not taken into consideration. For the rest of the seasons, the data on the result (Won or Lost) of every league stage match played by the champions of respective seasons, its venue (Home or Away) and the toss outcome of the match (Won or Lost) are collected.

Explanation: KKR won the IPL in 2012. So, from the 2012 season, only the league stage matches played by KKR is considered. Like this, from a particular season, only the matches played by the team which went on to win the IPL in that season are considered.

The collection of the data is done manually from www.wikipedia.com . The overview of the data is presented in the next table:

Season	Champion Team	Home Matches Won / Home Matches Played	Away Matches Won / Away Matches Played	Tosses Won/ Matches Played
2008	Rajasthan Royals	07-Jul	04-Jul	Oct-14
2010	Chennai Super Kings	04-Jul	03-Jul	Aug-14
2011	Chennai Super Kings	07-Jul	02-Jul	Jul-14
2012*	Kolkata Knight Riders	03-Jul	07-Aug	Jun-15
2013#	Mumbai Indians	08-Aug	03-Aug	Nov-15
2015	Mumbai Indians	04-Jul	04-Jul	Jun-14
2016	Sunrisers Hyderabad	04-Jul	04-Jul	Aug-14
2017	Mumbai Indians	05-Jul	05-Jul	Aug-14
2018	Chennai Super Kings	06-Jul	03-Jul	Oct-14
2019	Mumbai Indians	05-Jul	04-Jul	Aug-14
Total	IPL Champion Teams	53/71	39/72	82/143

Table 1.1: Overview of the collected data

*in the 2012 season, one of KKR's home game were abandoned due to rain. And this is the only season where we see that a champion team has a far better performance record in away games than home.

#the season 2013 had 10 teams in IPL, and a different format which resulted 8 home games and 8 away games for every team.

Analysis:

1. Analyzing the significance of the explanatory variables using GLM:

In our dataset, we have the data on 143 matches involving the teams which went on to win the IPL in respective seasons. Let us define:

A. Response Variable:

Y: Let Y be a random variable denoting the result of an IPL match played by the eventual champions of that season. (Y=1 denotes the champion team won the match, whereas Y=0 denotes they lost the match). In our study, this is the response variable.

Let, Y_i denote the result of the i^{th} IPL match from the dataset we considered. (Y_i=1 denotes the champion team won match, whereas Y_i=0 denotes they lost the match). [i=1(1)143]

B. Explanatory Variables or Covariates:

In our study we have two explanatory variables or covariates, x_1 and x_2.

x_{1i}: Let x_{1i} denote the "venue" of the i^{th} IPL match from the dataset we considered. (x_{1i}=1 denotes the match was a home game for the champion team involved in the match, whereas x_{1i}=0 denotes it was an away game.)

x_{2i}: Let x_{2i} denote the "toss outcome" of the i^{th} IPL match from the dataset we considered. (x_{2i}=1 denotes the champion team involved in the match won the toss, whereas x_{2i}=0 denotes they lost the toss.) [i=1(1)143]

Model: Logistic Model

In the logistic model, we take:

$$P(Y=1|x_1, x_2) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1+e^{\beta_0 + \beta_1 X_{1_+} \beta_2 X_2}}$$

Where *β_0, β_1, and β_2 are the unknown parameters which are to be estimated.*

- *Fitting the model*: By Maximum Likelihood Method
- *Goodness of fit measure*: Deviance
- *Software used*: R
- Output:

Parameters	Estimated value of the parameters	Standard Error	t = Estimated ValueStandard Error
β_0	0.21014	0.30201	0.696
β_1	0.92323	0.36411	2.536
β_2	-0.08371	0.36407	-0.23

Table 1.2: Output Table

Note:

Deviance measure = 179.66

Cut point = 1.96 (if |t|<1.96 for a parameter, we will mark it as insignificant for the model.)

Interpretation:

We are not really interested about the intercept parameter *β_0; rather, β_1 and β_2 will respectively state whether the covariates x_1 and x_2 are significant or not.*

- β_1:

1. *From the output table, we found that for β_1,* |t|>1.96 . So, firstly we can conclude that x_1, which denotes the "venue", is a significant covariate for this model. That is, whether a match is being played at home or away is an important factor even for the champion teams in order to win a match in IPL.

2. The estimated value *β_1* came out to be 0.92323 . So, we can also interpret that for the champion teams, the odds in favour of winning a match in IPL increases $e^{0.92323}$ times, i.e., 2.52 times, when a match is played at home. (keeping the other covariate fixed)

[Explanation:

Let, $Odds_1$ = The odds of winning an away match

$$= \frac{Pr(Y=1/X_1=0,X_2)}{1-Pr(Y=1/X_1=0,X_2)} \quad = e^{\beta_0 + \beta_1 * 0 + \beta_2 X_2} \quad = e^{\beta_0 + \beta_2 X_2}$$

Let, $Odds_2$ = The odds of winning a home match

$$= \frac{Pr(Y=1/X_1=1,X_2)}{1-Pr(Y=1/X_1=1,X_2)} \quad = e^{\beta_0 + \beta_1 * 1 + \beta_2 X_2} \quad = e^{\beta_0 + \beta_1 + \beta_2 X_2}$$

$$Odds\ Ratio = \frac{Odds_2}{Odds_1} \quad = \frac{e^{\beta_0 + \beta_1 + \beta_2 X_2}}{e^{\beta_0 + \beta_2 X_2}} \quad = e^{\beta_1}$$

Or, $Odds_2 = e^{\beta_1} * Odds_1$

So, odds in favour of winning a home match is e^{β_1} times than that of winning an away match.]

- β_2:

1. *From the output table, we found out that for β_2, $|t|<1.96$ so x_2 , which denotes the "toss outcome", is not a significant covariate for this model. That is,* for champion teams, whether the toss of a match is won or lost has not been important in order to win that particular match.
2. As x_2 comes out to be insignificant, we will not go into further analysis of its effects on the odds of winning a match.

Conclusions from the GLM analysis:

From the above analysis and calculations, we can conclude that taking the past seasons of IPL into account, seemingly:

1. The home games have been easier to win than the away games even for the champion teams. The odds of winning a home game is 2.52 times higher than the odds of winning an away game. So, losing home games or inconsistent performance at home is more fatal to a team's chances of qualifying for the playoffs or going all the way to win the IPL trophy, than losing the away games. (As one can make amends for their bad away performance by performing consistently at home but doing the exact opposite thing would be tough.)

2. Toss has not been a crucial factor for the champion teams in order to win a game in IPL. For some matches, toss is quite crucial depending on the weather conditions or pitch type. But in totality, toss outcomes cannot be blamed for a team's inconsistent performance, as for champion teams, winning or losing the toss has mattered little.

2. Conditional and Marginal Association Analysis using 3-Way Contingency Table

We have already come to a conclusion that toss is not a significant factor for the champion teams to win a game in IPL. Let us now consider "toss outcome" as a latent effect. By analyzing marginal and conditional associations from a 3-way table, we can check how the latent factor, i.e., the "toss outcome", is influencing the association between the response variable (result of the match) and the explanatory variable (venue of the match). Ultimately our goal is to see whether the outcomes from the analysis of a 3-way contingency table support the outcomes we have already got from the GLM analysis or are they contradicting.

Let us denote,

Y: Match Result (Y=1: Won, Y=0: Lost)

X: Match Location (X=1: Home, X=0: Away)

Z: Latent Effect: Toss Result (Z=1: Won, Z=0: Lost)

The corresponding 3-Way Contingency Table with the given data is given by,

		Match Result(Y)	
Toss Result(Z)	Match Location(X)	Won(Y=1)	Lost(Y=0)
Won(Z=1)	Home(X=1)	36	9
	Away(X=0)	17	20
Lost(Z=0)	Home(X=1)	17	9
	Away(X=0)	22	13
Total	Home(X=1)	53	18
	Away(X=0)	39	33

Table 2: 3-Way Contingency Table of the data

Using the above 3-Way Contingency Table we obtain the Partial and Marginal Tables for further analysis.

2.1 Conditional Association using Partial Tables

The Partial Table controlling the effect Toss Result Won, i.e., Z=1, is given by,

Match Location(X)	Match Result(Y)		Total
	Won(Y=1)	Lost(Y=0)	
Home(X=1)	36	9	45
Away(X=0)	17	20	37
Total	53	29	82

Table 2.1: Partial Table controlling the latent effect "Toss Result: Won", i.e., Z=1

From Table 2.1,

The odds of winning a match given that the match is played at home, when the toss is won is given by,

$$O_1 = \frac{P(Y = 1|X = 1, Z = 1)}{\mathrm{P(Y = 0|X = 1, Z = 1)}} = \frac{36}{9} = 4$$

The odds of winning a match given that the match is played away, when the toss is won is given by,

$$O_2 = \frac{P(Y = 1|X = 0, Z = 1)}{\mathrm{P(Y = 0|X = 0, Z = 1)}} = \frac{17}{20} = 0.85$$

Hence, the odds ratio is given as,

$$OR_1 = \frac{O_1}{O_2} = 4.7 \approx 5$$

Hence, the odds of winning a match given that the match is played at home is 5 times the odds of winning the match given that the match is played away, when the toss is won.

Also,

The probability of winning a match given that the match is played at home, when the toss is won is given by,

$$P_1 = P(Y = 1|X = 1, Z = 1) = \frac{36}{45} = 0.8$$

The probability of winning a match given that the match is played away, when the toss is won is given by,

$$P_2 = P(Y = 1|X = 0, Z = 1) = \frac{17}{37} = 0.4595$$

Here we calculate,

$$[100 * (P_1 - P_2)]\% = 34.05\%$$

Hence, when the team won the toss, the team won the match 34.05% times more often when playing at home than playing away.

The Partial Table controlling the effect Toss Result Lost, i.e., Z=0, is given by,

Match Location(X)	Match Result(Y)		Total
	Won(Y=1)	Lost(Y=0)	
Home(X=1)	17	9	26
Away(X=0)	22	13	35
Total	39	22	61

Table 2.2: Partial Table controlling the effect of "Toss Result: Lost", i.e., Z=0

From Table 2.2,

The odds of winning a match given that the match is played at home, when the toss is lost is given by,

$$O_3 = \frac{P(Y = 1|X = 1, Z = 0)}{\mathrm{P(Y = 0|X = 1, Z = 0)}} = \frac{17}{9} = 1.89$$

The odds of winning a match given that the match is played away, when the toss is lost is given by,

$$O_4 = \frac{P(Y = 1|X = 0, Z = 0)}{\mathrm{P(Y = 0|X = 0, Z = 0)}} = \frac{22}{13} = 1.69$$

Hence, the odds ratio is given as,

$$OR_2 = \frac{O_3}{O_4} = 1.11$$

Hence, the odds of winning a match given that the match is played at home is 1.11 times the odds of winning the match given that the match is played away, when the toss is lost.

In sum, when the team lost the toss, the odds of winning a home game for that team is at par with the odds of winning an away game.

Also,

The probability of winning a match given that the match is played at home, when the toss is lost is given by,

$$P_3 = P(Y = 1|X = 1, Z = 0) = \frac{17}{26} = 0.6538$$

The probability of winning a match given that the match is played away, when the toss is lost is given by,

$$P_4 = P(Y = 1|X = 0, Z = 0) = \frac{22}{35} = 0.6286$$

Here we calculate,

$$[100 * (P_3 - P_4)]\% = 2.52\%$$

From the above two calculations,

By controlling the Toss Result, the percentage of matches WON was higher when the matches were played at home than when played away.

2.2 Marginal Association using Marginal Table

The Marginal Table ignoring the effect Toss Result is given by,

Match Location(X)	Match Result(Y)		Total
	Won(Y=1)	Lost(Y=0)	
Home(X=1)	53	18	71
Away(X=0)	39	33	72
Total	92	51	143

Table 2.3: Marginal Table ignoring the effect of Toss Result (Z)

From Table 2.3,

The odds of winning a match given that the match is played at home is given by,

$$O_1^* = \frac{P(Y=1|X=1)}{P(Y=0|X=1)} = \frac{53}{18} = 2.94$$

The odds of winning a match given that the match is played away is given by,

$$O_2^* = \frac{P(Y=1|X=0)}{P(Y=0|X=0)} = \frac{39}{33} = 1.18$$

Hence, the odds ratio is given as,

$$OR^* = \frac{O_1^*}{O_2^*} = 2.49 \approx 2.5$$

Hence, the odds of winning a match given that the match is played at home is 2.5 times the odds of winning the match given that the match is played away. (This was already calculated while interpreting the covariates in GLM analysis)

Also,

The probability of winning a match given that the match is played at home is given by,

$$P_1^* = P(Y = 1|X = 1) = \frac{53}{71} = 0.7465$$

The probability of winning a match given that the match is played away is given by,

$$P_2^* = P(Y = 1|X = 1) = \frac{39}{72} = 0.5417$$

Here we calculate,

$$[100 * (P_1^* - P_2^*)]\% = 20.48\%$$

Hence, the team won the match 20.48% times more often when playing at home than away.

From the above calculations,

By ignoring the Toss Result, the percentage of matches WON was higher when the matches were played at home than when played away.

Conclusions from Conditional and Marginal Association Analysis:

From both the Conditional and Marginal Associations,

1. We can say that the results by controlling the effect of Toss Outcome are on par with the results where we ignore the Toss Outcome. In both the cases, the odds in favour of winning a home match is higher. And this very much supports the outcomes from GLM analysis, where we found that the venue of a match (Home or Away) is a significant factor for the outcome of a match involving the champion team.
2. Although we find that "toss outcome" is an insignificant factor from the GLM analysis, we can clearly see that it has a telling effect as a latent factor. When the toss is won, the odds of winning a home game is 5 times higher than that of winning an away game, whereas when the toss is lost, the odds of winning a home game is nearly equal to the odds of winning an away game.

Ultimate Conclusion:

In order to become the champion, a team has to perform like a champion. Here we analyzed the performance of the past champions in order to get a clear picture of what it takes to win the IPL. From the analysis we can conclude that even the champion teams have been dependent on their home performance in order to win the IPL, away games have been quite tough even for them. So,the fact of making full use of the advantages one gets in the home games and performing consistently well at home is a major key to a successful IPL season.

We also found out that the toss has not been a major factor for the champion teams in order to win a game in IPL, but it has certainly helped as a latent factor. Here a few things should be taken into consideration. During the T20 world cup played in UAE and Oman in October-November 2021, if we consider the matches played between the top 8 test playing nations, we will see that in 13 out of the 16 games, the team who won the toss chose to bowl first and went on to win the game. Although the extreme weather and pitch conditions were the main culprits behind making the tournament solely dependent on the flip of a coin, the toss has generally been a much more significant factor in short tournaments, than in the IPL. Short tournaments contain way lesser number of group games which sets the margin for error very low, so a favourable toss outcome comes handy. Whereas in IPL, there are 14 group games (7 at home) which balances out the luck factor and ultimately a team tastes success based on its performance only. So, the conclusion stated here about toss outcome is strictly for the IPL, and not in general.

References:

1. Goon A.M., Gupta M.K., Dasgupta, B. (2005), Fundamentals of Statistics, Vol II, World Press, Calcutta.

2. Agresti, A. (2007), An Introduction to Categorical data analysis. Wiley.

Student ARTICLES

Selection of Significant Predictors in Statistical Analysis

- Hrithik Sen (3rd Year)

INTRODUCTION:-

In machine learning and statistics, variable subset selection is the process of selecting a subset of relevant predictors for use in model construction. Selection procedures are used for several purposes:

- Simplify any model to make it easier for the user.
- To reduce time for necessary analysis and computations.
- To avoid the problem of dimensionality.
- To make a huge dataset compatible for future analysis.

The main notion while using a subset selection procedure is that any data contains some features that are redundant in connection to the response variable and thus can be removed without producing much loss of information.

This process involves choosing a subset of all the given predictors that are related to the response variable. We fit a model for the response variable on basis of those chosen set of predictors using the method of least squares. *Then, we want to select the best single model among all those fitted models.*

This variable subset selection process includes 1) “best subset selection”, 2) “forward stepwise selection” and 3) “backward stepwise selection” methods.

Overall Comparison in the Application of Best Subset Selection, Forward and Backward Stepwise Selection Method :-

In the field of various statistical analysis, one cannot confidently prefer any particular method over the others. If the number of predictors is large, then we have to opt for the stepwise selection over the best subset selection for the ease of computation. But if the number of predictors is not very large, then we can apply best subset selection on the dataset in order to obtain the most significant subset of predictors for easier interpretation and further computational analysis. These methods will yield more or less similar results for obtaining significant predictors corresponding to the response variable.

Working Example Using Stepwise and Best Subset Selection on the Same Dataset :-

Let us consider the dataset “Acute Inflammations” from the website “UCI MACHINE LEARNING REPOSITORY”. The dataset was created by a medical expert as a dataset to test the expert system, which will perform the presumptive diagnosis of a disease of the urinary system.

It will be the example of diagnosing of the acute inflammations of urinary bladder. Acute Inflammations of urinary bladder is characterised by the sudden occurrence of pains in the abdomen region and the urination in form of constant pushing, micturition pains and sometimes the lack of urine keeping.

Now, for all the three selection procedures, we will use the same independent and dependent variables.

Dependent Variable : Inflammation of urinary bladder {yes = 1, No = 0}

Independent Variable :-

1. Temperature of patient
2. Occurrence of Nausea {yes = 1, No = 0}
3. Lumbar Pain {yes = 1, No = 0}
4. Micturition Pains {yes = 1, No=0}

The response variable "Inflammation" is labelled as y in our data. It is binary in nature. So, our data seems to fulfil the criteria of binary logistic regression. The independent variables "temp", "nausea", "lumbar pain" and "micturition pain" are labelled as x_1, x_2, x_3 and x_4 respectively in our data.

Hence, we will fit binary logistic model on y based on the predictors x_1, x_2, x_3 and x_4.

Calculations:-

1. Best Subset Selection Method:-

Stage- 1:-

Logistic Regression Containing only 1 predictor:-

Predictors	Fitted Regression Equation	Deviance
Temp (x1)	$\eta = 10.381 - 0.269*x1$	159.5006
Nausea (x2)	$\eta = -0.218 + 0.806*x2$	162.9483
Lumbar Pain (x3)	$\eta = 1.386 - 2.374*x3$	131.8947
Micturition Pain (x4)	$\eta = -1.159 + 2.618*x4$	124.9958

OUTPUT:- Here, the value of the deviance for the model fitted on the predictor "Micturition Pain" (x_4) is minimum. So, we can take the model fitted on "Micturition Pain" (x_4) as M_1.

Logistic Regression Containing only 2 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x1), Nausea (x2)	$\eta = 27.393 - 0.728*x1 + 2.764*x2$	140.11993
Temp (x1), Lumbar Pain (x3)	$\eta = 3.514 - 0.056*x1 - 2.297*x3$	131.69606
Temp (x1), Micturition Pain (x4)	$\eta = 12.295 - 0.349*x1 - 2.765*x4$	117.55299
Nausea (x2), Lumbar Pain (x3)	$\eta = 1.386 + 4.301*x2 - 5.099*x3$	95.98995
Nausea (x2), Micturition Pain (x4)	$\eta = -1.124 - 0.443*x2 + 2.781*x4$	124.43318
Lumbar Pain (x3), Micturition Pain (x4)	$\eta = 0.282 - 2.502*x3 + 2.736*x4$	99.16968

OUTPUT:- Here, the value of the deviance for the model fitted on the predictors "Nausea" (x_2) and "Lumbar Pain" (x_3) is minimum. So, we can take the model fitted on "Nausea" (x_2) and "Lumbar Pain" (x_3) as M_2.

Logistic Regression Containing only 3 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Nausea (x_2), Lumbar Pain(x_3)	$\eta = 53.357 - 1.349*x_1 + 9.038*x_2 - 6.946*x_3$	66.13639
Temp (x_1), Nausea (x_2), Micturition Pain (x_4)	$\eta = 18.462 - 0.513*x_1 + 1.114*x_2 + 2.399*x_4$	115.74781
Temp (x_1), Lumbar Pain (x_3), Micturition Pain (x_4)	$\eta = 9.324 - 0.234*x_1 - 2.399*x_3 + 2.991*x_4$	96.96275
Nausea (x_2), Lumbar Pain (x_3), Micturition Pain (x_4)	$\eta = 0.619 + 3.175*x_2 - 4.414*x_3 + 1.578*x_4$	88.45238

OUTPUT:- Here, the value of the deviance for the model fitted on the predictors "Temp" (x_1), "Nausea" (x_2) and "Lumbar Pain" (x_3) is minimum. So, we can take the model fitted on "Temp" (x_1), "Nausea" (x_2) and "Lumbar Pain" (x_3) as M_3.

Logistic Regression Containing all 4 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Nausea (x_2), Lumbar Pain (x_3), Micturition Pain (x_4)	*$\eta = 55.458 - 1.042*x_1 + 9.434*x_2 - 7.161*x_3 - 0.231*x_4$*	*66.12639*

OUTPUT:- We can take the model fitted on "Temp" (x_1), "Nausea" (x_2),"Lumbar Pain" (x_3) and "Micturition Pain" (x_4) as M_4.

Stage- 2:-

Model	Predictors	AIC
M_1	Micturition Pain (x_4)	128.99579
M_2	Nausea (x_2), Lumbar Pain (x_3)	101.98995
M_3	*Temp (x_1), Nausea (x_2), Lumbar Pain(x_3)*	*74.13639*
M_4	Temp (x_1), Nausea (x_2), Lumbar Pain(x_3), Micturition Pain (x_4)	76.06313

Conclusion:-

Hence, by the help of "Best Subset Selection Method", we can choose the logistic regression model fitted on the predictors "Temperature of patient", "Occurrence of Nausea" and "Lumbar pain" as the best single model among all the models fitted on these four predictors.

2. Forward Stepwise Selection Method:-

Stage- 1:-

Logistic Regression Containing only 1 predictor:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1)	$\eta = 10.381 - 0.269*x_1$	159.5006
Nausea (x_2)	$\eta = -0.218 + 0.806*x_2$	162.9483
Lumbar Pain (x_3)	$\eta = 1.386 - 2.374*x_3$	131.8947
Micturition Pain (x_4)	*$\eta = -1.159 + 2.618*x_4$*	*124.9958*

OUTPUT:- Here, the value of the deviance for the model fitted on the predictor "Micturition Pain" (x_4) is minimum. So, we can take the model fitted on "Micturition Pain" (x_4) as M_1.

Logistic Regression Containing only 2 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Micturition Pain (x_4)	$\eta = 12.295 - 0.349*x_1 - 2.765*x_4$	117.55299
Nausea (x_2), Micturition Pain (x_4)	$\eta = -1.124 - 0.443*x_2 + 2.781*x_4$	124.43318
Lumbar Pain (x_3), Micturition Pain (x_4)	*$\eta = 0.282 - 2.502*x_3 + 2.736*x_4$*	*99.16968*

OUTPUT:- Here, the value of the deviance for the model fitted on the predictors "Lumbar Pain" (x_3) and "Micturition Pain" (x_4) is minimum. So, we can take the model fitted on "Lumbar Pain" (x_3) and "Micturition Pain" (x_4) as M_2.

Logistic Regression Containing only 3 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Lumbar Pain (x_3), Micturition Pain (x_4)	$\eta = 9.324 - 0.234*x_1 - 2.399*x_3 + 2.991*x_4$	96.96275
Nausea (x_2), Lumbar Pain (x_3), Micturition Pain (x_4)	*$\eta = 0.619 + 3.175*x_2 - 4.414*x_3 + 1.578*x_4$*	*88.45238*

OUTPUT:- Here, the value of the deviance for the model fitted on the predictors "Nausea" (x_2), "Lumbar Pain" (x_3) and "Micturition Pain" (x_4) is minimum. So, we can take the model fitted on "Nausea" (x_2), "Lumbar Pain" (x_3) and "Micturition Pain" (x_4) as M_3.

Logistic Regression Containing all 4 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Nausea (x_2), Lumbar Pain (x_3), Micturition Pain (x_4)	*$\eta = 55.458 - 1.042*x_1 + 9.434*x_2 - 7.161*x_3 - 0.231*x_4$*	*66.13639*

OUTPUT:- We can take the model fitted on "Temp" (x_1), "Nausea" (x_2), "Lumbar Pain" (x_3) and "Micturition Pain" (x_4) as M_4.

Stage- 2:-

Model	Predictors	AIC
M_1	Micturition Pain (x_4)	128.99579
M_2	Lumbar Pain (x_3), Micturition Pain (x_4)	105.16968
M_3	Nausea (x_2), Lumbar Pain (x_3), Micturition Pain (x_4)	96.45238
M_4	*Temp (x_1), Nausea (x_2), Lumbar Pain(x_3), Micturition Pain (x_4)*	*76.06313*

Conclusion:-

Hence, by the help of "Forward Stepwise Selection Method", we can choose the logistic regression model fitted on the predictors "Temperature of patient", "Occurrence of Nausea", "Lumbar pain" and "Micturition Pain" as the best single model among all the models fitted on these four predictors.

3. Backward Stepwise Selection Method:-

Stage- 1:-

Logistic Regression Containing all 4 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Nausea (x_2), Lumbar Pain (x_3), Micturition Pain (x_4)	*$\eta = 55.458 - 1.042*x_1 + 9.434*x_2 - 7.161*x_3 - 0.231*x_4$*	*66.13639*

OUTPUT:- We can take the model fitted on "Temp" (x_1), "Nausea" (x_2),"Lumbar Pain" (x_3) and "Micturition Pain" (x_4) as M_4.

Logistic Regression Containing only 3 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Nausea (x_2), Lumbar Pain (x_3)	*$\eta = 53.357 - 1.349*x_1 + 9.038*x_2 - 6.946*x_3$*	*66.13639*
Temp (x_1), Nausea (x_2), Micturition Pain (x_4)	$\eta = 18.462 - 0.513*x_1 + 1.114*x_2 + 2.399*x_4$	115.74781
Temp (x_1), Lumbar Pain (x_3), Micturition Pain (x_4)	$\eta = 9.324 - 0.234*x_1 - 2.399*x_3 + 2.991*x_4$	96.96275
Nausea (x_2), Lumbar Pain (x_3), Micturition Pain (x_4)	$\eta = 0.619 + 3.175*x_2 - 4.414*x_3 + 1.578*x_4$	88.45238

OUTPUT:- Here, the value of the deviance for the model fitted on the predictors "Temp" (x_1), "Nausea" (x_2) and "Lumbar Pain" (x_3) is minimum. So, we can take the model fitted on "Temp" (x_1), "Nausea" (x_2) and "Lumbar Pain" (x_3) as M_3.

Logistic Regression Containing only 2 predictors:-

Predictors	Fitted Regression Equation	Deviance
Temp (x_1), Nausea (x_2)	$\eta = 27.393 - 0.728*x_1 + 2.764*x_2$	140.11993
Temp (x_1), Lumbar Pain (x_3)	$\eta = 3.514 - 0.056*x_1 - 2.297*x_3$	131.69606
Nausea (x_2), Lumbar Pain (x_3)	*$\eta = 1.386 + 4.301*x_2 - 5.099*x_3$*	*95.98995*

OUTPUT:- Here, the value of the deviance for the model fitted on the predictors "Nausea" (x_2) and "Lumbar Pain" (x_3) is minimum. So, we can take the model fitted on "Nausea" (x_2) and "Lumbar Pain" (x_3) as M_2.

Logistic Regression Containing only 1 predictor:-

Predictors	Fitted Regression Equation	Deviance
Nausea (x_2)	$\eta = -0.218 + 0.806*x_2$	162.9483
Lumbar Pain (x_3)	*$\eta = 1.386 - 2.374*x_3$*	*131.8947*

OUTPUT:- Here, the value of the deviance for the model fitted on the predictors "Lumbar Pain" (x_3) is minimum. So, we can take the model fitted on "Lumbar Pain" (x_3) as M_1.

Stage- 2:-

Model	Predictors	AIC
M_1	Lumbar Pain (x_3)	135.89468
M_2	Nausea (x_2), Lumbar Pain (x_3)	101.98995
M_3	*Temp (x_1), Nausea (x_2), Lumbar Pain (x_3)*	*74.13639*
M_4	Temp (x_1), Nausea (x_2), Lumbar Pain(x_3), Micturition Pain (x_4)	76.06313

Conclusion:-

Hence, by the help of "Backward Stepwise Selection Method", we can choose the logistic regression model fitted on the predictors "Temperature of patient", "Occurrence of Nausea" and "Lumbar pain" as the best single model among all the models fitted on these four predictors.

References :-

1. UC IRVINE MACHINE LEARNING REPOSITORY
2. AN INTRODUCTION TO STATISTICAL LEARNING by GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE and ROBERT TIBSHIRANI

Student ARTICLES

STATISTICS, AS A TOOL TO FORECAST WEATHER

- Tamasha Dutta (2nd Year)

"THE GOAL OF FORECASTING IS NOT TO PREDICT THE FUTURE BUT TO TELL YOU WHAT YOU NEED TO KNOW TO TAKE MEANINGFUL ACTION IN THE PRESENT."

~ Paul Saffo

INTRODUCTION:

Weather affects people in different remarkable ways, such as, from balmy breezes to stormy winds, sunshine to rain, scorching heat to icy cold, etc. There are many ways that weather affects different businesses, households, farms, etc. Thus, to decrease the bad effects in a different field, due to climate changes, the concept of weather forecasting has been introduced.

Weather forecasting is one the most important applied field of different scientific methods and technologies to predict the conditions of the Atmosphere, for a given area and time. Different Statistical methods influence weather forecasting. It is done by collecting quantitative data (The set of information or data that can be numerically recognized and analysed) about the current state of the atmosphere, land, and ocean. It is done by using Meteorology to project the conditions of the atmosphere, that will change at a given place, according to time.

Student ARTICLES

BRIEF HISTORY OF WEATHER FORECASTING:

❖ Forecasting in Ancient times:

The concept of Weather Forecasting is introduced in 650 BC, in Babylonia. They predicted the weather from the patterns of the clouds. In about 350 BC, Aristotle described weather patterns in Meteorological (The text that Aristotle believed to have been all the affections common to the air and water, on the different parts of our earth). Later, Theophrastus compiled a book on weather forecasting (Book Name: Signs). The traditional process of Chinese weather prediction extends it, back in 300 BC, which also became the same time ancient Indian astronomers developed weather-prediction strategies.

The Ancient weather forecasting process is usually based on the observed patterns of different events, which is known as pattern recognition. In ancient times, this experience accumulated over the generations for weather forecasting.

❖ Forecasting in Modern times:

As, not all of the previous systems of predictions are reliable, a new system of prediction is introduced. This is started with the invention of Electric Telegraphs, in 1835 (The Modern age of Weather Forecasting). Basically, Sir Francis Beaufort and Robert FitzRoy are famous for modern time weather forecasting.

Student ARTICLES

Sir Francis Beaufort developed the wind force scale and Weather Notation Coding, for weather prediction. He also develops reliable tide tables around British shores, expanded the weather record-keeping at 200 British coast guard stations.

(*1) Sir Francis Beaufort

(*2) Robert FitzRoy

Later, Robert FitzRoy develops the collection of weather data at oceans. For this, all ship captains were tasked with collating data on the weather and computing it, with the use of different tested instruments. Also, a storm in the year 1859, which causes the loss of the Royal Charter, inspired Robert FitzRoy to develop different charts on the basis of the weather prediction. Here in Figure 1, One of the Weather Forecasting Charts of Europe, is shown.

Student ARTICLES

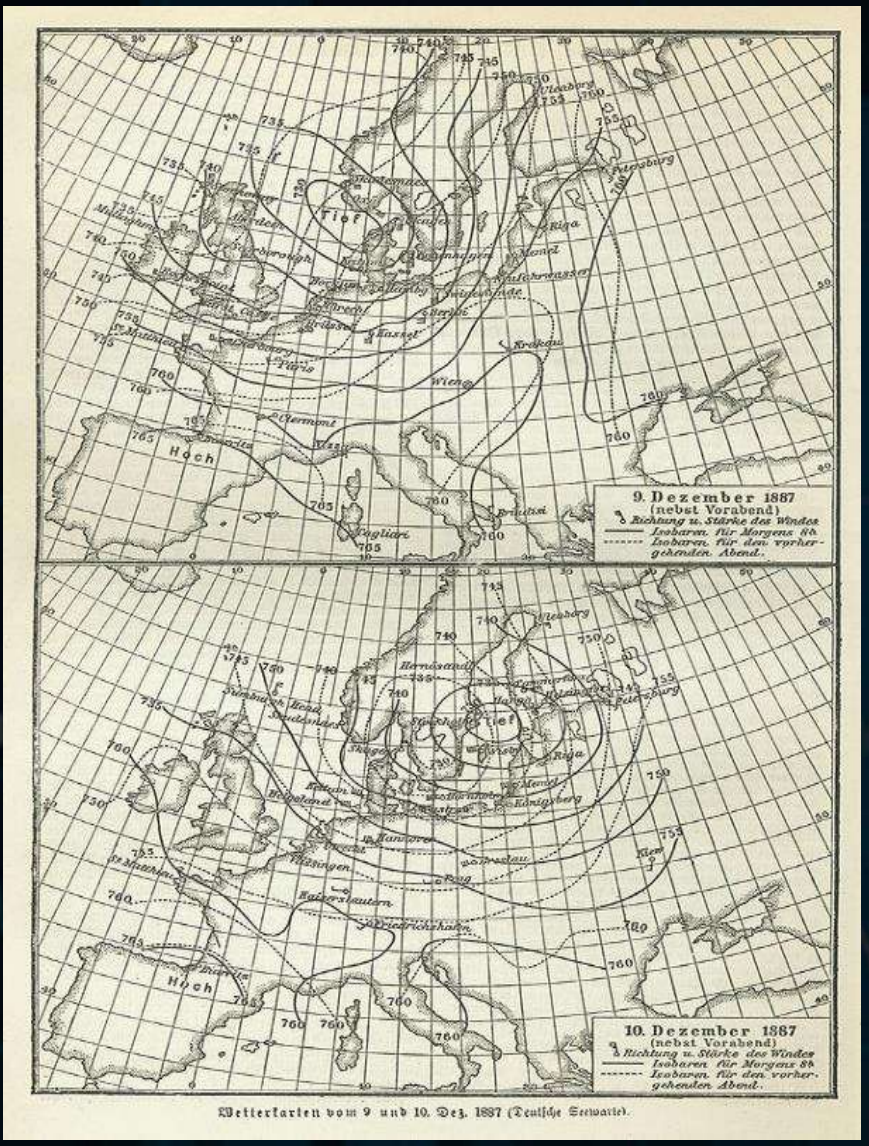


Figure 1

One of the Weather Forecasting Charts of Europe

After that, because the electric telegraph network distended, which permits to the more rapid dissemination of warnings, might be used to synoptic analyses. Completely different instruments endlessly record variations in meteorological parameters using different photographs. To convey accurate and correct information, it is necessary to process a typical vocabulary describing clouds.

TYPES OF DATA FOR WEATHER FORECASTING:

For the weather forecasting process, the collection of data has been divided into two main categories. These are:

- Surface-Weather Data:

Surface-Weather Data is the data used for different climatological reasons to forecast the weather and issue different warnings, worldwide. These Data can be taken by the weather forecasters, through computer observers, or by automated weather stations.

- Upper-Air-Weather Data:

Upper-Air-Weather Data is basically the measurements of temperature, humidity, and wind, over the surface, which are found by launching the radiosondes on the weather balloons. These data are also obtained by different aircraft, drop wind sondes, radar and satellites.

REGRESSION ANALYSIS IN FORECASTING:

In Weather forecasting, Regression analysis is one of the most important concepts. It is very useful in predicting future weather. There are basically, three usefulness for predicting the future weather. Firstly, the correlated variables do not approach to zero, until the data are taken some time apart, thus, the number of degrees of freedom becomes less than the sample size. Secondly, different variables are found to possess a relation for prediction, for the application of most models, and frequently include a large number of potential predictions. Thirdly, the process of prediction is highly skewed, so that, the weathering process, which is most important to predict, occurs very infrequently.

- Simple Linear Regression Model:

Here, the linear regression model is one of the most useful methods to predict future weather. It is used to model a relationship between two sets of variables. In 1997, sir Massie and Rose applied this model first, to predict the daily maximum temperatures. The form of the linear regression model used for this is,

$$Y = \propto + \boldsymbol{\beta} X + \boldsymbol{\varepsilon} \text{ ---------- (1)}$$

where, Y is the predictand, X is the predictor, $\propto$ and β are the parameters, which is to be estimated by the least square method, and ε be the error, due to prediction. Now, if, $\hat{Y}$ be the predicted value of Y, then, equation (1) becomes,

$$\hat{Y} = a + bX \text{ ---------- (2)}$$

Where, a and b are the estimates of $\propto$ and β.

Now, to find a and b, we will minimise the sum of square of the errors (e_i's), from the samples of size n, i.e., (X_1, Y_1), ..., (X_n, Y_n), where, X_i's and Y_i's are respectively the time periods and the corresponding observation, to be predicted. Then we have,

$$\sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=0}^{n} (Y_i - a - bX_i)^2 \text{ ---------- (3)}$$

Now, taking the partial derivative to equation (3), with respect to a and b, respectively, and equating it to 0, we get,

$$an + b \sum_{i=0}^{n} X_i - \sum_{i=0}^{n} Y_i = 0 \text{ ------- (4)}$$

$$a \sum_{i=0}^{n} X_i + b \sum_{i=0}^{n} X_i^2 - \sum_{i=0}^{n} X_i Y_i = 0 \text{ ---------- (5)}$$

Now, solving equation (4) and (5), we get,

$$a = (1/n) \sum_{i=0}^{n} Y_i - (b/n) - \sum_{i=0}^{n} X_i \text{ ---------- (6)}$$

$$b = (n \sum_{i=0}^{n} X_i Y_i - \sum_{i=0}^{n} X_i \sum_{i=0}^{n} Y_i) / n - \sum_{i=0}^{n} X_i^2 - (\sum_{i=0}^{n} X_i)^2 \text{ ------- (7)}$$

This predicting process is very helpful and useful to apply in different illustrative problems.

❖ Multiple Linear Regression Model:

In regression, there is another useful model to predict the weather. The model is known as the multiple linear regression model. This Multiple linear regression model is a very important field in applied statistical technique, that uses many instructive variables, which are used to predict the possible outcome of the response variable. The purpose of multiple linear regression is to model the relationship between the dependent and independent variables. Here, one has to form matrices from the numerical data, to get the model.

In the year 2012, sir Paras and sir Mathur used this Multiple Linear Regression model, firstly, to develop a model to forecast the weather. It was found that the proposed model is capable of forecasting the weather conditions for a particular station using the collected data.

❖ Other Useful Regression Models:

- Stepwise Regression:

Here, the stepwise regression models are used to predict variables, that are carried out by an automatic procedure. In every step, a variable is considered for addition or for subtraction, from the set of explanatory variables.

- Logit Regression:

Logistic regression is one of the most powerful and useful regression models, that is used when the dependent variable is binary. This regression model is used to describe the data and the relationship between a dependent binary variable and one or more nominal, ordinal, interval, and ratio independent variables.

In the year 2010, Sir Prasad et al., developed a multi-predictable logistic regression model for forecasting probabilistically, for the average rainfall on a certain monthly timescale for a certain region.

- Quantile Regression:

The quantile regression is used to get a more appropriate structure of the effect of the independent variables on the dependent variable. Rather than estimating the model with average effects, using the linear model, this regression process produces different effects along with the distribution of the dependent variable.

Sir Lauret et al., in the year 2017, proposed a linear quantile regression method to generate one hour to six-hour ahead probabilistic forecasts of solar irradiance at a site experiencing highly variable sky conditions.

REMARK: Still, there are some more regression models, employed to forecast weather, like, Optimal Subset Regression, Weighted Regression, Kalman filter Regression, Bayesian Regression, Ridge Regression, Lasso Regression, etc.

The advancement of technology has enabled us to obtain forecasts using complex mathematical models, which exhibit new directions with forecasting weather as a data-intensive challenge that involves inferences across space and time.

However, weather forecasts still have their limitations despite of using modern technology and improved techniques to predict future weather. But weather forecasting is still a complex and not always an accurate one, in its nature.

TIME SERIES ANALYSIS IN WEATHER FORECASTING:

Time series data is a sequence of observations in an ordered time. Mostly, these data are collected at equally spaced time intervals. Here, in this case, our basic assumption is that some aspects of the past pattern will continue to remain in the future.

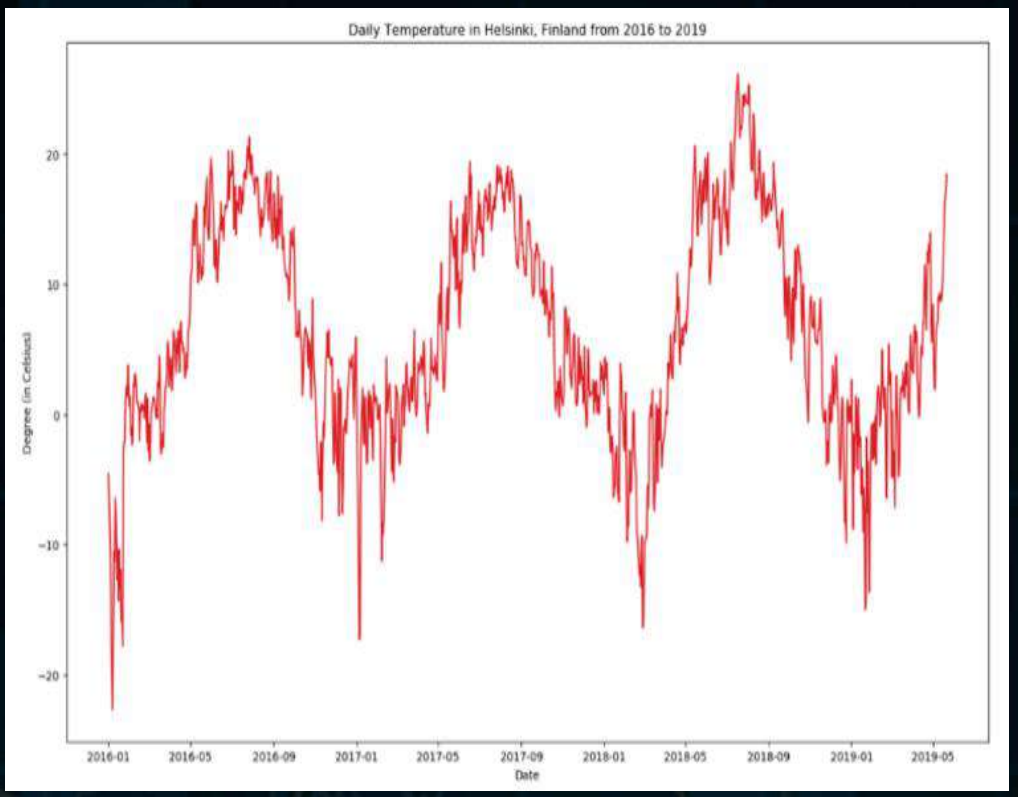


Figure 2

The time series plot shows the average daily temperature in Helsinki, Finland, from 2016 to 2019.

Thus, time series analysis can be used more easily for weather forecasting purposes. All these data are statistically dependent and time series modelling is used with the techniques for analysis of such dependent variables. Nowadays, these data are frequently used for the Weather Prediction process. Here, Figure 2, shows a time series plot of the average temperature in Helsinki, Finland, from the year 2016 to the year 2019, for each day.

By theoretical concept, time series analysis methods should be applied as the variables defining weather conditions like rainfall, temperature (like- maximum or minimum in temperature), relative humidity, etc., which continuously varies with respect to time.

CONCLUSION:

Using different terms of Statistical models in weather Forecasting, it is quite good to get a more or less correct prediction. These numerical data are made to the observations collected over many years. Weather forecasting also depends on the efficiency of the interplay of weather observation, the data analysis by meteorologists and computers, and communication systems. However, weather forecasting processes still have some limitations, as, Weather forecasting is not always accurate. But, the more efficient way to process large data with a good method of statistical model, will help to improve and extend the application of statistical forecasting in the future.

REFERENCES:

Sources for the Figures:

Figure 1:
https://en.m.wikipedia.org/wiki/Weather_forecasting

Figure 2:
https://medium.com/@llmkhoa511/time-series-analysis-and-weather-forecast-in-python-e80b664c7f71

Sources for the Equations in Simple Linear Regression Model:
https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.ecmwf.int/sites/default/files/elibrary/1982/9605-statistical-weather-forecasting.pdf&ved=2ahUKEwix4aO5xvzzAhU_ILcAHXmwBqMQFnoECDsQAQ&usg=AOvVaw2o1Chd_y9-vbtF-dbAv_Rz

(Page 14 & 15)

FOR FURTHER READINGS:

1. https://en.m.wikipedia.org/wiki/Weather_forecasting

2. https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.ecmwf.int/sites/default/files/elibrary/1982/9605-statistical-weather-forecasting.pdf&ved=2ahUKEwix4aO5xvzzAhU_ILcAHXmwBqMQFnoECDsQAQ&usg=AOvVaw2o1Chd_y9-vbtF-dbAv_Rz

Student ARTICLES

3. https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.linkedin.com/pulse/regression-analysis-weather-forecasting-chonghua-yin&ved=2ahUKEwjolsy22ID0AhWDfH0KHW4zAIEQFnoECAQQBQ&usg=AOvVaw3DcDkxpq7W8QpdxEgqO2OH

4. https://r.search.yahoo.com/_ylt=Awrx356VjNZhpgsAGga7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1641479445/RO=10/RU=https%3a%2f%2fen.wikipedia.org%2fwiki%2fFrancis_Beaufort/RK=2/RS=HI.Zqjn1dRaG.2sLLXyVOHHE.sI- (*1)

5. https://r.search.yahoo.com/_ylt=AwrxzhV9jdZhmjMAsx67HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1641479677/RO=10/RU=https%3a%2f%2fen.wikipedia.org%2fwiki%2fRobert_FitzRoy/RK=2/RS=ohIvVCaGV3weffolrOs746yHXuE- (*2)

6. https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.ijert.org/research/weather-forecasting-models-methods-and-applications-IJERTV2IS120198.pdf&ved=2ahUKEwiP85yl4u31AhXlheYKHcr7CSQQFnoECBkQAQ&usg=AOvVaw17tNP9Bs-1EwsVZHzxmvGD

7. https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.semanticscholar.org/paper/Operational-Weather-Forecasting-Inness-Dorling/1ad8c94961888430202f5969a43cde08f69f77ef&ved=2ahUKEwiP85yl4u31AhXlheYKHcr7CSQQFnoECBoQAQ&usg=AOvVaw3_97ka3jnpjQG0kxPLXAX6

8. https://www.scribd.com/book/350177633/Weather-Forecasting-Made-Simple

Student ARTICLES

Existence And Non-Existence of Minimum Variance Unbiased Estimator (MVUE)

- SATYAKI BASU SARBADHIKARY (3rd Year)

An unbiased estimator that has the least variance of all other unbiased estimator of the parameter is known as Minimum-variance unbiased estimator (MVUE) or uniformly minimum-variance unbiased estimator (UMVUE).

Before proceeding further let us define what is an unbiased estimator.

UNBIASED ESTIMATOR

DEFINITION:

An estimator of a given parameter is unbiased if its expected value is equal to the value of the parameter.

Estimable function

A parametric function $\gamma(\theta)$ is said to be an estimable function if $\exists$ an estimator T, such that

$E_\theta[T] = \gamma(\theta), \forall\ \theta \in \Theta.$

i) Unbiased estimator may not exist:-

Let us consider a random variable, $X \sim Ber(\pi)$ If possible, let T(X) be an unbiased estimator for π^2. Here, $\gamma(\pi) = \pi^2$, is the parametric function i.e., $E_\pi[T(X)] = \pi^2 \Rightarrow (T(1) \times \pi) + (T(0) \times (1 - \pi)) = \pi^2$

$\Rightarrow \pi^2 + \pi(T(0) - T(1)) - T(0) = 0$

This is a quadratic equation in π. Only two values of π satisfies the equation, but in order to be estimable, $E_\pi[T(X)] = \pi^2, \forall\ \pi \in (0,1)$.

This is a contradiction, so T(X) is not an unbiased estimator of π^2. i.e., for π^2 has no unbiased estimator.

ii) Unbiased estimator may exist, but it can be absurd:-

Let us consider a random variable,

$X \sim Poi(t)$

Let $T(X) = (-2)^X$

So, $E[T(X)] = \sum_{x=0}^{\infty}(-2)^x . \frac{e^{-t}t^x}{x!}$

$= e^{-t}.\sum_{x=0}^{\infty}\frac{(-2t)^x}{x!}$

$\Rightarrow E[T(X)] = e^{-t}. e^{-2t} = e^{-3t}$

∴ T(X) is an unbiased estimator of $e^{(-3\lambda)}$.

Now, $(-2)^X > 0$; if X is even

$(-2)^X < 0$; if X is odd

whereas e^{-3t} is a strictly positive quantity. So, it is absurd to use a negative estimator for a strictly positive quantity.

Minimum Variance Unbiased Estimator (MVUE):

An estimator T is said to be UMVUE of a parametric function $\gamma(\theta)$ if,

i) $E_\theta[T] = \gamma(\theta)\ \forall\ \theta \in \Theta$.

ii) For any other unbiased estimator T* of $\gamma(\theta)$, $Var_\theta(T) \leq Var_\theta(T^*)$

A necessary and sufficient condition for existence of UMVUE

Let M be the class of unbiased estimators of a parametric function $\gamma(\theta)$ and μ_0 be the class of unbiased estimators of 0. Then T is UMVUE of $\gamma(\theta)$ iff $cov_\theta(T,T_0) = 0, \forall\ T_0 \in \mu_0$, [$T_0$ is an unbiased estimator of 0]

Proof –

(If Part)

Here it is given that, $cov_\theta(T,T_0) = 0\ \forall\ T_0 \in \mu_0$ i.e., $E_\theta[T_0] = 0$ Also $E_\theta[T_0] = \gamma(\theta)\ \forall\theta \in \Theta$ Here we are required to prove that T is UMVUE of $\gamma(\theta)$. Let T^* be any other unbiased estimator of $\gamma(\theta)$. i.e., $E_\theta\ [T^*] = \gamma(\theta)$ i.e., $E_\theta[T - T^*] = E_\theta[T] - E_\theta[T^*] = \gamma(\theta) - \gamma(\theta) = 0$; i.e.,

$(T - T^*) \in \mu_0$ Thus, $(T - T^*)$ is an unbiased estimator of 0. So, $cov_\theta(T, T - T^*) = 0$

$[\because (T,T_0) = 0$ where $T_0 \in \mu_0] \Rightarrow Var_\theta(T) - Cov_\theta(T, T^*) = 0 \Rightarrow Var_\theta(T) = Cov_\theta(T, T^*)$ . . .(i)

Again, $Var_\theta(T - T^*) \geq 0 \Rightarrow Var_\theta(T) + Var_\theta(T) - 2Cov_\theta(T, T^*) \geq 0$

$\Rightarrow Var_\theta(T) + V ar_\theta(T) - 2Var_\theta(T) \geq 0$ [from (i)]

So, $Var_\theta(T^*) \geq Var_\theta(T)$

∴ T is the UMVUE of γ(θ). (Proved)

Only if part:-

Here T is the UMVUE of γ(θ). Now, let T_0 be an unbiased estimator of 0.

If possible, let

$Cov_\theta(T, T_0) \neq 0$

Define, $T^* = T + \lambda T_0$, where λ is an arbitary constant

i.e., $E_\theta[T^*] = E_\theta[T] + \lambda E_\theta[T_0] \Rightarrow E_\theta[T^*] = E_\theta[T] \Rightarrow E_\theta[T^*] = \gamma(\theta)$ i.e., T^* is an unbiased estimator of γ(θ).

Again, $Var_\theta(T^*) = Var_\theta(T) + \lambda^2 Var_\theta(T_0) + 2\lambda Cov_\theta(T, T_0)$

$\Rightarrow Var_\theta(T^*) = Var_\theta(T) + Var_\theta(T_0)[\lambda^2 + 2\lambda Cov_\theta(T, T_0) / Var_\theta(T_0)]$

$\Rightarrow Var_\theta(T^*) = Var_\theta(T) + Var_\theta(T_0) [\lambda^2 + 2\lambda Cov_\theta(T, T_0) / Var_\theta(T_0) + \{ Cov_\theta(T, T_0)/ Var_\theta(T_0) \}^2 - \{Cov_\theta(T, T_0)/ Var_\theta(T_0)\}^2]$

$\Rightarrow Var_\theta(T^*) = Var_\theta(T) - Var_\theta(T_0) \{Cov_\theta(T, T_0)/ Var_\theta(T_0)\}^2 + Var_\theta(T_0) \{ \lambda + Cov_\theta(T, T_0)/ Var_\theta(T_0) \}^2$

$\Rightarrow Var(T^*) = Var_\theta(T) - (Cov_\theta (T,T_0)^2/ Var_\theta(T_0)$ [we take $\lambda = -Cov_\theta(T,T_0)/ Var_\theta(T_0)$]

Now,

$\text{Cov}_\theta(T,T_0) \neq 0$ [initial assumption]

$\therefore (\text{cov}^2_\theta(T,T_0))^2 / \text{Var}_\theta(T_0) \geq 0$

i.e., $\text{Var}_\theta(T^*) < \text{Var}_\theta(T)$

This is a contradiction as T is UMVUE

$\therefore \text{Cov}_\theta(T,T_0)=0$ (proved)

Thus, the necessary-sufficient condition for an unbiased estimator to be the UMVUE, is that the unbiased estimator has to be uncorrelated with all unbiased estimator of zero.

UMVUE according to Lehmann Scheffé Theorem

Lehmann Scheffé Theorem: Let h be any arbitrary unbiased estimator of a parametric function γ(θ) and T be a complete sufficient statistic for γ(θ). Then E(h|T) is the UMVUE of γ(θ).

Proof:-

Let h_1 and h_2 be two arbitrary unbiased estimators of γ(θ). Let, T be a complete sufficient statistic for γ(θ). Then by Rao-Blackwell theorem:- i) $E(H_1|T)$ and $E(H_2|T)$ both are unbiased estimators of γ(θ). ii) Var $[E(h_i|T)] \leq \text{Var}(h_i)$; i = 1, 2 Now,

$E_\theta[E(h_1|T)] - E_\theta[E(h2|T)] = \gamma(\theta) - \gamma(\theta) = 0$

Define, $g(T) = E(h_1|T) - E(h_2|T)$, which is a function of complete sufficient statistic T i.e., $E_\theta[g(T)] = 0 \Rightarrow g(T) = 0$ almost everywhere $\Rightarrow E(h_1|T) - E(h_2|T) = 0$, almost everywhere $\Rightarrow E(h_1|T) = E(h_2|T)$, almost everywhere i.e., this conditioning is unique,

So, $E(h_1|T) = E(h_2|T) = E(h|T)$ is the UMVUE of $\gamma(\theta)$.

Implication:- If any function of complete sufficient statistic is unbiased for $\gamma(\theta)$ then that function is the UMVUE of $\gamma(\theta)$.

Non-Existence of Minimum-variance unbiased estimator (MVUE):

Sometimes there may not exist any MVUE for a given data. This can happen because of the two reasons as follows:

1) Non-existence of unbiased estimators
2) Even if unbiased estimator exists, none of them gives uniform minimum variance.

We consider having three unbiased estimators e_1, e_2 and e_3 which estimates a parameter θ.

Let the unbiased estimates be
$\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ respectively.
The diagram below shows two cases for the existence of MVUE among the three estimators.
In diagram (a), the third estimator gives uniform minimum variance compared to other two estimators.
In diagram b, none of the estimator gives minimum variance that is uniform across the entire range of θ.

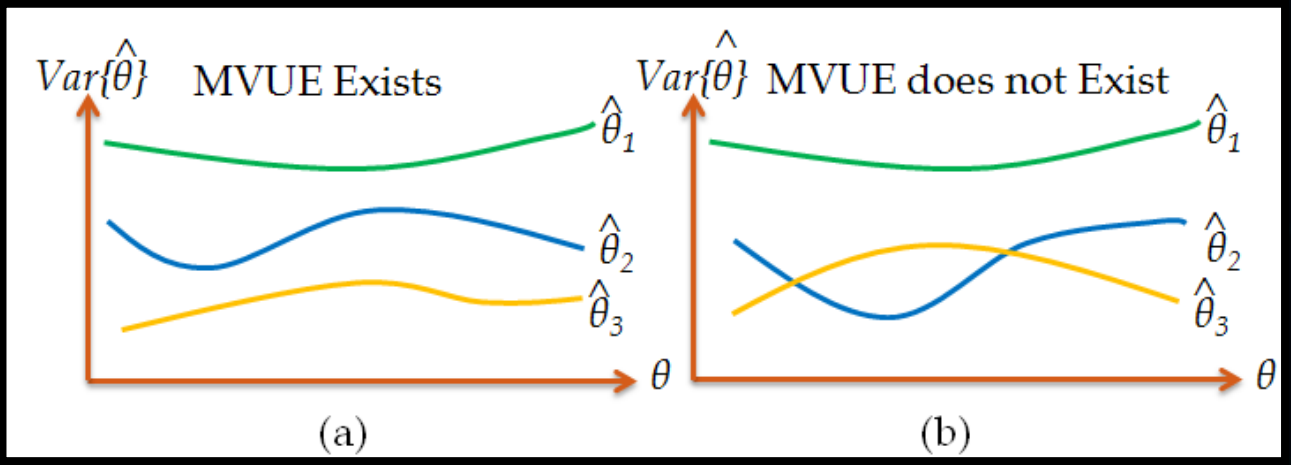


Diagram depicting existence and non-existence of Minimum Variance Unbiased Estimator (MVUE)

We take a single observation X following U (θ, θ+1). Suppose we wish to see whether UMVUE of a parametric function g(θ) actually exists or not,

If possible, suppose T is UMVUE of g(θ). Let U_0 be the class of all unbiased estimators of zero.

Obviously for every f ∈ U_0,

$$\int_{\theta}^{\theta+1} f(x)dx=0, \forall\ \theta \in R$$

Now on differentiating both sides of the above equation with respect to θ we get

$$f(\theta+1) = f(\theta) \quad \text{------(1)}$$

Because T is UMVUE, E_θ (Tf)=0 for all θ and for all f∈U_0.

Or, $Tf \in U_0$ whenever $f \in U_0$. So, similarly as (1) we have

$$T(\theta+1).\ f(\theta+1) = T(\theta).f(\theta) \text{----- (2)}$$

And (1) implies

$$T(\theta) = T(\theta+1) \text{ --------(3).}$$

Also, since T is unbiased for θ,

$$\int_{\theta}^{\theta+1} T(x)dx = g(\theta),\ \forall\ \theta \in R \text{ -------(4)}$$

Differentiating both sides of equation (4) w.r.t θ and equation (3) gives

$$g'(\theta) = T(\theta+1) - T(\theta) = 0.$$

This shows that g(θ) does not take UMVUE for any g (g non constant is assumed)

Therefore, if we take g(θ)=θ then T=X-1/2 is unbiased for θ but it is not UMVUE. Similarly for n observations,

$\overline{X_n}$-1/2 is unbiased but not a UMVUE.

Student ARTICLES

List of References:

1.) Statistical Inference (Second Edition) by George Casella, Roger Berger

2.) Introduction to The Theory of Statistics by Alexander M. Mood, Franklin A. Graybill, Duane C. Boes

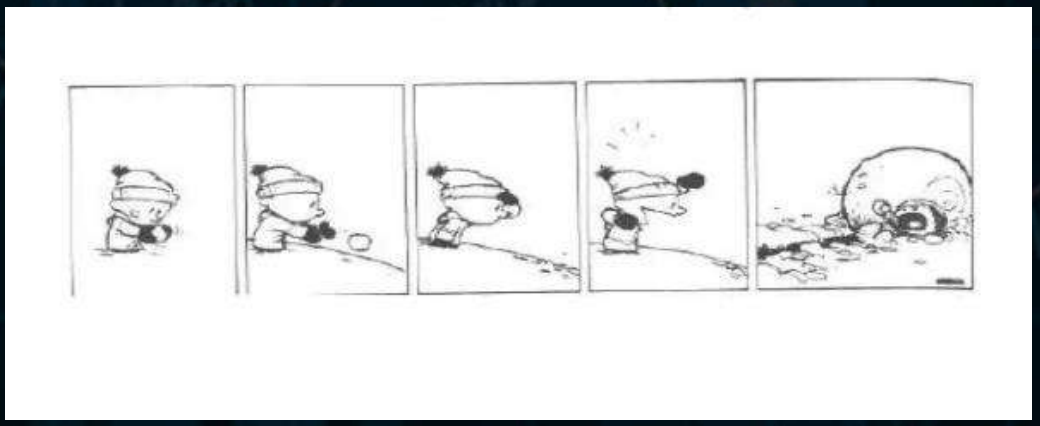
3.) An Outline Of Statistical Theory (Volume2) by A. M. Gun, M. K. Gupta, B. Dasgupta

Student ARTICLES

Snowball Sampling- a Sampling Technique Less Talked About

- Soham Chatterjee (2nd Year)

You are rolling a little snowball down a hill covered in snow. As it continues to roll, it gathers more snow, increasing its surface area and bulk. The snowball effect is what it is called. In figurative terms, it refers to anything insignificant that is gaining a lot of traction with the passage of time.



What does this have to do with statistics?

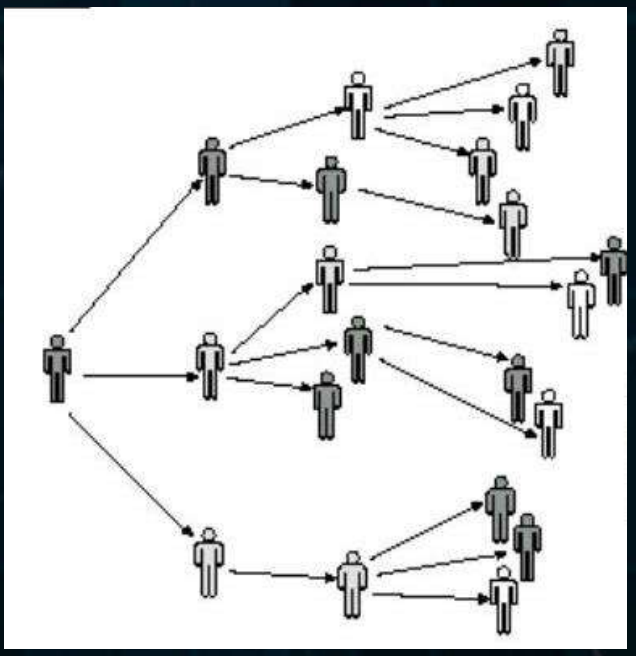
Snowball sampling is a technique that is based on this simple concept.

Student ARTICLES

We have learnt about different types of probabilistic sampling, such as simple random, stratified, systematic, and so on. We do, however, use non-probabilistic sampling approaches in certain problematic scenarios. Snowball Sampling is one such innovative approach. Now what are those difficult situations in this case? We will come to it soon.

The first roll down the hill:

To begin, you must first identify a snowball, or a single subject that is relevant to your study project. He/she then suggests another person who would be a good fit for the survey/project. The same procedure is used this time: they select their friends/acquaintances who share the same psychographic characteristics. Snow accumulates in the snowball.



As the sample size grows, enough data for the survey/project is acquired.

Student ARTICLES

Purpose:

What would a surveyor do if he wanted to find the 20 persons in his country, or the 200 people in the world, who have an exceptionally rare disease? Or, for example, how would he research certain illegal or subterranean activities/drug addicts? How, for example, would someone do health research on disadvantaged or stigmatized populations like sex workers or communities that are hesitant to engage in any survey? What about the homeless? People with uncommon abilities?

Such populations are difficult to find due to their scarcity. These are often referred to as "hidden populations" or "concealed populations", groups for which no official data exists. The only and best strategy to use is snowball sampling. Given the option, it is easy to see why some researchers prefer or require snowball samples.

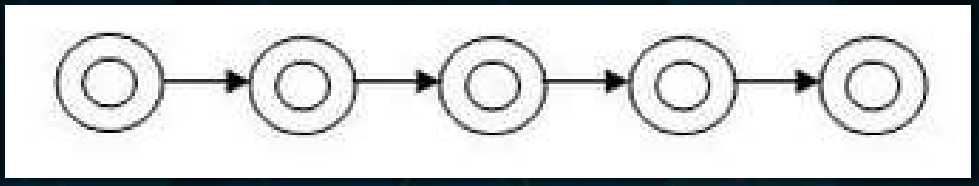
The growth of the snowball:

Chain sampling or Chain-referral sampling are other terms for snowball sampling. Greater links in each chain will provide more information about a given sample and may also allow access to the most difficult-to-find samples.

There is no sampling frame in contrast to probabilistic sampling. This sampling can be classified into three categories:

1. Linear Chain-referral:

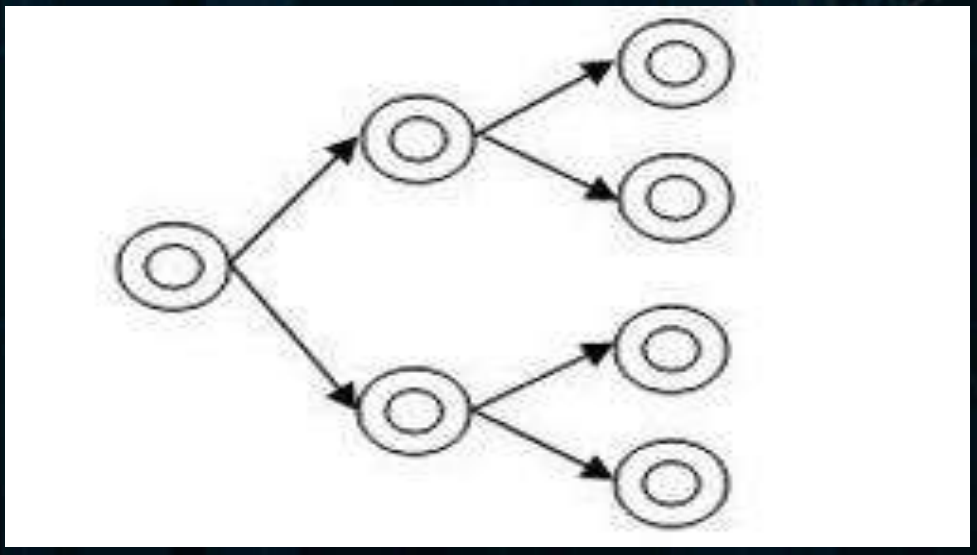
The sample is created with only one subject, and the subject only provides one recommendation, who then refers one more potential subject, and so on. (See Figure 1)



(Figure 1)

2. Exponential non-discriminative snowball sampling:

Unlike the preceding one, the initial subject pertains to a number of different participants. Each new referral is examined until a sufficient quantity of samples has been obtained. (See Figure 2)

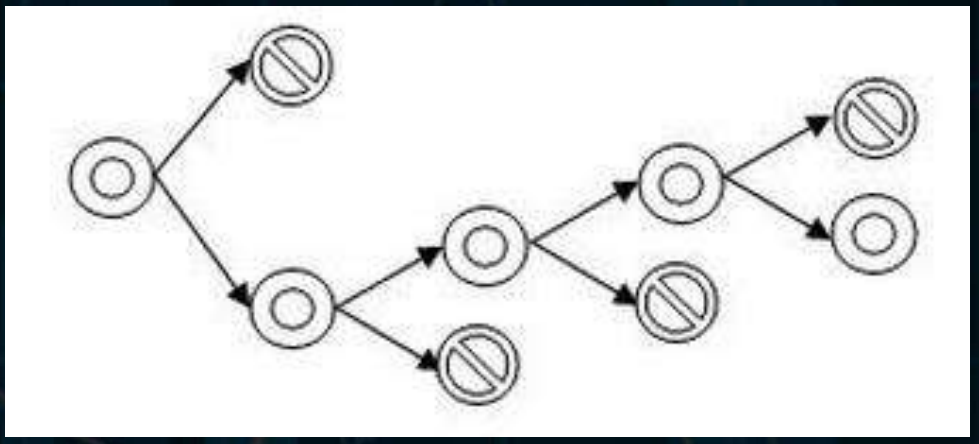


(Figure 2)

3. Exponential discriminative snowball sampling:

Subjects can suggest numerous people, but only one new person is recruited.

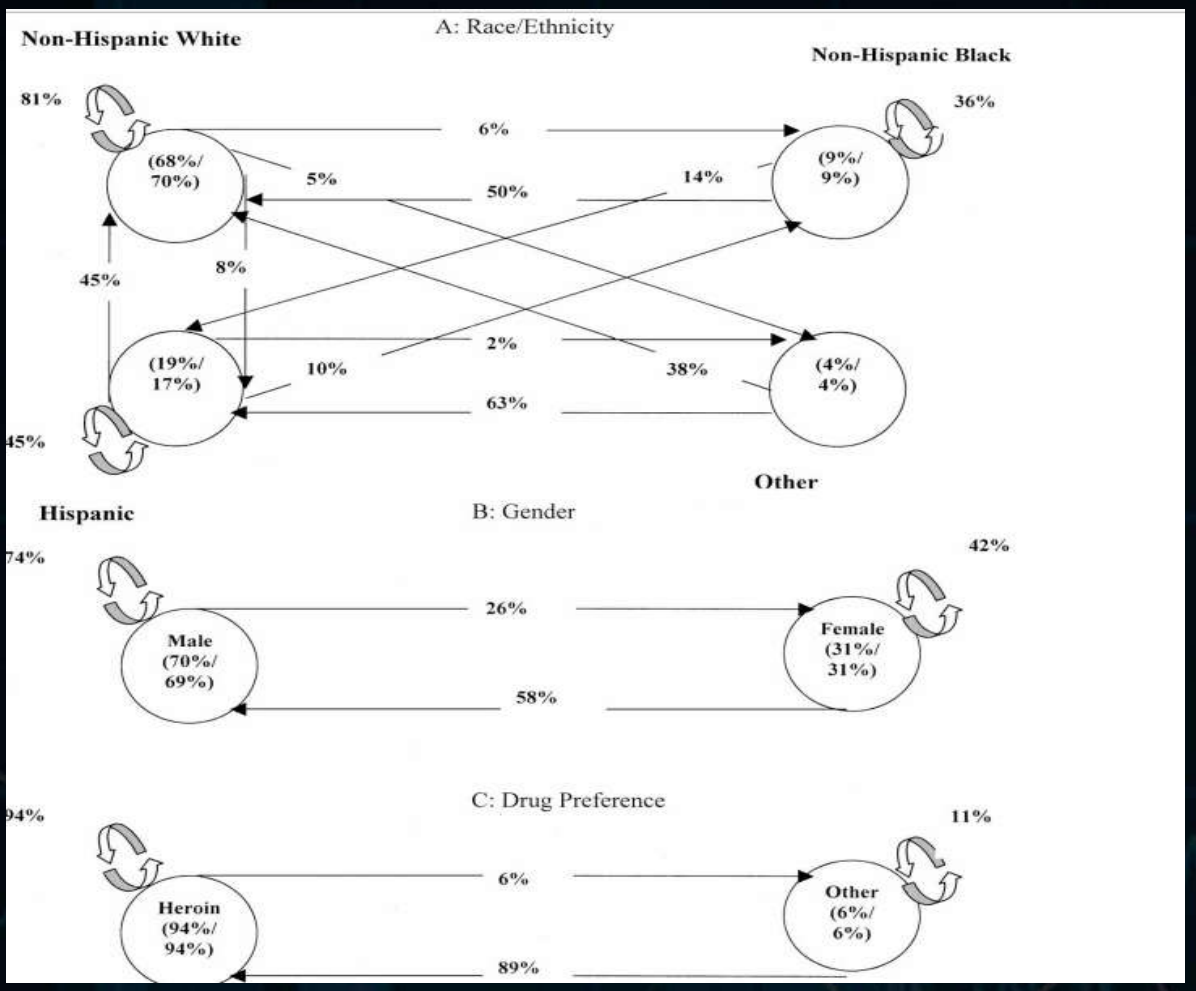
(The research's goal and objectives can help you choose a new subject.) (See Figure 3)



(Figure 3)

Bias- the biggest disadvantage:

There are biases in every sampling technique, whether probabilistic or non-probabilistic. Because no attempt is made to obtain a random sample, it is easy to observe that snowball samples are biased. There is a bias in the community. The study becomes extremely reliant on the initial volunteers; it's conceivable that many of them will have similar qualities or characteristics (a phenomenon known as homophily), which may be unrepresentative of the whole community. As a result, snowball sampling is frequently used in exploratory data analysis.



(Figure 4)

FIGURE 4:

Varying levels of homophily (similar features bias), i.e., a desire for links within the group, can be shown in drug injector recruiting patterns. The percentages in each node refer to the post sample, with the latter being the amount that approaches as the number of response waves grows. The arrows reflect the likelihood of each response group being recruited.

P.S. Respondent-driven snowball sampling is a modification that, in addition to chain referral, sets sample weighting to negate the preliminary non-random picking, which adjusts for the bias. Another alternative that reduces selection bias is Peer Esteem Snowballing (PEST). Formulating models to eliminate selection bias is still an ongoing endeavor in statistics research. In network data analysis, social networks use this concept to unbiasedly estimate population count.

References:

1. *Why would anyone use a biased snowball sample for legitimate research? | Canadian Viewpoint Inc - Your Market Research Tech Partner (canview.com)*
2. *(PDF) Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research (researchgate.net)*
3. *Snowball sampling (research-methodology.net)*
4. *Snowball sampling - Wikipedia*
5. *Journal of Studies in Social Sciences ISSN 2201-4624 Volume 5, Snowball Sampling Completion Irina-Maria Dragan.*
6. *Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18–25 Douglas D. Heckathorn,1,5 Salaam Semaan,2 Robert S. Broadhead,3 and James J. Hughes.*

Student ARTICLES

Kurtosis: Diving in its Controversy

- Subhajit Karmakar (2nd Year), Anik Chakraborty (3rd Year)

The Controversies . . .

"Kurtosis measures the degree of 'peakedness' of a distribution..."

"Kurtosis tells us nothing about the 'peak' of a distribution, it gives an idea whether there are outliers..."

"Kurtosis can be interpreted as the degree of deviation from normality in comparison to the 'peakedness'..."

"Kurtosis measures the probability concentration inside the range $(\mu \pm \sigma)$*..."*

Textbooks and Journals: What they say?

Textbooks and journals describe kurtosis as,

"The distribution with higher peak than normal distribution has positive kurtosis $(\gamma_2 > 0)$ *and negative kurtosis indicates that the distribution has a lower peak than normality."*

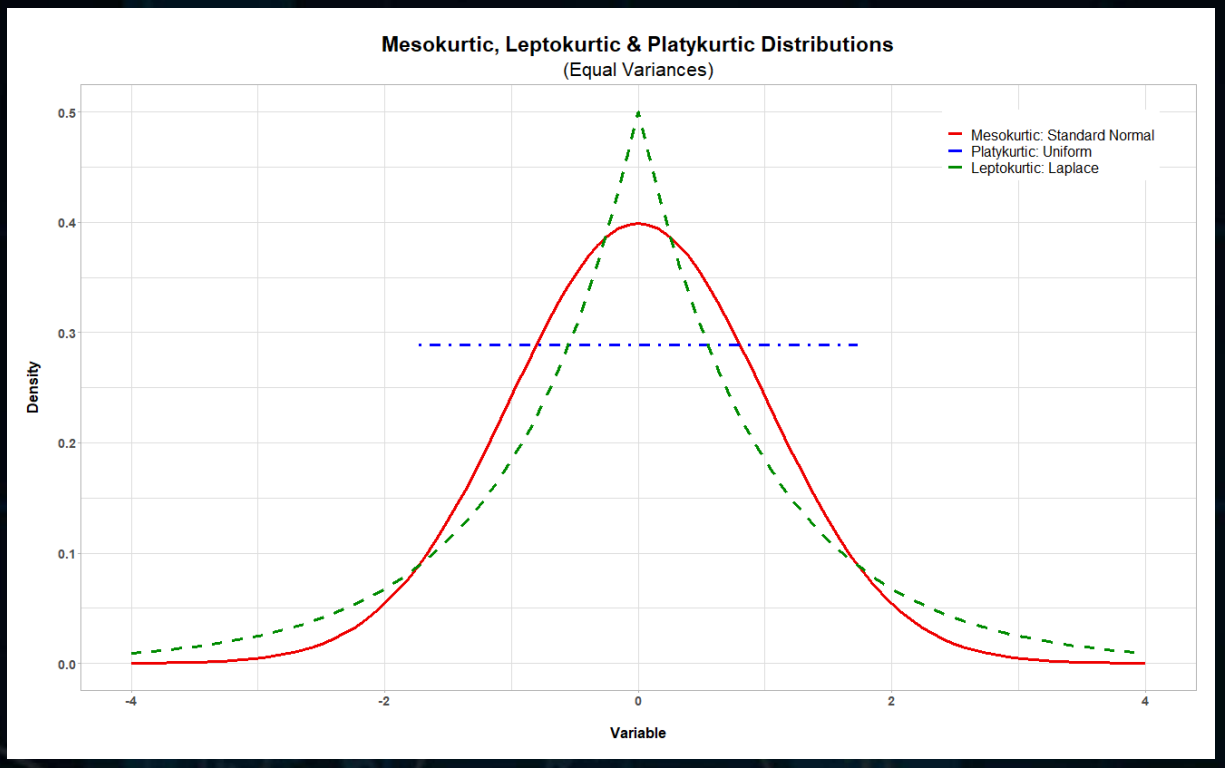


Figure 1: Peakedness & Kurtosis

Revealing the Truth . . .

Suppose X is our variable of interest with $\mathrm{E}(\mathrm{X}) = \mu$ and σ be the standard deviation. If the kurtosis of the distribution is γ_2' then,

$$\gamma_2 = \frac{E(X-\mu)^4}{\sigma^4} \quad ,$$

Now, for a given distribution σ is a constant. Hence, the value of γ_2' will depend only on the expression $E(X-\mu)^4$.

What does the expression $\mathbf{E(X-\mu)^4}$ try to tell us?

Note that, (X – μ) gives the deviation of the values of the random variable X from its mean. This deviation will be larger in magnitude if the values of X are too far away from μ (i.e., outliers) and will be smaller for values which are nearer to μ.

Now, as we raise the values (X – μ) to their 4th power, the larger values of the deviation will become more larger and smaller values (i.e., values nearer to 0) will become smaller. As a result of which, we observe that on an average the extreme values (outliers) of X contribute more to $E(X-\mu)^4$ and for the central part of the distribution the contribution is almost nil.

Hence, we observed that, kurtosis gives an idea of presence of outliers or *'tailedness'* rather than anything about *'peakedness'* of a distribution.

Sample Counterpart

The sample analogue of the measure of kurtosis for n observations $(x_1, x_2, \ldots, x_n)$ is given by,

$$g_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4, \qquad where\ s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

We have $E(g_2) \approx \gamma_2$, for large 'n'

Essentially, γ_2 measures the propensity to produce outliers from the probability distribution and g_2 measures whether the sample contains any outlier or not.

Tailedness & Peakedness: Work Together!

Let, us visualize the theoretical concept discussed above through diagrams to get a better grip of the discussion.

We consider here the *Student's t* distribution.

For this distribution the population kurtosis is,

$$k = \frac{6}{m-4},\ m > 4$$

where, m is the degrees of freedom of the t distribution.

From the theoretical expression it can be readily seen that, as m increases, kurtosis of t_m distribution decreases. We now choose three values of m = 5, 15, 25 and plot their respective theoretical densities to verify the result *graphically*.

Case I: m = 5

Here, for m = 5, $\gamma_2 = \frac{6}{5-4} = 6$

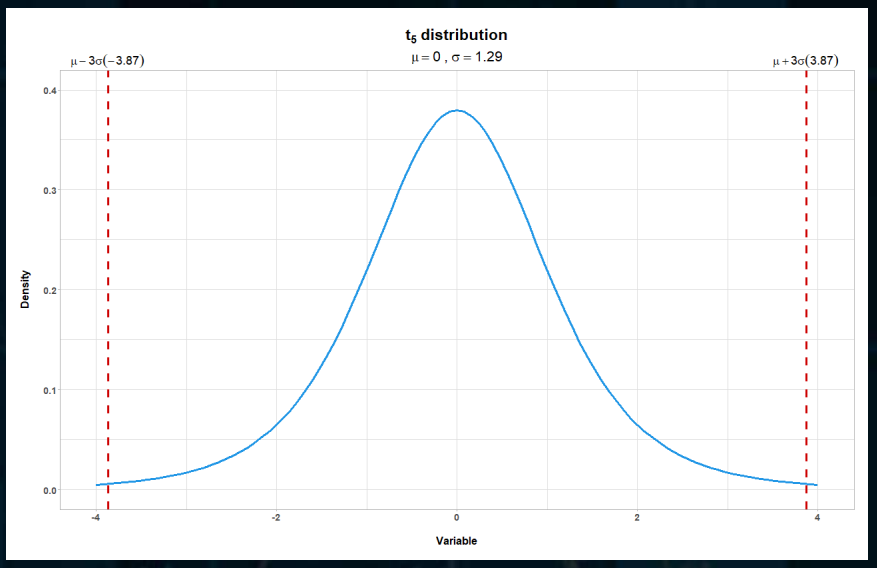


Figure 2: PDF of t_5 distribution

Case II: m = 15

Here, for m = 15, $\gamma_2 = \frac{6}{15-4} = 0.55$

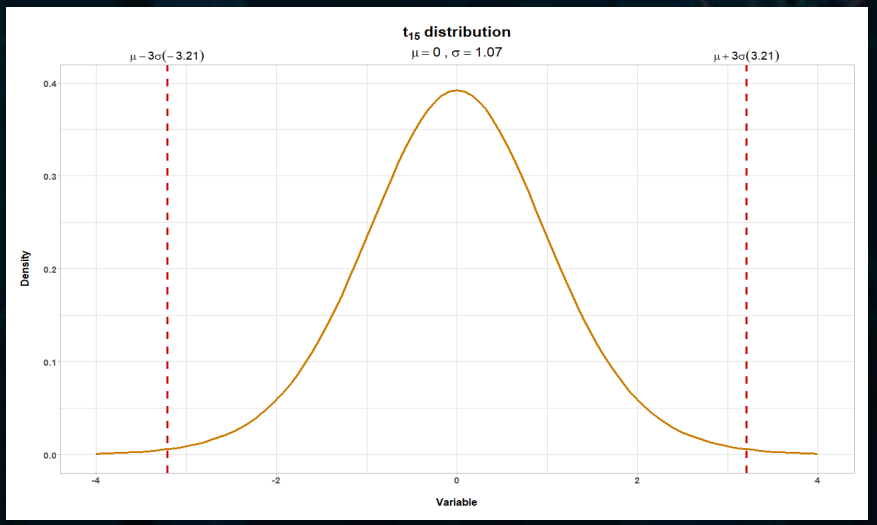


Figure 3: PDF of t_{15} distribution

Case III: m = 25
Here, for m = 25, γ_2 = 6/(25-4) = 0.286

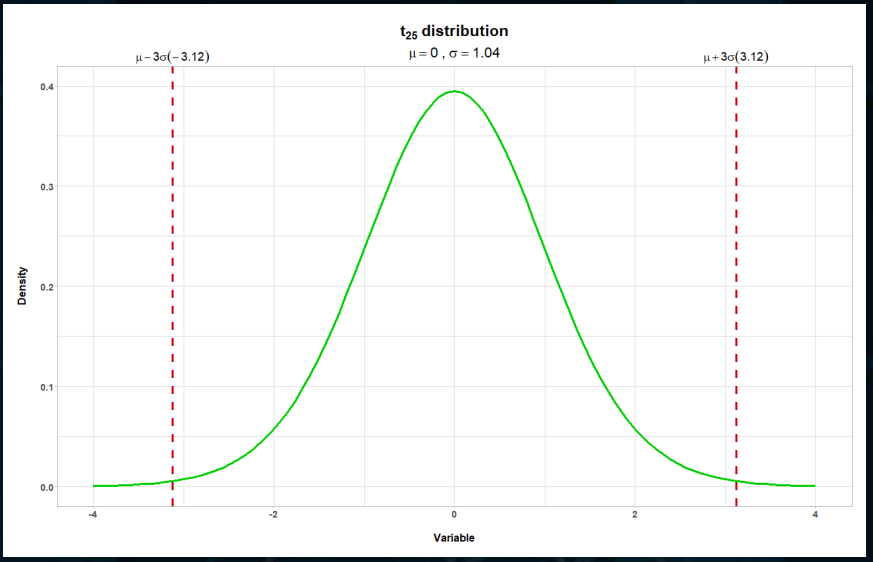


Figure 4: PDF of t_{25} distribution

Observations

From 'Figure 2' we observe that, there are too many extreme values in the distribution i.e., the tail of the distribution can be considered as '*thick*'.

Note that, in 'Figure 3' we can observe the presence of extreme values in t_{15} distribution but the densities at extremities are less than that of t_5 distribution.

Lastly, in 'Figure 4' fewer extreme values can be observed than both t_5 and t_{15} distributions.

Hence, from the above three cases, it seems that *as the degrees of freedom increases, tail of the distribution become lighter, and kurtosis also decreases*.

Observe that, the peakedness also decreases as the curve tends to bell shaped for large degrees of freedom.

Normal & t_5 : A Comparison

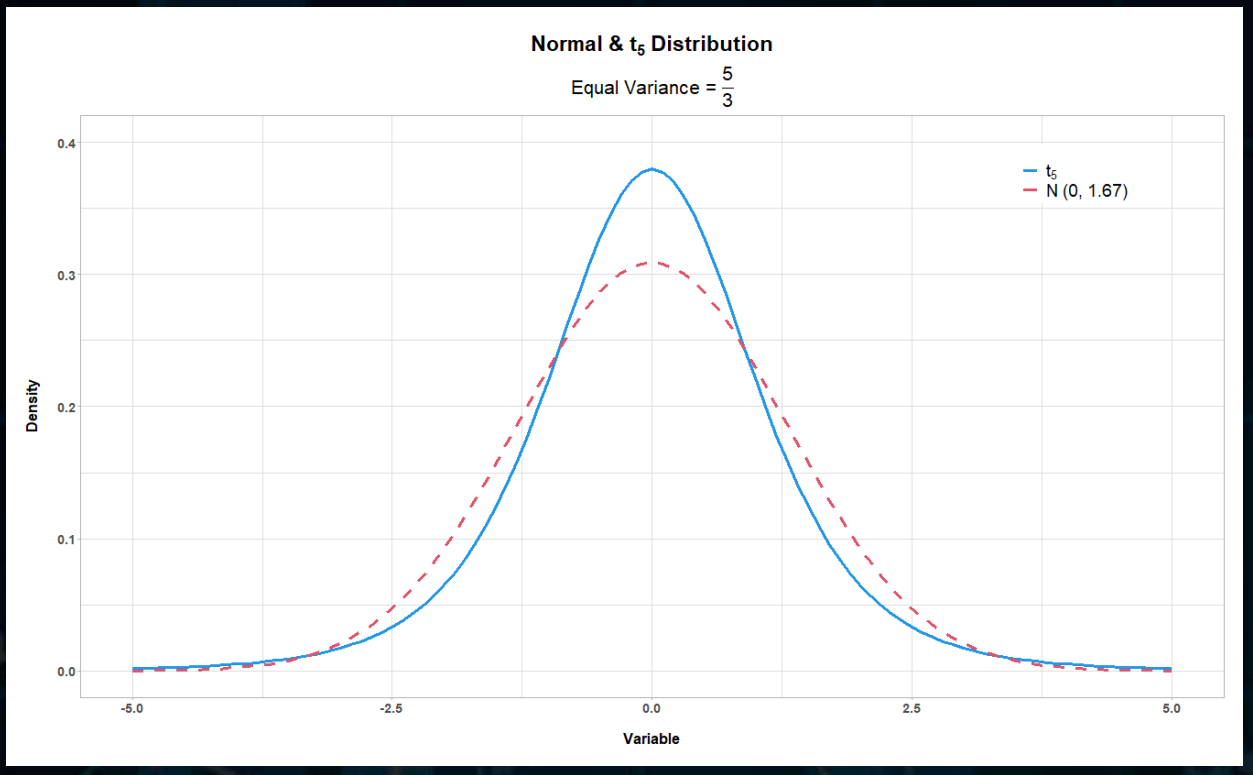


Figure 5: Comparing the Kurtosis

Note that the t_5 distribution has a variance of $\frac{5}{3}$, and the normal distribution shown in the 'Figure 5' is scaled to also have a variance of $\frac{5}{3}$.

Hence, from the above theoretical density plots (Figure 5), we can observe that t_5 has a thicker tail and a higher peak as compared to $N(0,\frac{5}{3})$ distribution. Thus, it matches with both the theory of *"tailedness"* and *"peakedness"* as the kurtosis is higher in case of t_5 distribution than $N(0,\frac{5}{3})$.

Lawrence T. DeCarlo in his article 'On the Meaning and Use of Kurtosis' told us, "...the t_5 distribution crosses the normal twice on each side of the mean, that is, the density shows a pattern of higher-lower-higher on each side, which is a common characteristic of distributions with excess kurtosis."

Hence, it can be said that both the concepts, "tailedness" and "peakedness" coexist here. In general, according to Karl Pearson, for the symmetric bell-shaped curves the idea of "peakedness" go with the kurtosis of a distribution.

The Parting of the Ways . . .

Let us now lean onto some cases where *"peakedness"* and kurtosis do not go hand in hand. If there is negative excess kurtosis for a probability distribution, then from the concept of *"peakedness"*, we are supposed to conclude that the distribution is flat-topped but that is not the case always.

Let us go through some examples –

1. In a short note Kaplansky (1945) drew attention to four examples of distributions with different values of kurtosis, where behavior was not consistent with the interrelation between kurtosis and peakedness.

2. Westfall (2014) showed examples of some distributions where this well-known connection between *"peakedness"* and *"kurtosis"* was quite doubtful.

We consider here, two distributions to visualize this –

- Beta (0.5, 1)

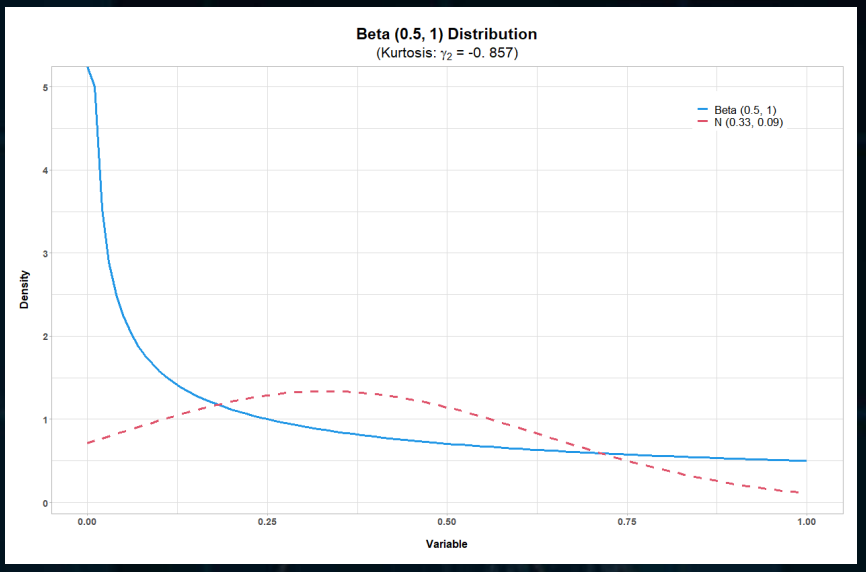


Figure 6: PDF of Beta (0.5, 1) Distribution

The distribution has kurtosis -0.857 (= γ_2) which is less than a normal distribution (γ_2= 0) and is therefore supposed to be "less peaked" than the normal but it is an infinitely peaked distribution.

- Mixture of Normal Distributions

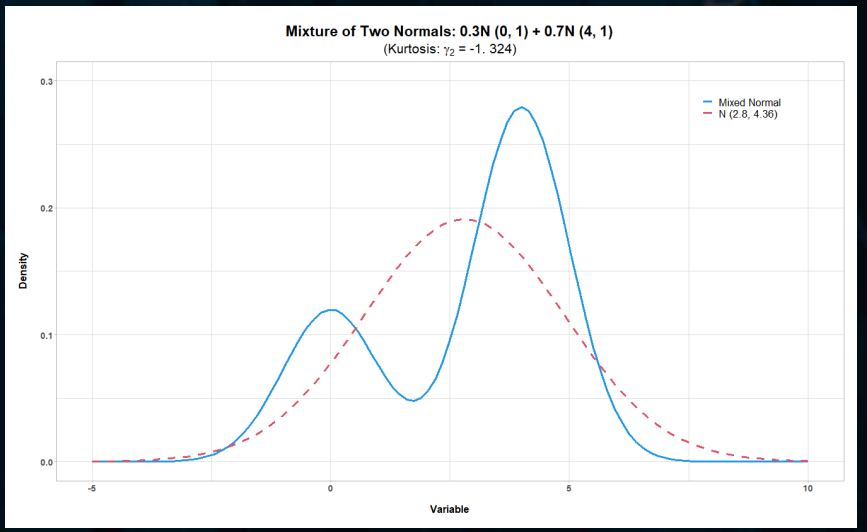


Figure 7: PDF of Two Normal Mixtures

The distribution has kurtosis -1.234 (= γ_2) which indicates it should be "less peaked" than the normal, but it has relatively higher peak than normal as observed from the diagram. Again, the distribution of mixture of two normal distributions has a lighter tail than the normal.

This supports the fact that Kurtosis tells nothing about the 'peak' of a distribution rather it deals with the tails and outliers of a distribution.

Does data also support the concept of "Tailedness"?

Now, we shall deal with some numerical examples and let's see what emerges from this –

Suppose the data values are - 5, 3, 3, 4, 5, 2, 5, 6, 4, 4, 1, 2, 2, 4, 3, 4, 3, 4, 6, 3.

The average of these values is 3.65 and g_2 = –0.596.

In this case there is no outlier and hence it gives small kurtosis.

Now, consider this data: 42, 2, 3, 1, 1, 1, 7, 1, 1, 0, 0, 3, 6, 2, 7, 722, 1, 11, 0, 1.

The average of this values is 40.6 and g_2 = 14.93.

In this case there are potential outliers and hence it gives high kurtosis.

Conclusion

In case where there are potential outliers [as in case (b)] there will be some extremely large z_i^4 values where, $\frac{(x_i - \bar{x})}{sd(x)}$ which, when averaged with all the other z_i^4 values, gives a high kurtosis. For the values which are close to the peak (near the mean), z_i^4 values are extremely small and contribute a little to the kurtosis.

Data near the middle do not contribute much to the kurtosis statistic i.e., kurtosis does not measure the "peakedness" rather it is simply a measure of detecting outlier in the sample and in case of distributions it measures the proneness of producing outliers compared to "mesokurtic distributions" (such as normal).

In particular cases where the data exhibit high kurtosis, it can be said that when we draw the histogram, the peak will occupy a narrow vertical strip of the graph. The reason is there will be a very small proportion of outliers which will spread over most of the horizontal axis, which will generate such a histogram that will have sharp peak, implies a high concentration toward the mean.

Lawrence T. DeCarlo also pointed out the errors persist in many textbooks, in his article, "...a number of textbooks, ranging from introductory to advanced graduate texts, describe positive kurtosis as indicating peakedness and light (rather than heavy) tails and negative kurtosis as indicating flatness and heavy (rather than light) tails. This is a serious error, because it leads to conclusions about the tails that are exactly the opposite of what they should be."

A perspective of this study: Financial Risk Management

As we have clearly come to know that kurtosis of a distribution reflects the presence of outliers, it helps a lot in financial risk. If we analyze the kurtosis of the *"distribution of returns",* we can get an idea about the risk of investing. In particular, if the kurtosis is high enough then the tail of the *"distribution of returns"* is thicker which indicates high risk for an investment since there is high chance of getting significantly high or small values of return.

If we can model the *"distribution of return"* by the distributions having lower kurtosis, then the risks become moderate. E.g. - *uniform, normal* distributions etc.

Otherwise, if we have to model it by a distribution having higher kurtosis, then the risk increases. E.g.– *Laplace* distribution, *Student's t* distribution (with lower degrees of freedom) etc.

Consider that we have a dataset of returns of ITC for the last 5 years. After standardizing it, it is modelled by a *normal* and a *t-distribution* with appropriate parameters.

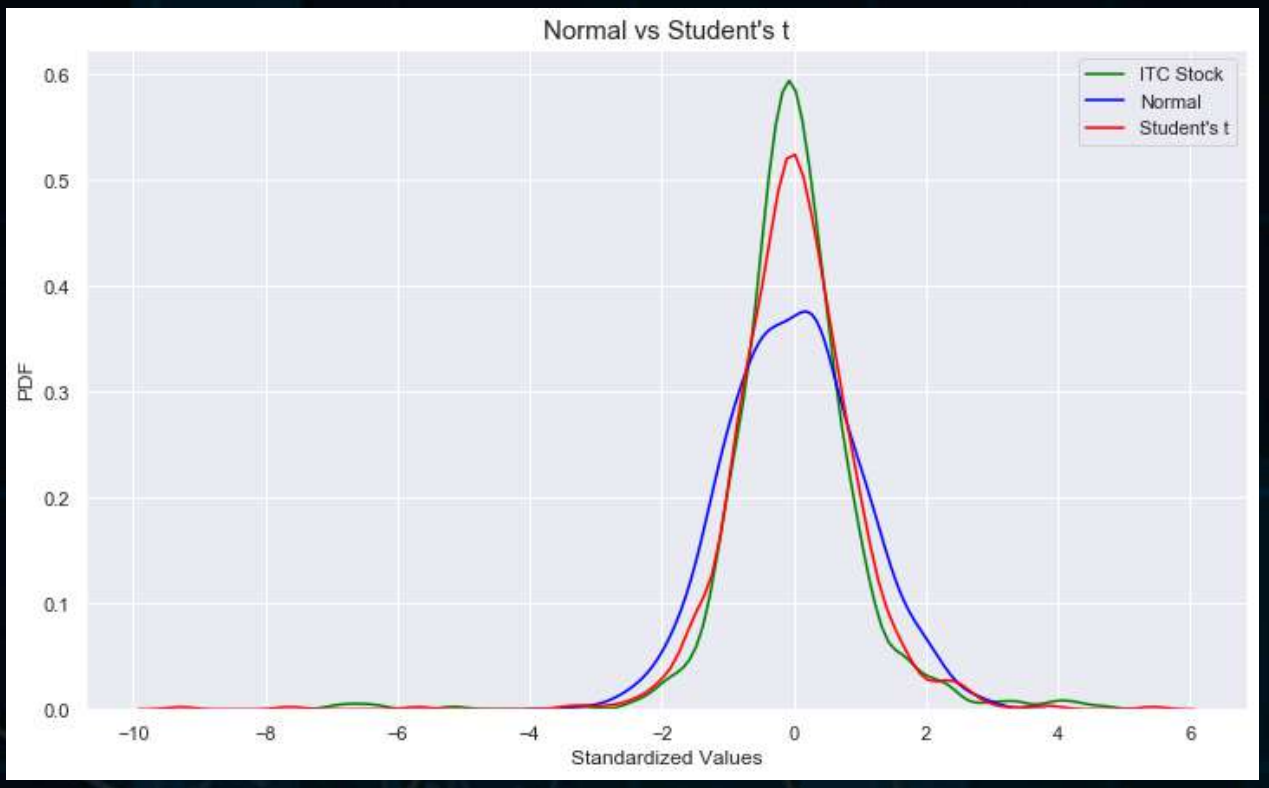
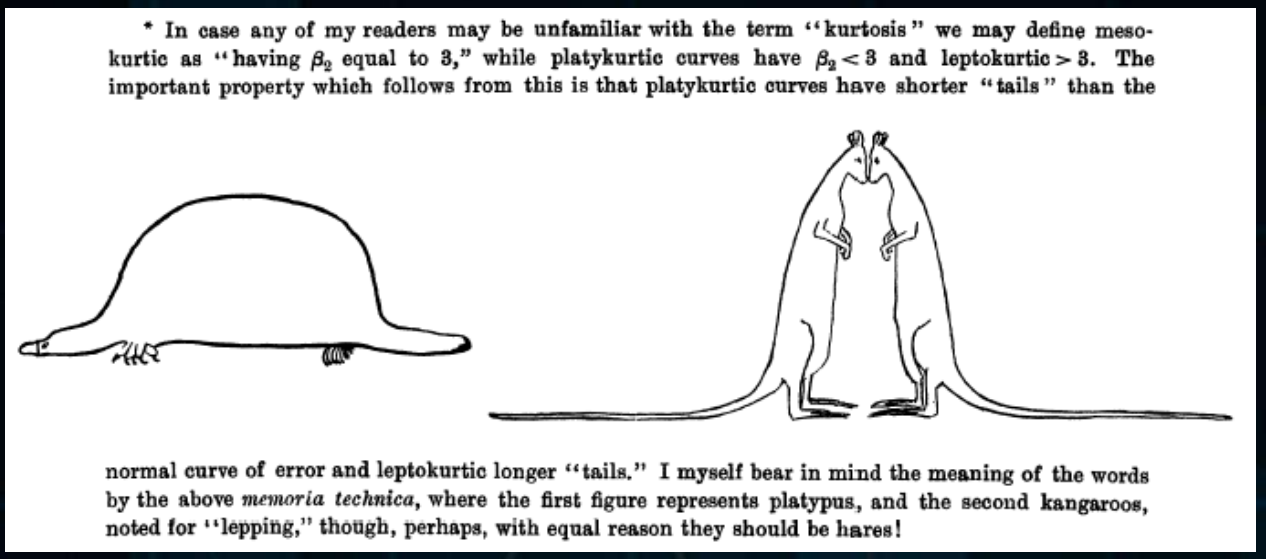


Figure 8: ITC Stock Returns

It is quite clear that the Student's t-distribution can model the data more accurately than normal distribution. The tail region of the distribution of return is captured well by t-distribution, the reason is that Student's t-distribution has thicker tail than that of a normal distribution for having a higher kurtosis. By modelling it by a normal distribution, we are basically underestimating the risk of extreme events. So, the distributions which reflects the risks more efficiently i.e., the leptokurtic ones are advantageous while calculating the risks.



* In case any of my readers may be unfamiliar with the term "kurtosis" we may define mesokurtic as "having β_2 equal to 3," while platykurtic curves have $\beta_2 < 3$ and leptokurtic > 3. The important property which follows from this is that platykurtic curves have shorter "tails" than the normal curve of error and leptokurtic longer "tails." I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for "lepping," though, perhaps, with equal reason they should be hares!

Figure 9: Kurtosis according to William Gosset [5]

References:

1. *Wikipedia: https://en.wikipedia.org/wiki/Kurtosis*
2. *"On the Meaning and Use of Kurtosis", Lawrence T. DeCarlo [1997]*
3. *"Kurtosis as Peakedness, 1905 – 2014. R.I.P.", Peter H. Westfall [2014]*
4. *"A Common Error concerning Kurtosis", Irving Kaplansky [1945]*
5. *"Errors of Routine Analysis", William S. Gosset (a.k.a. Student) [1927]*

FACULTY MEMBERS

Left to Right: Prof. Debjit Sengupta, Prof. Ayan Chandra, Prof. Durba Bhattacharya, Prof. Pallabi Ghosh, Prof. Madhura Dasgupta, Prof. Surabhi Dasgupta, Prof. Surupa Chakraborty.

OUR STUDENTS

THIRD YEAR

BATCH OF 2022

SECOND YEAR

BATCH OF 2023

FIRST YEAR

BATCH OF 2024

EPSILON DELTA 2022

Student Committee

Amrita Bhattacherjee
Convenor

Rajnandini Kar
Editor

Xavier Abhishek Rozario
Co-convenor

Abhinandan Bag
Associate Editor

Tithi Sharon Sarkar
Event Head

Srija Mukhopadhyay
Cultural Head

Mehuli Bhandari
Event Co-head

Shamie Dasgupta
Design Head