



St. Xavier's College (Autonomous), Kolkata

DEPARTMENT OF STATISTICS

PRAKARSHO

VOL - XIII

2021

P R A K A R S H O 2 0 2 1

ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA

DEPARTMENT OF STATISTICS

PRAKARSHO

2021

13TH EDITION

P R A K A R S H O 2 0 2 1

PRAKARSHO

2021

13TH EDITION



SCAN TO GET THE E-COPY

EMAIL: stsa@sxccal.edu
PHONE: 2255-1270

CONTENTS

Messages

- Message from the Principal 6
- Message from the Vice Principal 7
- Message from the Dean of Science 8
- Message from the Head of the Department 9
- Message from the Editor's Desk 10

Editorial Board 11

Departmental Report: 2020-2021 13

In Conversation with Prof. Asis Kumar

Chattopadhyay 18

From the Desk of a Senior Statistician: 27

- Statistics - Unparallel in Research and Development

-Prof. Manisha Pal

Column of Alumnus: 29

- What are in the names Data Science and Machine Learning? Statistics would be sweeter!

-Prof. Sourabh Bhattacharya

Articles

- The Unfair Game 38

-Sayantan Deb Barman

- A Case Study on Class Joining Pattern of Students During COVID-19 42

-Ishani Karmakar, Souhardya Mitra

- Data Science - As A Game Changer in Drone Engineering 58

-Ritoban Sen, Somjit Roy

CONTENTS

• <u>Can You Win by Losing?</u>	67
-Rajdeep Kundu, Sambit Das	
• <u>From Gambling to Probability: Modelling Chance Games</u>	71
-Somjit Roy	
• <u>On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis</u>	84
-Shabnam Dutta	
• <u>Randomness</u>	94
-Subharanjan Mandal	
• <u>Data Visualization : Glimpses of Today's World in the Light of Data</u>	96
-Adrija Saha, Shrayan Roy	
• <u>Determination of Gender in Dinosaurs: Statistics Concludes</u>	113
-Shantanu Nayek	
• <u>Statistics: Lies, Damned Lies?</u>	115
-Saptarshi Chowdhury, Utsya Chakraborty	
• <u>Misleading Statistics of COVID-19</u>	121
-Soham Chatterjee, Atreyee Roy	
• <u>Interpretation of Statistics in Different Fields of Science</u>	129
-Samapan Kar	
• <u>Lady With the Data</u>	134
-Abhay Ashok Kansal	
• <u>Strength lies in numbers, does wisdom?</u>	139
-Tathagata Banerjee	
<u>Our Professors</u>	143
<u>Our Students</u>	144
<u>Committee of the Epsilon Delta 2021</u>	145

Message from the PRINCIPAL

Rev. Dr. Domíníc Savío S.J.

Principál

St. Xavier's College (Autonomous), Kolkata

"It brings me a whole lot of jubilation, yet again, to learn that the Department of Statistics of our college is releasing its 13rd edition of the annual departmental magazine, 'PRAKARSHO'.

It is a matter of great appreciation that the department nurtures its tradition of excellence and distinction, which is manifested in the departmental magazine. Since its inception in 1996, the department has always endeavoured to propagate new skills and research abilities. The magazine has always provided an opportunity to the students to dive into the world of research apart from their curriculum.

My heartiest congratulations to the entire department, its faculty and its students. I wish them success in their efforts in publishing this issue of the magazine and many more subsequent ones in the years to come.

God bless you all! Nihil Ultra!"



PRINCIPAL

Message from the VICE PRINCIPAL

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

"Yet again, it is time to congratulate the entire department of statistics for bringing forward the latest edition of their departmental magazine, PRAKARSHO 2021.

The magazine is tangible proof of the department's stance in striving towards excellence in academic innovation and research. It bears testimony to the department's immense academic tenacity over all these years. This edition, like previous editions, brings the students a common podium to display their enthusiasm and passion for research, analysis and innovation.

My heartiest congratulations go out to the students, the faculty members and the editorial board members of the Department of Statistics for their conjoint effort towards making this year's edition a success."



VICE PRINCIPAL

Message from the DEAN OF SCIENCE

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

"I am immensely delighted to learn that the Department of Statistics is ready to publish the latest edition of the Departmental Magazine, PRAKARSHO 2021.

The magazine has always been a platform for the students to be creative and innovative in their field of study. It gives them a glimpse of the research world, when they explore various domains to write on different topics. These articles speak volumes about the interests and motivations of the students that the department inculcates in them. It is heart-warming to see that they go beyond their curriculum to make this happen.

I would like to congratulate the collaborative effort of the entire department that resulted in the fruition of the magazine and I wish them success for their future endeavours."



DEAN OF SCIENCE

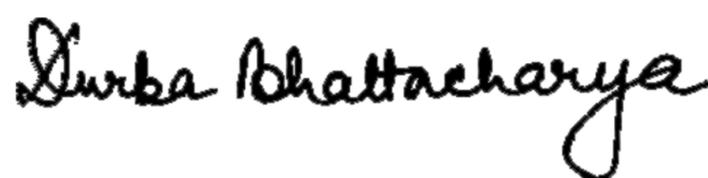
Message from the HEAD OF THE DEPARTMENT

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

"It is indeed a very satisfactory and proud moment for us to see our students successfully bring out the 13th edition of the Departmental Magazine, PRAKARSHO. Yet again, we have been able to reflect the essence of this department in the magazine, which would not have been possible without the relentless determination and untiring efforts of the students amidst the pandemic.

I would like to extend my heartfelt gratitude to Father Principal, Vice-Principal, Dean of Science and Dean of Arts for their perennial guidance and encouragement. Sincere thanks goes to the Programme and Publication Committee, for their support. I wish to applaud the Student Publication Committee for the hard work, enthusiasm and devotion with which they have overcome all the challenges to make this issue a reality.

My sincere thanks and appreciation go to my colleagues, whose dedication and efforts as a team has helped us come together to unveil yet another achievement of our department."



HEAD OF THE DEPARTMENT

“Statistics out of COVID-19” FROM EDITOR’S DESK

COVID-19 has put an intense spotlight on statistics. The numbers of daily and cumulative cases, deaths and infection rates are reported by governments and media daily all over the world and discussed by this very mankind. So, on one hand, apparently, Statistics have never enjoyed such recognition as a critic of favourable decision-making and providing public accountability. Yet official statistics enjoy a much more dichotomous existence, as on the other hand, we see around the world that statistics are ignored by public leaders and replaced with misinformation and rumours for various political and administrative causes. Today, in the midst of the data revolution, official statistics are competing with myriad sources of information, all claiming to be authoritative and to dictate the civilization. Science is at risk of being drowned out by a sea of irrelevance.

Amidst the global COVID-19 pandemic, the Department of Statistics, St. Xavier’s College (Autonomous), Kolkata has continued its work to cope up with the odds and come back with new challenges. The department has witnessed remarkable achievements by its own students even in this difficult time.

Hence, like every year we did not fail to publish yet another volume of our annual departmental magazine PRAKARSHO. Before you get driven away by the mesmerizing findings of statistics let us take the privilege to extend our heartfelt gratitude to the Patron, Advisory Committee, beloved and respected professors and the editorial team, without whom it was not possible to make this thought reality.

PRAKARSHO is not only an issue with articles, but also a weapon to express the excellence and revolution of minds of the students of our department. It fills up the blanks, fights against the odds and evil in the society and makes us one.

Hence, we take pride in presenting to you the 13th edition of PRAKARSHO.

NIHIL ULTRA!

Editorial BOARD

Patron

Rev. Dr. Dominic Savio S.J.
Principal

Advisory Board

Prof. Bertram Da Silva
Vice Principal

Dr. Tapati Dutta
Dean of Science

Argha Banerjee
Dean of Arts

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

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

Srijan Sen
Student Editor

Rajnandini Kar
Associate Student Editor

Editorial BOARD

Student Editorial Committee

Ritoban Sen
Somjit Roy
Amrita Bhattacharjee
Rajnandini Kar
Tathagata Pain
Mehuli Bhandari
Tithi Sarkar
Sweata Majumder

Srija Mukhopadhyay
Adrija Bhattacharya
Debolina Bhattacharya
Shrayan Roy
Adrija Saha
Pratyusha Mukherjee
Xavier Rozario
Abhinandan Bag

Student Designing Committee

Soham Biswas
Rajnandini Kar
Abhinandan Bag

Srijan Sen
Amrita Bhattacharjee
Xavier Rozario

Departmental Report: 2020-2021

Departmental Activities:

➤ Epsilon Delta held on 14th March, 2020

The Department conducted the 2020 version of its Annual program “Epsilon Delta” on March 14th, 2020. The program commenced with the launch of the 12th edition of the Departmental Magazine Prakarsho. Organized through the day were a Chess competition event, a science quiz “Inquizzit”, a theme quiz “Expellianswers”, paper presentations by students “Proectura” and a cultural program by the students of the Department.



Departmental Report: 2020-2021

Departmental Activities:

➤ Panel Discussion on 14th March, 2020

The Department organized its first Panel Discussion on March 14th, 2020. The panel comprised Mr. Prithwis Mukerjee, Director, Praxis Business School; Prof. Subhadip Basu, Professor, Department of Computer Science and Engineering, Jadavpur University, Kolkata; Mr. Somnath De, Executive Director, KPMG, India and Mr. Angshuman Bhattacharya, Founder and CEO, Sibia Analytics. The question before the house was "Is Data Science and AI the Future of Statistics?". The session was attended by more than 230 students and professionals from different colleges and institutes.



Departmental Report: 2020-2021

Departmental Activities:

➤ International Webinar on 13th August, 2021

The Department had organized an International Webinar, streamed on YouTube, on “Understanding Covid-19 Progression with Data Science” on 13th August, 2020. Prof. Arni S.R. Srinivasa Rao, Medical College of Georgia, St. Augusta, USA and Prof. Sourish Das from the Chennai Mathematical Institute, India were the speakers.



Departmental Report: 2020-2021

Students' Achievements:

Awards & Recognition (3rd Years):

1. **Dhruvi Mundra** - Runner's up at "National Level Debate Competition" organised by IIM - Ahmedabad, 25th October 2020.
2. **Ishita Pandey** - Best Athlete, Gold Medals in 400 m and 1500 m Flat Race, Discus Throw, Javelin Throw in Annual Sports 2020, St. Xavier's College, Kolkata, 02nd February 2020.
3. **Soham Majumder** - Winner from "Anti-Hero League" in XPL Quiz organized by SXCSC and Department of Sports, 16th June 2020.

Awards & Recognition (2nd Years):

1. **Saptarshi Chowdhury** -
 - a) Second Runner's Up in the event "X-Quiz (Sports Quiz)" organized by St. Xavier's College Student's Council with the Department of Sports, 18th June 2020.
 - b) Runner's Up in the event "Short Story Contest" organized by Booklustic Incorporated, 12th May 2020.
2. **Sandip Chakraborty** -
 - a) Won Gold Medal Table Tennis (Doubles) "Invictus TT Championship", JD Birla Institution, February 14th, 2020.
3. **Soham Ganguly** -
 - a) Gold Medal in Table Tennis (Doubles) in "Invictus TT Championship", JD Birla Institution, 14th February 2020.
 - b) Participated and completed "10 km Marathon organized by IDBI Federal Life Insurance", 02nd February 2020.

Departmental Report: 2020-2021

Awards & Recognition (1st Years):

1. **Tarushee Agarwal** - Chairperson at United Nations General Assembly at Sri Sri Model United Nations, 16th October – 18th October 2020.
2. **Yuvraj Dutta** - Secured the position of Special Mention in UNICEF Committee of Flagship 5.0 Gujarat Model United Nations, 29th January – 30th January 2021.
3. **Yenisi Das** - JBNSTS Bigyani Kanya Medha Britti Award, 2020, Jagadish Bose National Science Talent Search, 19th February 2021.
4. **Srinjoy Chaudhuri** - Campus Ambassador internship ELAN & NVISION, IIT, Hyderabad, 07th February 2021.

Placement Details (3rd Years):

1. **Ishani Karmakar** and **Shreya Singhee** were offered the post of “Associate Analyst” at Deloitte US India.
2. **Dhruvi Mundra** was offered the post as “Analyst at Management consulting research and Analytics team” at PricewaterhouseCoopers.

In Conversation With **PROF. ASIS KUMAR CHATTOPADHYAY**

Dr. Asis Kumar Chattopadhyay is the Pro Vice Chancellor for Academic Affairs of the University of Calcutta and the coordinator of IUCAA Resource Centre, Kolkata spending his career with a pool of extraordinary knowledge 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 Srijan Sen and Soham Biswas, on behalf of the editorial committee, on the 25th of February, 2021, at the office of Pro Vice Chancellor for Academics, University of Calcutta.

(The following is the written and abridged form of the interview, that has been prepared, in consultation with Dr. Asis Kumar Chattopadhyay.)



Interview

Srijan & Soham: Good Evening, Sir. Thank you so much for managing a time for us out of your busy schedule and given pandemic situation. We have come to have a forthright conversation with you regarding Statistics and its impact on your life, for the 13th edition of our departmental magazine Prakarsho 2021.

Dr. Asis Kumar Chattopadhyay: Thank you for coming and taking interest in having a talk with me. Please convey my sincere thanks and gratitude to all your teachers and friends in the Department for giving me this wonderful opportunity to speak my mind. Best wishes.

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

Q1. Sir, can you please tell us about your childhood, schooling, and early education?

Ans: I grew up in Dumdum and back then, the practice was to join a school which was in your locality so that it is easier to attend classes every day. So, I attended the Motijheel School, which was close to where I lived. Back then, it was a very well-reputed school with compassionate and dedicated teachers who took great care in teaching us. I really enjoyed my school life. At that time, the education system was divided into 11+3 years, that is, 11 years of schooling followed by 3 years of Undergraduate Study. The division of students into the fields of Science, Commerce and Humanities used to take place in the 9th standard. So, I bifurcated into the Science stream from Class 9. The following three years of high-school were probably the most important years of my life, because that's when I was exposed to the various topics of science. I started to develop an interest in Physics, Mathematics and Mechanics. After passing the Higher Secondary Examination, I joined the Presidency College as a student of Statistics (Hons).

Q2. How did you come to know about Statistics as a subject and what were the reasons for choosing it as your Major?

When I graduated from high-school, I had little to no exposure to Statistics as a subject. We had a paper on Core Mathematics in high school which contained some basic elements of Statistics, such as, the concepts of mean, median and mode. Apart from that, I had no knowledge about the subject or its prospects. In fact, when I went to take admission in the Presidency College, there was an option for applying for two subjects, and I had selected Physics as my first choice. But one of my school teachers was acquainted with a person, who was the Head of the Department of Statistics in Presidency at that time. This person, as I later came to know, was Atindra Mohan Gun, widely known as Prof. A.M. Gun. Now, my school teacher urged me to meet this professor and talk to him before making my final decision. So, I met him before the admission test and after talking to him, I changed my mind. In just a single conversation, he painted a picture of the subject so vividly that I decided to take it up as my Major. I sat for the entrance examination and fortunately cleared it, and hence, went on to pursue Statistics Honours in Presidency College.

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

Q3. How would you describe your graduation and post-graduation days?

The three years in Presidency was the golden part of my life, not only in terms of academics, but also in terms of the environment, the friends I've made, the teachers I learned from and the seniors who guided me. The faculty was exceptional. There is a very popular book titled 'Fundamentals of Statistics', written by Gun-Gupta-Dasgupta. All three of these writers were professors of Statistics in Presidency. However, just before I joined college, Prof. Gupta left Presidency and joined Kalyani University. But we had the pleasure of being taught by Prof. A.M.Gun and Prof. Bhagabat Dasgupta. Besides, there were other professors, viz., Prof. Biswanath Das and Prof. Shankar Ghosh, who were excellent teachers. So, that's where the foundation of Statistics was laid.

One part of the curriculum consisted of how to look at data and analyse it. The other part consisted of basic mathematics, analysis, probability theory and so on; and, this combination was very balanced. It ensured that we developed both theoretical and practical knowledge. In fact, the practical classes were very interesting. At that time, we did not have powerful computational tools. We just used some hand machines, known as facet machines, which was a whole new experience. So, I would say that the initial days forged my life in the field of statistics.

After graduating from Presidency College, I joined M.Sc. Statistics in the Ballygunge Science College. Here, the atmosphere was very different from that of Presidency. Back in college, the atmosphere was much more open and diverse. There were multiple events and activities happening all year round. However, when I went for my post-graduation, the environment became much more rigid in terms of student interaction and activities. This was a tough transition to go through. We had to attend classes starting from around 10:30-11:00 in the morning till 5:00-5:30 in the evening, at a stretch. On top of it, the course was very heavy. At that time, even 60% marks was something very difficult for us to score. This was the time when I was exposed to concepts of Advanced Statistics. The topics were extremely rigorous and precise. In fact, I still refer to those notes. Among our teachers, the senior-most professors were Prof. Shoutir Kishore Chatterjee and Prof. Shyamaprasad Mukherjee. In our first year in Masters, we had Prof. P.K. Bose, the first Pro-Vice-Chancellor for Academic Affairs of Calcutta University, as our teacher. After them, many new teachers joined. We had Prof. Rahul Mukherjee, Prof. Uttam Banerjee, Prof. Kalyan Das, Prof.

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

Samindranath Sengupta, Prof. Nripesh Kumar Mandal, and many others. So, I enjoyed my post-graduate days, but the course content was much more advanced and rigorous as compared to the undergraduate level.

Q4. Back then, did you feel the dearth of adequate textbooks or reference books as compared to what is available nowadays?

Of course! But it had its boons and curses. At that time, we did not have many books to refer to. As a result, we had to read several journals from the libraries which was tough but it helped us to know what was happening in the field of Research. Nowadays, everything is easily available. There are hundreds of books available on the internet. But browsing through the libraries and reading journals is a different experience. It helps you to know what's happening in the world around you, it gives you information regarding the relevant research work that has been done and facilitates students to learn and extract new ideas. Of course, looking through so many journals and advanced research papers was difficult, but the exposure eventually turned out to be beneficial. But of course, the lack of textbooks was deeply felt at that time.

Q5. How were the job prospects back then in the field of Statistics?

When I studied Statistics, it was not possible to assess the job market, since we did not have computers. After computers came into the picture, the entire job scenario changed. In my time, there were no software-based jobs. However, the job market was better for Statisticians (as compared to other general fields of study) in the sense that the Indian Statistical Service used to offer prestigious jobs which were coveted. In fact, out of, say 100 students from Kolkata, almost 80 of them aimed for the ISS. Another big sector was the teaching community. Institutions like the Statistical bureau and ISI hired Statisticians to teach aspiring students. So, job prospects were slightly better compared to other subjects. But the software market was not prevalent.

Q6. Why did you prefer teaching over the other available job profiles?

During that time, teaching was a very prestigious vocation. The best students, say, top 20-25% students used to go for research-oriented projects. This was a general convention among students of Statistics. These days, the best students opt for going abroad for research or jobs. But, back then, the best students continued to research in the

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

department or in ISI. They would then take up a job as a professor in some good University. According to me, becoming a professor seemed to be much more prestigious, compared to the other jobs that were available. So, that was one of the motivations for joining the teaching community.

Q7. Would you care to share a brief account of your teaching career and experiences?

After completing my Ph.D., I joined Calcutta University in 1987. So, it has been a long time since I started teaching. I have taught in many places in India and abroad. For a year, I was an associate professor in ISI as well. So, during my time as a teacher, I have met many different types of students which increased my interest in this subject even more. I realized that I had very little idea about many parts of statistics as a student; parts which I grew more proficient in, as I kept teaching. I strongly believe in the saying; teaching is the best way of learning. I feel, if you don't teach, it is very difficult to know the subject properly because interacting with so many students strengthen the fundamentals in a way that books can never do. So, in my opinion, for doing better research, one should always have some experience in teaching.

Q8. Being in an important administrative chair do you find it difficult to balance your academic and administrative work?

Frankly speaking, it is very tough. Nowadays, most of my time is dedicated towards maintaining the administration. Once you get into administration, it is very difficult to go back to the classroom and take classes on a regular basis. I try desperately to take out some time for my students. In fact, today I have a class at 7:30 PM, after this interview. So, balancing both sides is a very tough process for me. I sometimes regret the fact that I had to move away from the subject for tackling this job in administration.

Q9. Do you think there has been any significant difference in the teaching/learning process of the subject now as compared to how it was during your time?

Yes. The most important difference is the mode of teaching in terms of availability of gadgets. At our time, the only mode of teaching consisted of a blackboard, some chalks and a duster. Nowadays, different options are

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

available, starting from PowerPoint Presentations to other electronic modes of teaching like lecture videos. Evidently, the Internet helps a lot. However, I feel much more comfortable in teaching a class of students in person, writing the theories and concepts down on the blackboard as I go. It somehow ensures a certain presence of both the professor and the students. But when I use, say, a PowerPoint Presentation, I think it is difficult for the students to just look at a previously prepared slide and understand the entire topic. The online mode limits communication. So, for me, chalk and blackboard is a better way of teaching than PowerPoint presentations.

Q10. How much do you enjoy this online teaching procedure?

To be honest, I do not enjoy the online teaching processes at all. The students are not in front of me, I cannot see their faces and I cannot take questions from the class directly. It gets very difficult to communicate. So, after a certain point, I'm not sure whether my students are being able to follow whatever is being taught. But, given the current situation, we do not have a choice.

Q11. What are your views on the present CBCS curriculum?

The main advantage of the CBCS curriculum is the wideness and variety of choices that a student is offered. The main objective is to let a student choose his/her subjects according to their likings. I may choose, say, Literature and Statistics. But unfortunately, several institutions are still not being able to offer this range of choices due to various limitations. To facilitate this system, you need proper infrastructure in terms of the faculty, availability of classrooms and several other factors. But if we can do it properly, I feel that the CBCS curriculum is a very good initiative.

Q12. Having attended several national and international symposia and conferences, how would you rate our country in terms of its advancement in the field of Statistics?

During the 1950s - 1960s, India was dominating the world of Statistics. In fact, if you look at the statistics departments in the universities of US or Canada, you will find many Indians (and especially Bengalis) in the

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

important administrative positions. Because, at that time, systematic educational infrastructure for teaching Statistics was very limited all over the world. In India - and perhaps in Asia, if I remember correctly - statistics was first taught as a separate subject in the University of Calcutta. So, we had an upper hand. Gradually, many other universities started opening their own Departments of Statistics. Even then, I would say that Calcutta has a distinctive advantage in terms of faculty and guidance offered to students in statistics. ISI, for example, is an excellent institution. The only drawback in the field of academics today, is the fact that good students do not stay in academia after completing their higher studies. So, finding good professors is becoming increasingly difficult.

Q13. As you said, many students nowadays are going for corporate jobs after studying Statistics. Therefore, the number of good teachers is decreasing. How would you like to address this issue?

See, it totally depends on the particular person's choice, and we should not ask a student to choose a career they do not want to pursue. However, in my opinion, when you just complete your post-graduation, or any other degree, the student might not be matured enough to make a prudent decision. They might feel that the corporate life is attractive - most of the students are drawn towards a heavy pay check, and an affluent life abroad. But, when I talk to my old students, say 10 years down the line, they often tell me how difficult it gets to continue with a job in the corporate sector. This is mainly because, most of the jobs are not very challenging and the students, particularly the good ones, tend to lose interest after a certain point of time. The process makes them lose their originality, their knowledge and, having studied such a heavy subject, not being able to implement those theories in practice is very frustrating. So, until and unless you are working for a company which really drives you to use your knowledge, it is very difficult to obtain job satisfaction. However, it is true that these corporate sectors offer readily available jobs for students as soon as they graduate from college or university. So, this is beneficial for people who are in need of a job right after completing their education. On the other hand, teaching jobs are very limited and difficult to land. But, if you are patient and determined to become a teacher, if you are driven by the thought that you must contribute to the field of academics by imparting your knowledge and passing it on to the future generations, then you should definitely stick to your dream and become a professor. These students are the ones who uphold the legacy and passes their knowledge

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

on to the next generations to come.

Q14. How did you develop an interest in Astro-Statistics?

Well, my wife is from Astrophysics. She was an associate of IUCAA, Pune. This one time, a professor of Astrophysics from IUCAA visited Calcutta and met me. When I told him I study Statistics, he said that he was looking for a statistician to do some analysis of the data that they handle in astronomy. At that time, the data sciences were gradually coming into the picture. They had large amounts of data collected from the different satellites etc., which needed proper analysis. For this, they needed a statistician. So, I was the first Statistician to become an associate of IUCAA. I did many joint projects with my wife on astrophysics and have worked with other people from IUCAA as well. From there, I came to know about Prof. Jogesh Babu, who was in the Pennsylvania State University. Prof. Jogesh was from Astrostatistics and there was a centre on Astrostatistics in the USA (founded by Prof. Eric Feigelson, astrophysics), where he was working. So that's how I came into this field. With time, they organized several workshops and we visited many such places where people worked with astrostatistics. In fact, there is a centre in France as well.

Q15. After you became the Pro-VC of this university, how do you get time to associate with IUCAA Kolkata?

To be honest, I do not have any time at all. After being the Pro-VC, I have never but once visited the IUCAA Kolkata. However, I am trying to get back to it desperately; the COVID-19 situation has made it even more difficult.

Q16. What according to you are the job and research prospects in Statistics for the current generation of students?

Do you see that the direction is gradually changing? The field of mathematical statistics is now almost completely directed towards Data Science. On this context, I don't like the domination of computer scientists in Data Science. I believe, statisticians must be the leaders here, because most part of Data Science is about statistics. But we don't see any progress in that direction and I think there should be a change in the way things are taught. There must be a change in terms of the teaching methodology, in order to make the students get better acquainted with

In Conversation With PROF. ASIS KUMAR CHATTOPADHYAY

the concepts of Data Science and on how to pursue that subject, because data science and machine learning is the future. You will find, in the coming five to ten years, they will dominate the job market.

Q17. How did you use to handle large spectral data without the software that are easily available now?

Yes, we had a thing called the 'spectral analysis', which is an important part of data analysis. We started working on it using some large computers, which we got access to from some co-authors at the Grenoble University under IPAG. So, we used their computer facilities to analyse the spectral data. Another option in IUCAA was cluster computing, where they had some very good computers for dealing with big data. So, that is how we worked with spectral data.

Q18. Do you feel that a 'Brain -Drain' is rampant in the country? Is there anything that can be done and in what capacity to prevent it?

It started long back, and it will be there because for students of today, the market is all over the world. So, it is not wise to complain about the fact that the students are going abroad because, if they get good opportunities, then of course they will go. We must create a healthy and challenging working atmosphere, since that is the basic requirement for any job in the industry. They are looking for a comfortable and secure workspace. So, these are necessary steps to prevent the Brain Drain. There are no other ways or shortcuts about it.

Q19. Sir, in conclusion may we have a parting word of advice from you for our students and the younger generation of Statisticians?

I think you all are much more well-versed in terms of what you want to achieve. So, what I would like to say is that, choose your teachers wisely, from whom you can gain considerable knowledge and whom you can follow. Try to extract as much knowledge from them as possible, so that you have a strong fundamental concept of your subject. Try to look at the society from a statistical angle, how one can apply statistics in the society, what are the basic needs in the world and so on. Until and unless you apply your concepts to solve real-world problems, you will not understand a subject to its entirety. I know that theory and mathematics is important for your subject, but you should always find some practical application of what you learn and then try to project it onto real-world problems.

From the Desk of a SENIOR STATISTICIAN

Statistics - Unparallel in Research and Development

Prof. Manisha Pal,

Department of Statistics, University of Calcutta

Statistics has been neglected by students- at least at the very beginning - as 'more mathematics'. They failed to look at its prowess in developing quantitative reasoning skills, devising tools to make inferences, assessing limitations, and in detecting errors and uncertainty in data. However, with the "happy marriage of Statistics and Computer Science", it has paved its way into the good books of all. The two have almost merged together, as the practice of Statistics has moved onto our electronic devices in the form of programming. The use of languages like R and Python has grown enormously over the past few years, especially in the academic circle. Moreover, proficiency in R or Python is highly desired by many employers.

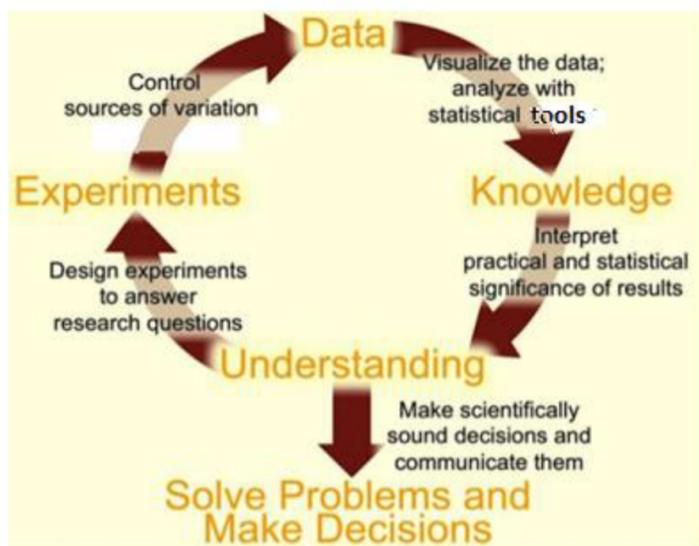


When one thinks of embracing Statistics, the thought that comes foremost to mind is what can he/she gain by studying the subject? It is very important to note that, apart from its theoretical development, Statistics is highly sought in research and development in almost all areas. In order to bring in new concepts, look for improvement in the old ones, it is essential to carry out experiments leading to data in the form of outcomes. Statistics now comes into the picture to interpret the data, analyze it, look into the level of uncertainty, predict the chance of success, ... and the list goes on. Be it in academia, information technology, medicine, economics, business, or construction, statistical research has now become a compulsion to progress forward.

Why do we need to pair scientific research with Statistics? This is because Statistics helps from the very nascent stage of a research project. It helps in identifying the variables to be included in the study, in deciding on the size of sample to be taken, type of data to be collected and the method to be adopted for collecting data. The data collected is, more often than not, "messy, ugly and incomplete". This is where Statistics saves the day! It helps to scrutinize it, clean it and come up

From the Desk of a SENIOR STATISTICIAN

with salient features using inferential tools like hypothesis testing and estimation, and interprets the results. Diagrammatically, the role of statistics in research and development is as follows:



Carrying out a project without statistical help has many pitfalls, which can lead to unconvincing and incorrect conclusions. Even at this stage, a statistician can help by identifying the flaws in the experiment. As the great Statistician Dr. Ronald Fisher said... *“To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.”*

All said and done, Statistics is not foolproof. It cannot solve all problems under the sun. It depends heavily on modeling and inferential procedures, which have their limitations owing to design and formulation. One needs to choose the statistical tools with great care in a given situation, or else the conclusions drawn may be highly erroneous. Choosing the correct tool for a given experiment must be based on expertise in the subject and in the matter under study. Like all other tools, Statistics may be misused either deliberately or by researchers who have little knowledge about statistical concepts and procedures. Being indispensable in research and development, Statistics needs to have the “handle with care” tag when trying to match the best techniques in a specific research study.

Column of ALUMNUS

What are in the names Data Science and Machine Learning? Statistics would be sweeter!

*Prof. Sourabh Bhattacharya,
Interdisciplinary Statistical Research Unit, ISI, Kolkata*

1 Introduction

To quote the Bard: “What’s in a name? That which we call a rose by any other name would smell as sweet.” The modern era is captivated by the sweet scent of scientific success of data science and machine learning – apparently buzzwords that travel hand-in-hand in an inseparable Romeo-Juliet way.

Indeed, the craze for machine learning and data science has engulfed the modern world, so much so that almost every college, university, institution across the globe is emphasizing courses and research initiatives surrounding these new avatars, not to speak of the industries and business organizations, which have already given in to the enchanting powers of these buzzwords. But somewhat disconcertingly, the real essence of machine learning and data science still seems to be obscure to the most. According to some, machine learning is a computer (machine)-oriented, specialized way of analyzing data, to some others it is a hotch-potch of adhoc algorithms for data analysis. Again, to some others, it is just sub-discipline of statistics. “Up-down, up-down”, is also another name for machine learning given by students of a reputed institution, following a superhit Bollywood movie clip! Data science is plainly referred to as science of data by many, remaining vague about the explicit meaning or if statistics if being referred to. But when it is reminded that data science is also concerned with big data, then most think it is related to some advanced technology. It is also perceived by many that data science and machine learning are the same. Such has been the level of enchantment and vagueness of the buzzwords that often business organizations hire data science and machine learning without even knowing what problem to solve!

In this article, our goal is two-fold. We first attempt to dispel some of the vagueness that emerged in the above discussion. Then we uphold statistics, specifically, the Bayesian paradigm, as having the genuine promise for influencing the modern and future scientific world, compared to either data science or machine learning. We begin with data science, arguing



Column of ALUMNUS

that this is indeed no more than a mere buzzword in essence.

2 What is data science?

2.1 A new paradigm, or just a new name?

Data science refers to design, collection and analysis of data that are typically big (at least one terabyte). This description itself shows that data science is no different from statistics. The “big data” scenario is certainly not a new paradigm qualifier for data science which only requires enhanced computer resources. New paradigms must be distinguished by new philosophies, such as frequentist and Bayesian statistical paradigms. Hence, not surprisingly, most statisticians are of the opinion that data science is only a new name for statistics. Interestingly, for the first time in 1985, Jeff Wu, in a lecture given to the Chinese Academy of Sciences in Beijing, suggested that statistics be renamed data science, and in 1997, he again suggested the same, reasoning that the new name would more appropriately indicate that statistics is more than just accounting or data description. Many others are in the absolute delusion that statistics is about quantitative data only while data science deals with qualitative data (for example, images, emails, social media posts, audio, video), as well.

To us, however, statistics has always been far more than what Jeff Wu or the others might have perceived, and does not need a new name.

2.2 Big data

The big data scenario of data science is challenging in the sense that traditional methods of capturing, storing, transfer, visualization and analyzing huge amount of data are rendered inadequate. Examples of such big data are social networks, web logs, internet search indexing, sensor data, call detail records, military surveillance, as well as complex data in astronomy, biogeochemistry, genomics and environmetrics. Technological revolution of computer hardware in the forms of fast processors, cheap memory, massively parallel processing architectures, and the specialized soft wares such as Hadoop and MapReduce, have made collection, storage and empirical analyses of big data feasible.

To provide a glimpse of some major successes of big data analysis even in an area usually considered far from science and technology (even education!), note that Barack Obama’s successful 2012 re election campaign and the success of BJP in winning the Indian General Election in

Column of ALUMNUS

2014, owe much to big data analysis. Hence, it is easy to anticipate the impact of big data analysis in general.

2.3 Statistical approaches for big data

However, realistic statistical modeling and analysis in the big data situations are hitherto unexplored. The reason is that, serious modeling of the underlying complex data-generating phenomenon requires realistic incorporation of dependence structure in the model in some appropriate stochastic process framework, and the dependence structure often becomes responsible for rendering the computational complexity infeasibly high in big data situations, particularly when matrix inversions are involved. Basis function expansions, in conjunction with parallel computing skills, may be invaluable to avoid such computational infeasibility.

There is also a subtle question regarding the requirement of big data in genuine statistical analysis. A well-collected sample is necessary, along with a sound understanding of the data-generating phenomenon for modeling. In the Bayesian paradigm, a well-chosen prior also obviates much of the big data requirement, and this author's experience has been to outperform big data methods in astronomy using judicious, sophisticated Bayesian analysis with small data.

The field of machine learning is more interesting than the buzzword data science, as it can be directly related to artificial intelligence (AI). Indeed, the machine learning story is deeply ingrained in that of AI. Moreover, the essence of machine learning lies in the deterministic algorithms, so that unlike data science, machine learning may be considered a paradigm with respect to determinism. To proceed towards machine learning, we begin with a brief overview of AI.³ A simplistic briefing of AI

The field of AI may be considered as the investigation and creation of systems that have perception of its environment and can act in a way to maximize its chance of success. Such a system may be considered intelligent. John McCarthy, often considered the father of AI, and the one who coined the term in 1955, defined it as "the science and engineering of making intelligent machines". Ultimately, such intelligent machine must at least be able to emulate human capabilities. It has been argued that it is possible to copy the human brain directly into hardware and software, so that a simulated brain need not be any different from the original. This of course requires extremely precise description of the human brain. Not everyone believes that this is possible and various counter arguments in many different forms exist. Gödel's proof of his first incomplete theorem

Column of ALUMNUS

demonstrates that it is always possible to create statements that a formal system can not prove, while the same does not apply to human beings. On the other hand, there are many around the globe who are even apprehensive of creation of malevolent AI that may spell destruction of the human race (remember the Terminator movie series?). Stephen Hawking was also one of the proponents of such an omen.

Setting aside further debate whether or not true AI is achievable, let us look at the current state-of-the-art of AI. As is usual for large scientific endeavors, AI is split into sub-problems, each associated with traits the intelligent system must be endowed with. We shall take a peek into some of such important sub-problems, with the hope to kindle interest in the approaches to solving such sub-problems, which we discuss subsequently.

3.1 Reasoning and problem solving

It is desirable of AI to imitate human beings in solving puzzles, making logical deductions and to prove mathematical theorems. Early AI research developed logic-based algorithms that proceed in a step-by-step basis and have been useful in demonstrating some of these abilities. However, such algorithms require enormous computational resources to solve difficult problems, which necessitated further research in this area.

3.2 Learning

In robotics, learning is considered to be skill acquisition by a robot through interaction with human teachers, autonomous self-exploration, using guidance mechanisms such as imitation and active learning via participation and interaction with the learning process.

3.3 Natural language processing

The ability of machines to read and understand human-spoken languages is provided by natural language processing. Common applications of natural language processing are Google Translate, Microsoft Word and Grammarly that check grammatical correctness, Interactive Voice Response (IVR) in call centers, personal assistant applications such as OK Google, Siri, Cortana and Alexa.

3.4 Perception

Perception of AI is its ability to deduce aspects of its environment. In practice, input from sensors (such as cameras, microphones, tactile sensors, sonar, etc.) can be processed to this end. For example, visual input

Column of ALUMNUS

can be analyzed by computer vision, while speech, facial and object recognition are other perceptions desired of AI.

4 Embodiment of machine learning in AI

4.1 Supervised and unsupervised learning

Machine learning can be described as the study of computer algorithms that automatically improve as their experience increases. As can be easily anticipated from the goals of AI, machine learning plays the key role in AI research, supervised and unsupervised learning constituting the important divisions associated with labelled and unlabelled data, respectively. Important examples of supervised learning comprises classification and regression, while clustering and density estimation are important examples of unsupervised learning.

4.2 Classifiers and controllers

In AI, the simplest applications are constituted by classifiers and controllers, where the latter also use classifying conditions before taking actions. In other words, classifiers are central to many AI applications. Broadly, classifiers are nothing but functions that determine the best class for the current observation or pattern, based on previous experience, that is, training based on the remaining observations. The best class can be perceived as the best decision.

Among very many classifiers, the most popular ones used in AI are the neural network, support vector machine, k-nearest neighbor algorithm, Gaussian mixture model, naive Bayes classifier and the decision tree. Importantly, the “no free lunch” theorem formalizes that no single classifier can be expected to outperform all others on all problems.

4.3 Neural networks

The human brain has been the object of curiosity since time immemorial. But only in 1943, research on neural networks was initiated by Warren McCulloch, a neurophysiologist, and Walter Pitts, a mathematician, who used electrical circuits to model a simplistic form of the neural network. Donald Hebb then demonstrated in 1949 that neural pathways gained strength after every use. The first effort to simulate a neural network was due to Nathaniel Rochester from the IBM research laboratories, in 1950, which later provided the impetus to intense research on AI and neural

Column of ALUMNUS

networks. The oldest neural network, namely, the perceptron, surprisingly grew out of research on the eye of a fly, conducted by Frank Rosenblatt, a neuro-biologist. Development of the backpropagation algorithm for fitting neural networks by Paul Werbos was also instrumental in popularizing research in this area.

Neural networks can be classified into feedforward and recurrent neural networks. The former refers to forward-directional signal passing, while the latter allows feedback. Popular feedforward neural networks are perceptrons, multi-layer perceptrons and radial basis networks, and the Hopfield net is a popular example of the recurrent neural network.

The single-layered feedforward neural network can be extended to multiple-layered networks, and this extension is often referred to as “deep learning”. For fitting such multiple-layered networks, each layer may be pre-trained using an unsupervised restricted Boltzmann machine, followed by fine-tuning using supervised backpropagation.

The problem of intelligent control (for robotics) make extensive use of neural networks method ologies.

4.4 Applications of machine learning to AI

4.4.1 Application to reasoning and problem solving

Human beings have the ability to solve problems by intuitive judgments and do not often need to proceed step-by-step. Although previous AI approaches used the step-by-step algorithmic approach, it was inefficient and required vast computational resources. In the modern times it has been replaced with neural net simulation to emulate human brain structures believed to have such intuitive skills.

4.4.2 Application to language processing

Natural language processing for AI make extensive use of support vector machines, Bayesian net works, neural networks and deep learning, among supervised learning methods, while clustering plays the most important role among various unsupervised methods.

4.4.3 Application to learning

Almost all machine learning algorithms are generally applied to the learning problem of AI.

4.4.4 Application to perception

Column of ALUMNUS

AI perception makes heavy use of deep learning tools for speech, face and object recognition.

5 Triumph of statistical methods in AI

When it comes to dealing with incomplete or uncertain information, deterministic machine learning algorithmic tools are inadequate for AI. Such issues that arise in reasoning, learning and perception problems of AI, require quantification of uncertainty. In such situations, probabilistic and statistical approaches are employed, setting aside the determinism of the traditional machine learning algorithms. Since quantification of uncertainty is achieved most coherently in the Bayesian paradigm, it is not surprising that the Bayesian approach, specifically, Bayesian networks and dynamic Bayesian networks, are considered indispensable in the aforementioned situations. Various traditional statistical methods are also used for filtering, prediction, smoothing required for perception systems to analyze time-dependent processes such as hidden Markov models.

To aid the agent in making choices and plans, decision theory, Markov decision processes, dynamic decision networks, game theory and mechanism designs are incorporated in AI.

Indeed, incorporation of the statistical approaches in AI design led to unprecedented success in the 1990s so much so that such a statistical movement has been referred to as truly scientific, "nothing less than a revolution" and the "victory of the neats". Yet there is an ongoing debate regarding usefulness of statistical approaches in AI, particularly with respect to language processing. The interested reader is referred to <https://sites.tufts.edu/models/files/2019/04/Norvig.pdf> for details.

6 Foresight and Bayesian AI

In the context of AI it is important to point out that human beings are distinguished from the other beings, not only by taking the best decisions in the face of the present uncertainties, but also by their ability to foresee the future. For instance, human beings are expected to foresee that the current global warming trend is only temporary and poses no threat to the future world. However, other beings would attach too much importance to the current trend and infer continuation of global warming till the earth melts!

The current trend in AI is to use machine learning methods to predict the

Column of ALUMNUS

future. However, none of the existing deterministic methods including machine learning, can create any system having both these qualities. Even with a huge amount of data, the future can not be captured by determinism. Only appropriate Bayesian analyses using stochastic processes, can provide hope in this regard. [1] demonstrated that machine learning methods are convincingly outperformed by classical statistical time series methods, in both one-step ahead and multi-step ahead forecasts. The Bayesian forecasts are naturally expected to outperform both, because of the highly sophisticated and coherent uncertainty quantification ability of this paradigm.

As much as future global warming is concerned, the deterministic approaches of the global climate models have produced future predictions that are in keeping with the current global warming trend, and are responsible for creating great consternation among people, and (generally nonsensical and biased) policies by the International Panel on Climate Change to control future global warming. On the other hand, our Bayesian approaches ([2], [3]) realistically forecast a normal future world, where no doom is spelt by global warming.

7 Conclusion

We began this article by comparing data science and machine learning to Romeo-Juliet. Indeed, machine learning dwells in the heart of data science – machine learning libraries are central to systems like Hadoop, built for data science, and has very many applications for exploratory data analysis. Machine learning is also the better half as it may deserve to qualify for the paradigm of deterministic algorithms and its wide applicability in AI.

Data science and machine learning have together conquered the world with their sweet scent of success. The common person, however educated, lauds their success, almost completely unaware, that statistics has been far more successful than machine learning in AI and holds more promise for the future, as we pointed out. According to the Bard, the rose, by any other name, would smell just as sweet. Statistics and data science differ only by names, yet data science smells sweeter. Alas! The Bard was wrong!

We sign off with an earnest prayer to the future in the same vein as our very own Bard – *“dao phirey se statistics, loho e buzzwords”*.

Acknowledgment

We acknowledge various internet sources, and the book [4], for information required for this article.

Column of ALUMNUS

References

- [1] S. Makridakis, E. Spiliotis, V. Assimakopoulos, *Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward*, PLoS One 13 (2018) 1–26.
- [2] D. Chatterjee, S. Bhattacharya, *How Ominous is the Future Global Warming Premonition?*, arXiv preprint (2020).
- [3] S. Roy, S. Bhattacharya, *Bayesian Appraisal of Random Series Convergence with Application to Climate Change*, arXiv preprint (2020).
- [4] *Introduction to Machine Learning: The Wikipedia Guide*, available at https://www.academia.edu/41157657/Introduction_to_Machine_Learning_The_Wikipedia_Guide.

Articles from STUDENTS

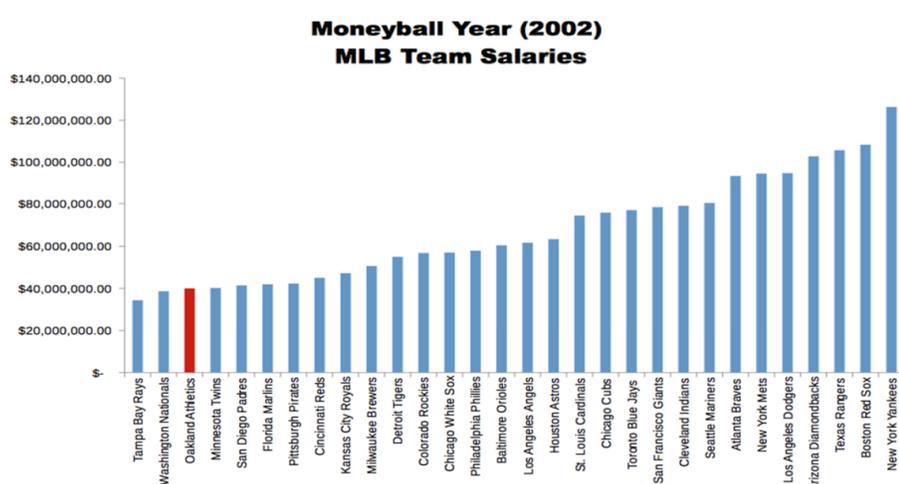
The Unfair Game

Sayantana Deb Barman (3rd Year)

Three of your top players leave the team as free agents; a 41 million dollar team teams up against teams with budgets as high as 125 million dollars. That's the Unfair Game.

And that was what Billy Beane and Paul DePodesta were up against. So did Billy Beane, Oakland Athletics general manager team up with Paul DePodesta, a Yale graduate economist to boost the crippling Athletics to the top of the American League Division Series? Nah, real life is seldom like that. This isn't another David-beat-Goliath story. The Athletics lost, but they created history. They became the first team to go on a 20 match win streak in 100 years.

The 2002 team salaries looked like this:



So how did they do it with the small budget? They used something called Sabermetrics- a statistical method to radically revolutionise the old scouting techniques. In baseball, the scouts reach out to players and access them based on attributes such as strength, generating bat speed,

head stays on ball, lack of fear, etc. But Sabermetrics challenged this methodology.

The main attributes DePodesta took into account were:

1. RS — Runs Scored
2. RA — Runs Allowed
3. W — Wins
4. OBP — On Base Percentage
5. SLG — Slugging Percentage
6. BA — Batting Average
7. Playoffs — Whether a team made it to playoffs or not
8. OOBP — Opponent's On Base Percentage
9. OSLG — Opponent's Slugging Percentage

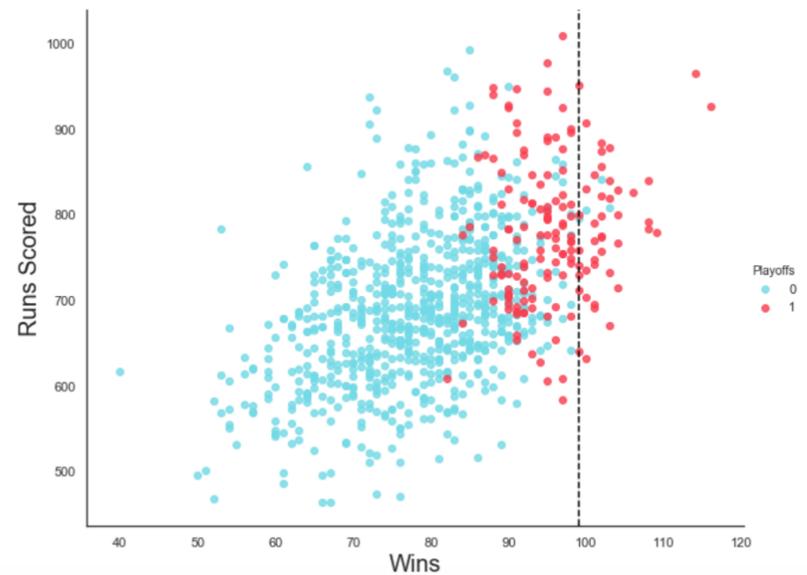
The Unfair Game

Now previously Batting Average was considered to be the most deciding factor and DePodesta realised that scouts often overlooked the other attributes like OBP and SLG. Hence Oakland Athletics was able to secure players with high OBP and SLG at steal prices.

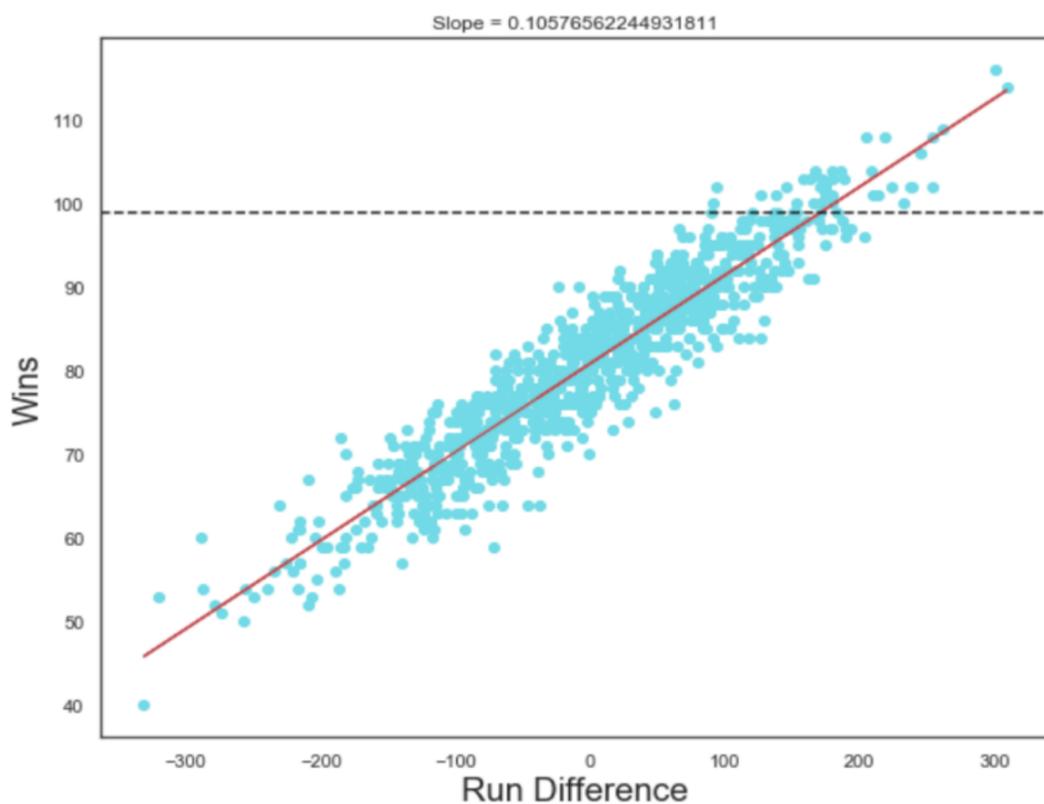
	Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOPB	OSLG
0	ARI	NL	2012	734	688	81	0.328	0.418	0.259	0	NaN	NaN	162	0.317	0.415
1	ATL	NL	2012	700	600	94	0.320	0.389	0.247	1	4.0	5.0	162	0.306	0.378
2	BAL	AL	2012	712	705	93	0.311	0.417	0.247	1	5.0	4.0	162	0.315	0.403
3	BOS	AL	2012	734	806	69	0.315	0.415	0.260	0	NaN	NaN	162	0.331	0.428
4	CHC	NL	2012	613	759	61	0.302	0.378	0.240	0	NaN	NaN	162	0.335	0.424

This is just a portion of the team data and the original dataset can be found on Kaggle.

Here we see that only 3 teams that have won 99+ games have not been able to make it to the playoffs. So DePodesta calculated that to win 99+ games they will have to score 814 runs, conceding 645, allowing a run difference of 169.



Let us visualise the importance of Run Difference (RS-RA) on Wins using a scatterplot.



Thus we see a positive linear association between Wins and Runs Difference. The correlation coefficient between them is 0.938515, indicating a very strong relationship.

The Unfair Game

Amongst the other attributes, the correlation matrix between them is:

	OBP	SLG	BA	RS
OBP	1	0.806154	0.854055	0.904909
SLG	0.806154	1	0.814068	0.926384
BA	0.854055	0.814068	1	0.831625
RS	0.904909	0.926384	0.831625	1

Thus we see that Batting Average is least correlated to the very important factor, Runs Scored. The scouts had overvalued Batting Average and undervalued the more important attributes of SLG and OBP.

Now we try to fit a Runs Allowed and Runs Scored model using linear regression on these attributes:

$$RS = -804.627 + (2737.768 \times (OBP)) + (1584.909 \times (SLG))$$

$$RA = -775.162 + (3225.004 \times (OBP)) + (1106.504 \times (SLG))$$

Subsequently, our Wins model takes the form of:

$$W = 84.092 + (0.085 \times (RD))$$

Now we plug in the pre-season values of these attributes of the players who made the Oakland Athletics team of 2002 in the model.

(OBP: 0.339; SLG: 0.430; OOBP: 0.307; OSLG: 0.373)

We compare our predictions with the actual results of 2002.

Attribute	Model Estimate	Actual
Runs Scored	805	800
Runs Allowed	628	653
Wins	99	103

Despite the limitations, Sabermetry (or as it is popularly called, Moneyballing) revolutionized not only baseball but made its way into other sports too. Leicester City used Moneyballing to find N'Golo Kanté, a relatively unknown player who played for newly promoted Caen in France. The underdogs made their way to their maiden EPL win. The odds against them were an astronomical 5000/1.

The Unfair Game

Liverpool FC and its analytical team scooped up the likes of Andy Robertson for 8 million euros only. They boosted their way to a UCL win and an EPL win the next season (their first in 30 years).

So data science has already revolutionised baseball and made its way into football. So which sport is next? Nothing seems off-limits here.

Articles from STUDENTS

A Case Study on Class Joining Pattern of Students During COVID-19

Authors:

Ishani Karmakar, STSA (3rd year)

Souhardya Mitra, STSA (3rd year)

Data collected by:

Xavier Abhishek Rozario, STSA (1st year)

Shantanu Nayek, STSA (2nd year)

Souhardya Mitra, STSA (3rd year)

Ishani Karmakar, STSA (3rd year)

Nirnisha Pramanik, MCBA (3rd year)

Madhumita Choudhury, CMSA (3rd year)

INTRODUCTION:

2020 started as was expected. We were back into the loop of alarm – running to college – classes – exhausted – dozing off – alarm. The CORONA VIRUS had started to spread around the world and soon was declared a pandemic by WHO in March 2020.

On March 14, our college declared a two-week holiday owing to the ongoing CORONA VIRUS pandemic. We were quite excited to have a break from the loop of the monotonous day and also, we would get some time for our upcoming end-semester examination.

This two-week holiday has now been extended to almost a year. Physical classes had to be replaced by online classes.

We have collected data on the number of students who have joined 5 minutes before the start of the first class of the day and the first class of the second half (i.e. post-lunch break) for the three years of our department (STATISTICS) and the third years of two other departments, viz. MICROBIOLOGY and COMPUTER SCIENCE for the third week of February 2021.

A Case Study on Class Joining Pattern Of Students During COVID-19

In our article, we have worked on the following:

- Comparison between the daily average number of students joining 5 minutes before the 1st class of the day and 5 minutes before the 1st class of second half (i.e. post-lunch break)
 - Students of 3rd year of the Statistics department
 - Students of the Statistics Department (taking all the three years together)
 - Students of 3rd year of the three departments considered in our experiment (viz. Statistics, Microbiology and Computer Science)
- Test for homogeneity over the three years of Statistics department
 - Joining time for the 1st class of the day
 - Joining time for the 1st class of 2nd half (i.e. post-lunch break)
- Test for homogeneity over three departments (viz. Statistics, Microbiology and Computer Science)
 - Joining time for the 1st class of the day
 - Joining time for the 1st class of 2nd half (i.e. post-lunch break)
- Test for the independence of year and joining time in the online class
 - Joining time for the 1st class of the day
 - Joining time for the 1st class of 2nd half (i.e. post-lunch break)
- Test for the independence of department and joining time in the online class
 - Joining time for the 1st class of the day
 - Joining time for the 1st class of 2nd half (i.e. post-lunch break)

TESTING PROBLEMS:

- 1.1 **Comparison between the daily average number of students of 3rd year of the Statistics department joining 5 minutes before the 1st class of the day and 5 minutes before the 1st class of second half (i.e. post-lunch break):**

A Case Study on Class Joining Pattern Of Students During COVID-19

Let X be the random variable denoting the number of students who have joined 5 minutes before the 1st class of the day and let Y be that for the 1st class of 2nd half (i.e. post-lunch break).

Let us assume that $X \sim \text{Poisson}(\lambda_1)$ independent of $Y \sim \text{Poisson}(\lambda_2)$.

To test, $H_0 : \lambda_1 = \lambda_2$ against $H_1 : \lambda_1 \neq \lambda_2$.

Let random samples of sizes n_1 and n_2 be drawn respectively from the distributions of X and Y , independent of each other.

We have,

$X_i \sim \text{Poisson}(\lambda_1), i = 1 (1) n_1$, independently of
 $Y_i \sim \text{Poisson}(\lambda_2), i = 1 (1) n_2$.

Define, $\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$ and $\bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$

By Variance Stabilizing Transformation (Square Root Transformation of Poisson mean), we have,

$\sqrt{n_1} \left\{ \sqrt{\bar{X}} - \sqrt{\lambda_1} \right\} \xrightarrow{d} N\left(0, \frac{1}{4}\right)$, independently of
 $\sqrt{n_2} \left\{ \sqrt{\bar{Y}} - \sqrt{\lambda_2} \right\} \xrightarrow{d} N\left(0, \frac{1}{4}\right)$, for moderately large n_1, n_2

Test statistic: Define $T = \sqrt{\bar{X}} - \sqrt{\bar{Y}}$

$$E(T) = \sqrt{\lambda_1} - \sqrt{\lambda_2}$$

$$V(T) = \frac{1}{4n_1} + \frac{1}{4n_2}$$

$$\therefore \frac{(\sqrt{\bar{X}} - \sqrt{\bar{Y}}) - (\sqrt{\lambda_1} - \sqrt{\lambda_2})}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)} \xrightarrow{d} N(0, 1), \text{ for moderately large } n_1, n_2.$$

$$\text{Under } H_0, Z = \frac{(\sqrt{\bar{X}} - \sqrt{\bar{Y}})}{\left(\frac{1}{4n_1} + \frac{1}{4n_2}\right)} \sim N(0, 1), \text{ asymptotically for moderately large } n_1, n_2.$$

A Case Study on Class Joining Pattern Of Students During COVID-19

Test rule: Reject H_0 iff $|Z_{obs}| > \tau_{\alpha/2}$, where α is the level of significance of the test.

Computation:

$$n_1 = 6, n_2 = 6, \bar{X} = 25.67, \bar{Y} = 20.67, \alpha = 0.05, \tau_{\alpha/2} = 1.95996$$

$$Z_{obs} = 1.8019$$

Clearly, $|Z_{obs}| < \tau_{\alpha/2}$

Therefore, we accept H_0 at a 5% level of significance.

Based on the data that we have collected it seems that the daily average number of students of 3rd year of the Statistics department joining 5 minutes before the 1st class of the day and 5 minutes before the 1st class of second half (i.e. post-lunch break) are equal.

1.2 Comparison between the daily average number of students of the Statistics Department (taking all the three years together) joining 5 minutes before the 1st class of the day and 5 minutes before the first class of second half (i.e. post-lunch break):

Let X be the random variable denoting the number of students who have joined 5 minutes before the 1st class of the day and let Y be that for the 1st class of 2nd half (i.e. post-lunch break).

Let us assume that $X \sim \text{Poisson}(\lambda_1)$ independent of $Y \sim \text{Poisson}(\lambda_2)$.

To test, $H_0: \lambda_1 = \lambda_2$ against $H_1: \lambda_1 \neq \lambda_2$.

Proceeding similarly as in **1.1**, we arrive at the following conclusions.

Computation:

$$n_1 = 6, n_2 = 6, \bar{X} = 79.50, \bar{Y} = 67.67, \alpha = 0.05, \tau_{\alpha/2} = 1.95996$$

$$Z_{obs} = 2.3913$$

Clearly, $|Z_{obs}| > \tau_{\alpha/2}$

A Case Study on Class Joining Pattern Of Students During COVID-19

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the daily average number of students of the Statistics Department (taking all the three years together) joining 5 minutes before the 1st class of the day and 5 minutes before the first class of second half (i.e. post-lunch break) are equal.

1.3 Comparison between the daily average number of students of 3rd year of the three departments considered in our experiment (viz. Statistics, Microbiology and Computer Science) joining 5 minutes before the 1st class of the day and 5 minutes before the 1st class of second half (i.e. post-lunch break):

Let X be the random variable denoting the number of students who have joined 5 minutes before the 1st class of the day and let Y be that for the 1st class of 2nd half (i.e. post-lunch break).

Let us assume that $X \sim \text{Poisson}(\lambda_1)$ independent of $Y \sim \text{Poisson}(\lambda_2)$.

To test, $H_0: \lambda_1 = \lambda_2$ against $H_1: \lambda_1 \neq \lambda_2$.

Proceeding similarly as in 1.1, we arrive at the following conclusions.

Computation:

$$n_1 = 6, n_2 = 6, \bar{X} = 75.17, \bar{Y} = 64.67, \alpha = 0.05, \tau_{\alpha/2} = 1.95996$$

$$Z_{obs} = 2.1765$$

Clearly, $|Z_{obs}| > \tau_{\alpha/2}$

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the daily average number of students of 3rd year of the three departments considered in our experiment (viz. Statistics, Microbiology and Computer Science) joining 5 minutes before the 1st class of the day and 5 minutes before the 1st class of second half (i.e. post-lunch break) are equal.

A Case Study on Class Joining Pattern Of Students During COVID-19

2.1 Test for homogeneity over the three years of Statistics department:

2.1.1 Joining time for the 1st class of the day

Consider the three years of Statistics department as 3 independent populations, each classified into 2 classes based on joining time.

Let p_{ij} denote the population proportion of members in the i^{th} class of the j^{th} population, $i = 1, 2, j = 1, 2, 3$.

To test, $H_0 : p_{i1} = p_{i2} = p_{i3}, i = 1, 2$ against $H_1 : \text{not } H_0$

Let a random sample of size n_j be drawn from the j^{th} population, $j = 1, 2, 3$.

Let f_{ij} denote the number of members in the i^{th} class of the j^{th} population, $i = 1, 2, j = 1, 2, 3$.

$$\therefore \sum_{i=1}^2 f_{ij} = n_j$$

Test statistic: Define, $\chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(f_{ij} - n_j p_{i0})^2}{n_j p_{i0}}$

Under H_0 , $\chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(f_{ij} - n_j p_{i0})^2}{n_j p_{i0}}$, where p_{i0} is the common value of p_{i1}, p_{i2}, p_{i3} under H_0 , $i = 1, 2$.

We estimate p_{i0} as, $\hat{p}_{i0} = \frac{\sum_{j=1}^3 f_{ij}}{\sum_{j=1}^3 n_j}$,

$$\therefore \chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(f_{ij} - n_j \hat{p}_{i0})^2}{n_j \hat{p}_{i0}} \sim \chi_{(3-1)(2-1)}^2$$

Test rule: Reject H_0 iff $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$, where α is the level of significance of the test.

A Case Study on Class Joining Pattern Of Students During COVID-19

Computation:

Joining time \ Year	1 st	2 nd	3 rd	Total
Joined 5 mins before	135	188	154	477
Not joined 5 mins before	195	124	206	525
Total	330	312	360	1002

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 29.3166$$

$$\text{Clearly, } \chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$$

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the three years of Statistics department are homogeneous with respect to their joining time for the 1st class of the day.

2.1.2 Joining time for the 1st class of 2nd half (i.e. post-lunch break)

Proceeding similarly as in **2.1.1**, we arrive at the following conclusions.

Computation:

Joining time \ Year	1 st	2 nd	3 rd	Total
Joined 5 mins before	98	184	124	406
Not joined 5 mins before	232	128	236	596
Total	330	312	360	1002

A Case Study on Class Joining Pattern Of Students During COVID-19

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 65.6395$$

Clearly, $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the three years of Statistics department are homogeneous with respect to their joining time for the 1st class of the 2nd half (i.e. post-lunch break).

Based on the data that we have collected it seems that there is not enough evidence to say that the three years of Statistics department are homogeneous with respect to their joining time in class.

2.2 Test for homogeneity over three departments (viz. Statistics, Microbiology and Computer Science) considered in our experiment:

2.2.1 Joining time for the 1st class of the day

Consider the three departments as 3 independent populations, each classified into 2 classes based on joining time.

Let p_{ij} denote the population proportion of members in the i^{th} class of the j^{th} population, $i = 1, 2, j = 1, 2, 3$.

To test, $H_0 : p_{i1} = p_{i2} = p_{i3}, i = 1, 2$ against $H_1 : \text{not } H_0$

Proceeding similarly as in **2.1.1**, we arrive at the following conclusions.

A Case Study on Class Joining Pattern Of Students During COVID-19

Computation:

Joining time \ Dept.	STSA	MCBA	CMSA	Total
Joined 5 mins before	154	97	200	451
Not joined 5 mins before	206	185	208	599
Total	360	282	408	1050

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 14.5574$$

Clearly, $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the three departments considered in our experiment are homogeneous with respect to their joining time for the 1st class of the day.

2.2.2 Joining time for the 1st class of 2nd half (i.e. post-lunch break)

Proceeding similarly to **2.2.1**, we arrive at the following conclusions.

Computation:

Joining time \ Dept.	STSA	MCBA	CMSA	Total
Joined 5 mins before	124	75	189	388
Not joined 5 mins before	236	207	219	662
Total	360	282	408	1050

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 29.3349$$

Clearly, $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$

A Case Study on Class Joining Pattern Of Students During COVID-19

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the three departments considered in our experiment are homogeneous with respect to their joining time for the 1st class of the 2nd half (i.e. post-lunch break).

Based on the data that we have collected it seems that there is not enough evidence to say that the three departments considered in our experiment are homogeneous with respect to their joining time in class.

3.1 Test for the independence of year and joining time in an online class:

3.1.1 Joining time for the 1st class of the day

Let the population of the students of the Statistics Department of the current year (2020 - 2021) is classified according to two attributes based on their year of study (1st, 2nd and 3rd years) (attribute A, say) and based on their joining time (attribute B, say).

Let $A_1, A_2,$ and A_3 denote classes corresponding to attribute A and let B_1 and B_2 be those corresponding to attribute B.

Let p_{ij} denote the population proportion of members who belong to the i^{th} class of A and the j^{th} class of B, $i = 1, 2, 3,$ $j = 1, 2.$

Define, $p_{i0} = \sum_{j=1}^2 p_{ij}$: the proportion of members in the population who belong to the i^{th} class of A, $i = 1, 2, 3,$ and,

$p_{0j} = \sum_{i=1}^3 p_{ij}$: the proportion of members in the population who belong to the j^{th} class of B, $j = 1, 2.$

To test, $H_0 : p_{ij} = p_{i0} * p_{0j}, \forall (i,j)$ against $H_1 : \text{not } H_0$

A Case Study on Class Joining Pattern Of Students During COVID-19

Let a random sample of size n be drawn from the population. Let f_{ij} denote the number of members in the sample who belong to the i^{th} class of A and the j^{th} class of B, $i = 1, 2, 3, j = 1, 2$.

Define, $f_{i0} = \sum_{j=1}^2 f_{ij}$, and, $f_{0j} = \sum_{i=1}^3 f_{ij}$

Test statistic: Define, $\chi^2 = \sum_{i=1}^3 \sum_{j=1}^2 \frac{(f_{ij} - n p_{i0} p_{0j})^2}{n p_{i0} p_{0j}}$

Under H_0 , $\chi^2 = \sum_{i=1}^3 \sum_{j=1}^2 \frac{(f_{ij} - n p_{i0} p_{0j})^2}{n p_{i0} p_{0j}}$

We estimate p_{i0} and p_{0j} as,

$$\hat{p}_{i0} = \frac{f_{i0}}{n}, i = 1, 2, 3, \text{ and, } \hat{p}_{0j} = \frac{f_{0j}}{n}, j = 1, 2.$$

$$\therefore \chi^2 = \sum_{i=1}^3 \sum_{j=1}^2 \frac{(f_{ij} - n \hat{p}_{i0} \hat{p}_{0j})^2}{n \hat{p}_{i0} \hat{p}_{0j}} \sim \chi_{(3-1)(2-1)}^2$$

Test rule: Reject H_0 iff $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$, where α is the level of significance of the test.

Computation:

Year \ Joining time	Joined 5 mins before	Not joined 5 mins before	Total
1 st	135	195	330
2 nd	188	124	312
3 rd	154	206	360
Total	477	525	1002

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 29.3163$$

Clearly, $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$

A Case Study on Class Joining Pattern Of Students During COVID-19

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the attributes 'year of study (Statistics department)' and 'joining time for the 1st class of the day' are independent.

3.1.2 Joining time for the 1st class of 2nd half (i.e. post-lunch break)

Proceeding similarly as in 3.1.1, we arrive at the following conclusions.

Computation:

Dept. \ Joining time	Joined 5 mins before	Not joined 5 mins before	Total
1 st	98	232	330
2 nd	184	128	312
3 rd	124	236	360
Total	406	596	1002

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 65.6401$$

$$\text{Clearly, } \chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$$

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the attributes 'Year of Study (Statistics department)' and 'joining time for the 1st class of the 2nd half (i.e. post-lunch break)' are independent.

Thus, based on the data that we have collected it seems that there is not enough evidence to say that the attributes 'year of study (Statistics department)' and 'joining time in class' are independent.

A Case Study on Class Joining Pattern Of Students During COVID-19

3.2 Test for the independence of department and joining time in an online class:

3.2.1 Joining time for the 1st class of the day

Let the population of the students of the three departments considered in our experiment, of the current year (2020 - 2021) is classified according to two attributes based on their department (viz. Statistics, Microbiology and Computer Science) (attribute A, say) and on their joining time (attribute B, say).

Let A_1 , A_2 , and A_3 denote classes corresponding to attribute A and let B_1 and B_2 be those corresponding to attribute B.

Let p_{ij} denote the population proportion of members who belong to the i^{th} class of A and the j^{th} class of B, $i = 1, 2, 3, j = 1, 2$.

Define, $p_{i0} = \sum_{j=1}^2 p_{ij}$: the proportion of members in the population who belong to the i^{th} class of A, $i = 1, 2, 3$, and,

$p_{0j} = \sum_{i=1}^3 p_{ij}$: the proportion of members in the population who belong to the j^{th} class of B, $j = 1, 2$.

To test, $H_0 : p_{ij} = p_{i0} * p_{0j}, \forall (i,j)$ against $H_1 : \text{not } H_0$

Proceeding similarly as in **3.1.1**, we arrive at the following conclusions.

Computation:

Dept. \ Joining time	Joined 5 mins before	Not joined 5 mins before	Total
STSA	154	206	360
MCBA	97	185	282
CMSA	200	208	408
Total	451	599	1050

A Case Study on Class Joining Pattern Of Students During COVID-19

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 14.5572$$

Clearly, $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the attributes 'department' and 'joining time for the 1st class of the day' are independent.

3.2.2 Joining time for the 1st class of 2nd half (i.e. post-lunch break)

Proceeding similarly as in **3.2.1**, we arrive at the following conclusions.

Computation:

Dept. \ Joining time	Joined 5 mins before	Not joined 5 mins before	Total
STSA	124	236	360
MCBA	75	207	282
CMSA	189	219	408
Total	388	662	1050

$$\alpha = 0.05, \chi_{\alpha; (3-1)(2-1)}^2 = 5.99146$$

$$\chi_{obs}^2 = 29.3342$$

Clearly, $\chi_{obs}^2 > \chi_{\alpha; (3-1)(2-1)}^2$

Therefore, we reject H_0 at a 5% level of significance.

Based on the data that we have collected it seems that there is not enough evidence to say that the attributes 'department' and 'joining time for the 1st class of the 2nd half (i.e. post-lunch)' are independent.

Thus, based on the data that we have collected it seems that there is not enough evidence to say that the attributes 'department' and 'joining time in class' are independent.

A Case Study on Class Joining Pattern Of Students During COVID-19

CONCLUSION:

From the above tests based on our data we arrive at the following conclusions:

- The daily average number of students joining 5 minutes before the 1st class of the day and 5 minutes before the 1st class of the second half (i.e. post-lunch break)
 - Seems to be equal for the students of 3rd year of the Statistics department
 - Seems to be unequal for the students of the Statistics Department (taking all the three years together)
 - Seems to be unequal for the students of 3rd year of the three departments considered in our experiment (viz. Statistics, Microbiology, and Computer Science)
- The three years of Statistics department
 - Does not seem to be homogeneous with respect to the joining time for the 1st class of the day
 - Does not seem to be homogeneous with respect to the joining time for the 1st class of 2nd half (i.e. post-lunch break)

Thus, the three years of the Statistics Department do not seem to be homogeneous with respect to the joining time in class.

- The three departments (viz. Statistics, Microbiology, and Computer Science)
 - Does not seem to be homogeneous with respect to the joining time for the 1st class of the day
 - Does not seem to be homogeneous with respect to the joining time for the 1st class of 2nd half (i.e. post-lunch break)

Thus, the three departments considered here do not seem to be homogeneous with respect to the joining time in class.

A Case Study on Class Joining Pattern Of Students During COVID-19

- The year of study (Statistics department) seems to be dependent on the
 - Joining time for the 1st class of the day
 - Joining time for the 1st class of 2nd half (i.e. post-lunch break)

Thus, the year of study (Statistics department) seems to be dependent on the joining time in class.

- The attribute department seems to be dependent on the
 - Joining time for the 1st class of the day
 - Joining time for the 1st class of 2nd half (i.e. post-lunch break)

Thus, the attribute department seems to be dependent on the joining time in class.

Articles from STUDENTS

Data Science - As A Game Changer in Drone Engineering

Ritoban Sen, Somjit Roy (3rd Year)

With the safe touchdown of the new Perseverance Rover on the surface of Mars, the discussion about technological achievements of Man has once again become the center of attraction. The most exciting upgrade over the Curiosity rover 10 years ago is the fully autonomous drone "Ingenuity" sent along with the Perseverance Rover. This rotorcraft by necessity is fully autonomous. The reason being that any kind of communication with Mars from Earth is carried on with a delay of 15 minutes, whereas Ingenuity has only 90 seconds of flight time. The autonomy of unmanned flying vehicles is an idea made possible only through ingenious application of Data Science.

Some Machine Learning methods are:

- **SUPERVISED LEARNING (SL):** This is the most common form of Machine Learning. In Supervised Learning the agents (computers) observe some input output pairs and learns a function that maps output from input.
- **UNSUPERVISED LEARNING (UL):** It has no labeled output; instead the agent focuses on observing the world and learning patterns without being given labeled data.
- **REINFORCEMENT LEARNING (RL):** It concerns to the determination of the optimal policy, which is learnt from feedbacks, these kinds of feedbacks are known as "Reward" or "Reinforcement". It repeats trial and error until the agent chooses which of the actions is appropriate to maximize the total reward.
- **DEEP LEARNING (DL):** It is also a type of Machine Learning method based on the intelligent behavior of the human brain. DL Algorithms can be classified into supervised, unsupervised and reinforcement learning according to their characteristics.

Some concepts of Data Science used in reaching autonomy of UAVs are discussed below.

Data Science: As A Game Changer in Drone Engineering

- Use of GPS in UAV

Global Positioning System or GPS is a key feature in navigation of UAV. It is a mode of Satellite Navigation that depends on the global network of over 30 satellites. At any instance, the UAV can calculate its distance from 3 or more satellites and find out the exact location on the face of the Earth. A GPS system includes an atomic clock as precise time keeping is necessary to control an UAV in even the most difficult places.

However, only a GPS based approach of drone navigation has many weaknesses and setbacks. A GPS system uses radio signals to transmit data and these signals are very susceptible to interference. In presence of high rises or mountains, line of sight connection with satellites may not be possible; in that case remote control of drones can easily be hijacked by third parties - GPS Spoofing. GPS navigation is also affected by sustained drift due to accumulative navigation errors, moreover this form of navigation fails in GPS-denied environments.

These problems have led to the most exciting innovation of image based autonomous navigation using the on-board cameras of the drones. At the core of this feat lies the set of algorithms and mathematics that are known as Neural Networks.

- Image Classification - An Eyesight for Drones

Moravec's Paradox says that "It is comparatively easy to make computers exhibit adult level performance and difficult or impossible to give them skills of a one-year-old". It refers to the fact that a computer can be taught to compute complex mathematics because we can teach the steps needed but how can a computer be taught to see, hear, or feel when we ourselves cannot comprehend the billions of years of evolution that has made us able to do those tasks.

For an UAV to achieve autonomy, it is vital that it can "see" in addition to compute. Image classification is a problem that teaches a computer to compute the probability that an image is part of a "class", where a class is a label such as 'house', 'human', 'car' etc. This task is accomplished with the help of Deep Learning which uses the concept of Neural Networks.

Data Science: As A Game Changer in Drone Engineering

- Neural Networks

The concept of neural network has its roots in the biology of human brain. Just like the human brain has millions of interconnected neurons - sending and receiving information to each other all at the same time processing it; neural networks are artificial systems that consist of connections, weights, biases, propagation function and a learning rule. A neural network must be fed data without any pre-programming from which it identifies the underlying characteristics. Neural Networks are application of supervised machine learning where an input and output is needed while it is iteratively making predictions on the data. Each "neuron" receives an input and produces an output, and each connection consists of weights and biases that decide how the output is going to be propagated to other neurons.

A special kind of Neural Network is **Convolutional Neural Network (CNN)**. This specializes in analysing and processing image data. Through the processing power of CNN, a computer can recognize the location and magnitude of different characteristics present in the image.

- Machine Learning in the Autonomy of UAVs

- i. Real life flying situations involve a great deal of uncertainty and non-linear dynamics. Using Neural Networks, it is possible to reduce the level of uncertainty and overcome problems caused by non-linearity in the dynamics.
- ii. High Dimensional data regarding flight situations are processed through Neural Networks and complex aerobatic manoeuvres are calculated with its use.
- iii. Real time - path planning and navigation strategies are important problems to overcome for a drone to achieve autonomy. Reinforcement Learning enables a drone to deal with these problems.
- iv. Object recognition with the help of concepts such as Image Classification makes a drone capable of performing tasks such as Collision Avoidance, checking for proper landing spot, classification of crops or disaster areas etc.

Data Science: As A Game Changer in Drone Engineering

- v. Like human brains, Neural Networks require huge amount of data to teach itself how to perform a task, and as a result, the computer must go through numerous failures before it can learn how to perform the task above a desired level of quality. Such heavy dependence on continuously acquired dataset makes it dangerous for an autonomous drone to fly under new flight conditions.

➤ ROLE OF UAVs IN THE REAL WORLD - PRACTICAL APPLICATIONS

Drones have recently gained its reputation on an interdisciplinary level. From its usage and applicability in Aerial Photography/Videography to the fact that it is successfully being employed in completing missions in outer space, the ideology of such a high-tech machinery is to be appreciated.

Keeping in sight the primary objective of the Unmanned Aerial Vehicles, which is pilotless flight or autonomous flight, the applicability of this piece of machinery and technology is immense.

Not only it has found its way into the customary groove or the regimen of work starting from "Payload Carrying", "Delivery of Items", "Constructional Activities", "Traffic Control and Surveillance", "Agriculture - Crop Monitoring", "Crowd Control", "Geological Surveys", etc. but also UAV's contribution to the more serious actions is noteworthy such as hosting the successful execution of the planned defense during wars and other emergency situations such as rescue from natural calamities and disasters, fire control, marine rescue, etc.

It can nearly be said that science has played its magic on an interdisciplinary level making Statistics, Machine Learning and Artificial Intelligence to coalesce with Physics and Computing, giving the real world, a technology whose applications and implementations are staggering.

• UAVs in Outer Space

- o "Ingenuity Helicopter, strapped to NASA's Perseverance Mars Rover, sends first status report - The First Helicopter to take flight outside Earth."

Data Science: As A Game Changer in Drone Engineering

On 18th February 2021, NASA's much awaited venture - The Perseverance Mars Rover entered the Martian Atmosphere and resulted into a successful touchdown. Along with the Rover, a robotic rotorcraft named "Ingenuity" was also sent to MARS, aimed at analyzing whether the Martian Atmosphere facilitates flying, in other words the primary objective of Ingenuity was to validate the theoretical assumptions of possible flight in MARS' thin atmosphere, thereby collecting data and information about the Red Planet.

Contribution of Data Science and AI has led to the successful implementation of Ingenuity resulting into autonomous flight and facilitating the collection of data and information in MARS through programming, making the rotorcraft to take decisions on its own - a non-viable scenario through usual engineering.

- o "Titan to witness NASA's Dragonfly Mission - A Future Prospect."

Yet another breakthrough is scheduled by NASA in 2026, where an Unmanned Aerial Multirotor Vehicle named "Dragonfly" is to be sent to the icy moon of Saturn - Titan, to investigate promising locations on Titan aimed at studying prebiotic chemical processes common to both Earth and Titan.

Not only this is a breakthrough in the Physical World, but also adding up to the numerous contributions made by Data Science, where this time Machine Learning and AI has made it possible for NASA to advance with their search of building blocks of life in a planet other than Earth.

- **UAVs in Combat**

- o "Indian Army leases 4 Unmanned Aerial Vehicles from Israel under Emergency procurement program."

Apart from the utility of UAVs to study the possibility of existence of life in outer planets, these technological hybrids are also being put to use in designing defense mechanisms and safety plans in situations of extreme emergencies such as combat, war, etc.

Data Science: As A Game Changer in Drone Engineering

Indian Army along with the US Army has approached towards the usage of these AI based UAVs, appreciating the concept of Autonomous or pilotless flight, i.e., taking flight without a pilot, which seems to be very useful, during combats and wars, as life risk to the pilot is averted.

These UAVs are put into action according to the different needs and objective of the missions.

One of the most notable uses of the UAVs and henceforth the power of data science was against the Talibans, where these Vehicles acted as intelligence, surveillance and reconnaissance (ISR) platform as well as an on-call strike platform.

Few of the UAVs used for Military applications are as follows:

- o *Global Hawk*
- o *Reaper*
- o *Shadow, etc.*

- **UAVs in Agriculture**

- o "Agriculture Ministry gets nod for using drones for crop loss assessment."

Data Science and Statistics have extended their helping hands towards the field of agriculture also. Apart from the theoretical applications of Statistics such as Designing of Experiments for better yield of crop, implementations of several ANOVA models to statistically analyze the annual harvest or yield of crop, Statistics has also contributed physically through Machine Learning and Data Science facilitating crop monitoring, Soil and Field Analysis, Manure Applications, Altering Dosages of the different fertilizers applied, irrigation, livestock farming, etc.

The UAVs have reduced the workload of the farmers by efficiently monitoring the essentials of crop plantations and productions, also taking a step forward towards intelligently studying and analyzing effective plantation techniques to optimize the yield of crop and earn greater profits by collecting farm yield data.

There are some drones or UAVs especially designed and manufactured for agricultural purposes, which are:

Data Science: As A Game Changer in Drone Engineering

- *Agras T16 designed by DJI Technologies.*
- *PHX Complete System developed by a Minnesota - based agricultural imagery software company named Sentera.*

- **UAVs as a miscellaneous invention**

Above are a handful of the objectives and benchmark met by an Unmanned Aerial Vehicles, popularly known as Drone. There are innumerable applications of the UAVs which when listed would go on and on.

- Earlier, the coverage of any recreational concert was done through video cameras, by a conventional camera man, which nowadays has been made easier and more flexible by the implementation of the UAVs which tend to cover the entire program without even showing the need to be controlled, as AI and Machine Learning has facilitated the opportunities of clicking pictures or recording a video whenever necessary.
- **Unmanned Aircraft System Traffic Management (UTM)** - According to the flight aviation data company - FlightAware, around 1.2 billion people around the world are airborne at any given point of time, thereby reflecting today's airspace busier than ever.

But thanks to Data Science and other advanced technologies, UAVs have made it possible to manage the traffic situations which are airborne, eventually recording data and intelligently suggesting ways of optimizing the traffic management on its own after analyzing the data collected.

“Real UTM Services are already starting to be implemented. The system we've put in place today and over the next few years will be the one that's in place for decades to come. It will affect the next generation of people using the air traffic control system.”

- Joe Polastre, Airbus UTM, Head of UTM Products.

Data Science: As A Game Changer in Drone Engineering

- ***Airborne Food Delivery System: Air Swiggy*** – What if the food we order online gets delivered via a drone in lieu of a delivery official. This seems to be a very plausible scenario in the coming days. Yes, it would surely affect the employment status of the delivery personnel, but this is quite an achievable target through the technologies and facilities brought into picture by Data Science. The UAV's could easily be programmed with the destination where the delivery is to be made and the task could easily be completed.

This type of a technology could be used by the restaurants or the franchises in the situations of heavy rush due to high demand, in other words, a large number of take away orders have been placed but there are not many delivery persons to fulfill or complete the orders. In such a scenario an UAV can be of immense help.

Likewise, we can encounter countable infinite applications of this rising technology – Unmanned Aerial Vehicles.

➤ UAVs WITH A FORSEEABLE FUTURE - CONCLUSION

In recent times, society has witnessed an enormous amount of technological developments, and among them the idea of "Autonomous Flight" has emerged as the most successful venture, which is likely to have a promising future in almost every field which can be thought of.

Starting from high - rise construction to autonomous food delivery, from combat in the battlefield to discovering life forms in outer space and other planets, the application of Machine Learning and AI has worked wonders for the UAVs, giving it an extremely applicable and balanced structure facilitating an usage on an interdisciplinary level. It is very difficult to think of a particular area where these intelligent drones cannot be put to use.

With the advancement in the fields of Data Science and Engineering, it is quite probable for these Unmanned Aerial Vehicles to book and mark a successful future ahead of its time for itself.

Data Science: As A Game Changer in Drone Engineering

Some of the companies interested towards the application of these autonomous drones are the mega - corps like:

- o Amazon
- o Uber
- o FedEx
- o Microsoft
- o Facebook
- o Apple and many more.

➤ REFERENCES

- i. Veena Ghorakavi - "Neural Networks | A beginner's guide".
- ii. Jason Brownlee - "How do convolutional layers work in deep learning neural networks".
- iii. K. Amer, M. Samy, M. Shaker and M. ElHelw, "Deep Convolutional Neural Network-Based Autonomous Drone Navigation"
- iv. "6 Things to Know About NASA's Ingenuity Mars Helicopter" - Jia Rui Cook, Jet Propulsion Laboratory, California, Pasadena & Grey Hautaluoma/Alana Johnson, NASA Headquarters, Washington.
- v. "Sky-Farmers: Application of Unmanned Aerial Vehicles (UAV) in Agriculture" - Chika Yinka - Banjo and Olasupo Ajayi, December 17th, 2019.

Articles from STUDENTS

Can You Win by Losing?

Rajdeep Kundu, Sambit Das (3rd Year)

It seems quite strange how possibly one can win by losing. Well, it is possible to wring out a win by using one's losses. There is a paradox which states that we can construct a winning strategy by using the combination of losing strategies. Let us see how this works by taking the following examples.

Example 1:

1. We consider two games, X and Y, with the following rules:

In both the games, you must start with a certain amount of money and for every round of game X, a biased coin with the probability of showing up head with 49.5% is tossed.

2. In game Y, if the amount of money you are left with is not divisible by 3, a coin with a 74.5% probability of showing up head is tossed and if the amount of money is divisible by 3, a coin with the probability of obtaining head with 9.5% is tossed.

In both games, for showing up head you will earn Re.1 and for showing up tail you will lose Re.1.

Which one should you play?

In game X, after each round, you are expected to lose Rs.0.01. If you start with Rs.100 and keep playing this game, you are expected to lose your entire money after 10000 rounds. So this is entirely a loss-making game in the long run and you should not be interested in playing this.

On the other hand, let us have a look at game Y. If you are left with an amount of money not divisible by 3, after each round you are expected to earn Rs.0.49 whereas for the amount divisible by 3, after each round you are expected to lose Rs.0.81. For game Y, by the analysis of Markov Chain for its state transition matrix, we have that the probability of having money left of the amount divisible by 3 is 0.3836 that is close to 40% time when you play this game. Due to this disproportionate share of the outcome, game Y also ends up being a loss-making game in long run.

Can You Win By Losing?

But if we play these two games alternately, then we can construct a winning strategy from two losing ones.

Let us take a simplified example to understand how it works.

Example 2:

Consider two games, A and B, with the following rules:

1. In game A you lose Re.1 every time you play.
2. Game B has two parts. If the amount of money left is an even number, then you win Rs.3, else you lose Rs.5

Both the games, A and B, are clearly losing games.

But if we combine the games and play them alternately, then we can make this game profitable. Suppose you start the game with Rs.100. After playing game A, you are left with Rs.99. Now if you play game B then you lose Rs.5. So instead you play game A again you are left with Rs.98. Then if we play game B we win Rs.3 and we make a Re.1 profit. Thus if we repeat to play the game this way, then we can make an infinite amount of profit.

But this is more of a cheat code (as you are not selecting the games randomly) that helps you win an infinite amount of money.

So where is exactly the paradox?

The paradox is that even if you play the two losing games randomly between themselves, it will still produce a winning game. This is known as Parrondo's Paradox named after its creator Juan Parrondo.

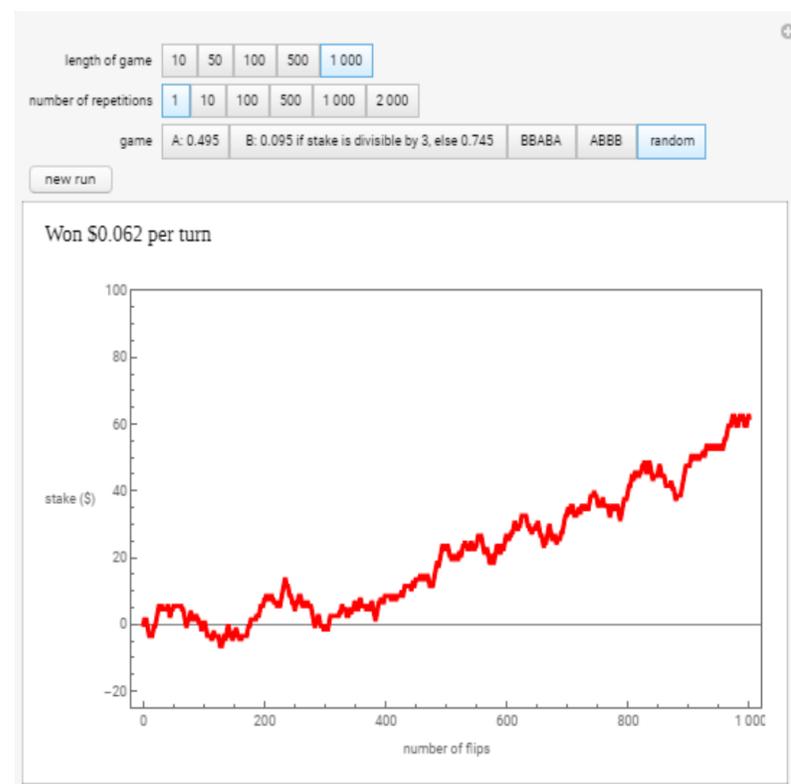
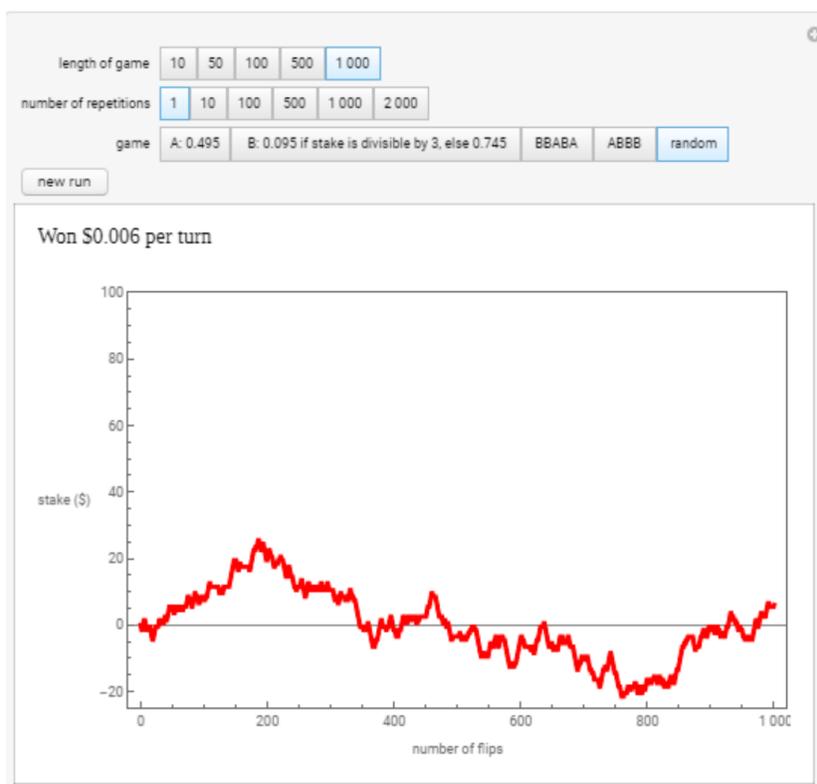
We can use the Parrondo Paradox wolfram simulator (available in the link www.wolframcloud.com/objects/demonstrations/TheParrondoParadox-source.nb) to prove it.

This simulator uses Parrondo's biased coin flip odds and allows us to set the number of times we want to flip the coin and how many times we want to repeat the experiment.

Can You Win By Losing?

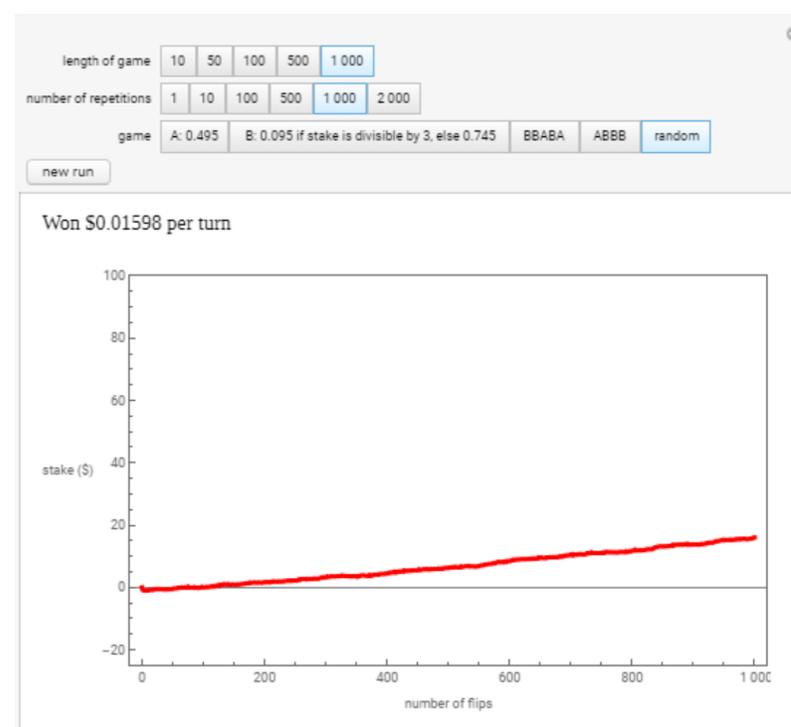
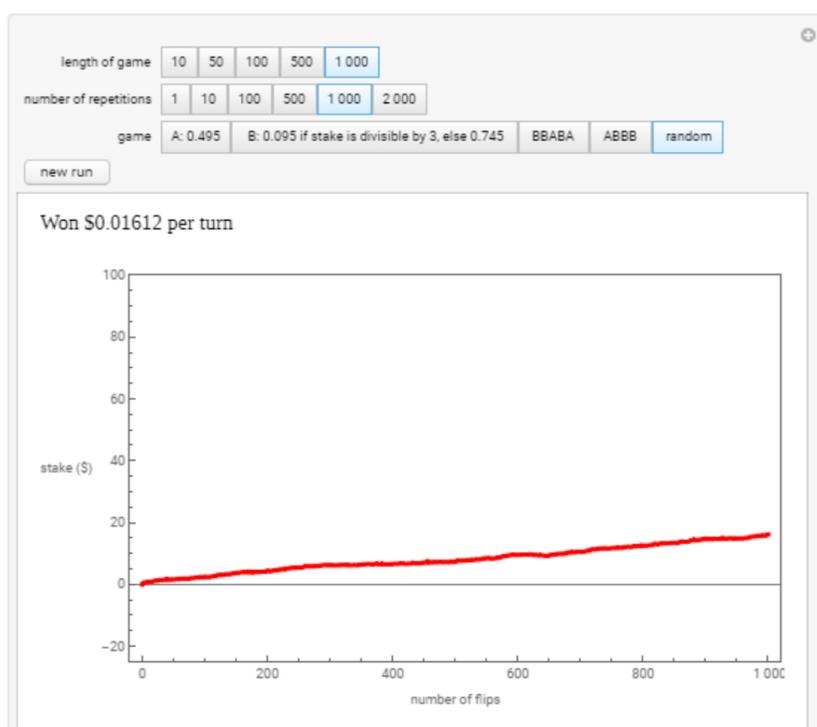
The Parrondo numbers used in the simulator for games A and B are identical to our coin tossing example.

We set the number of coin tosses to 1000 with only 1 repetition. We get the outputs for two random games as follows:



We can clearly see that sometimes we may win and sometimes we may lose.

But now if we set the number of repetitions to 1000, we see the following results:



Can You Win By Losing?

The graphs were generated using the Parrondo Paradox Wolfram Simulator [1].

We can see a positive slope every time we run the simulation. Despite playing the games A and B randomly, we see that in the long run it always yields in a win.

Thus even if it takes a while, even if it sounds weird, mathematically you can actually win from losing.

Reference: https://en.m.wikipedia.org/wiki/Parrondo's_paradox
[1] <https://www.wolframcloud.com/objects/demonstrations/TheParrondoParadox-source.nb>

Articles from STUDENTS

From Gambling to Probability: Modelling Chance Games

Somjit Roy (3rd Year)

I. BACKGROUND OF THE GAMBLER'S RUIN PROBLEM

The first origin of the problem was from a letter given by Blaise Pascal to Pierre De Fermat in 1656. In that very year, Pierre De Carcavi in his letter to Huygens modified the version of the problem as given by Pascal. The problem posed to Huygens was as follows:

"Let there be two players (A & B) playing a game with three dice. Player A wins a point if 11 is thrown on the three dice and Player B wins a point if 14 turns up. But in lieu of the normal accumulation of the points, a point shall be added to the player's score if his/her opponent is having 0 points otherwise it would be subtracted from his /her opponent's score. The Player who scores 12 points first, is the winner.

The question is, what are the relative chances of each player winning the game?"

Huygens's on the other hand once again reformulated the problem as:

"Each player starts with 12 points, and a successful roll of the three dice (11 for Player A and 14 for Player B) adds one point to the score of the winning player and subtracts a point from the opponent's score.

What is the probability of winning the game for each player, if the loser of this particular game is the first one to reach zero points?" [4]

The above two problem statements as proposed by Carcavi and Huygens respectively, are the derivative of what is known today widely as the "**Gambler's Ruin Problem**", one of the most frequently visited problems in the world of "Game of Chances".

From Gambling to Probability: Modelling Chance Games

II. INTRODUCING THE GAMBLER'S RUIN PROBLEM - STATEMENT & SETUP

Now instead of talking about the game of dice between two players, we shift ourselves to the world of Gambling and Betting. We take into consideration the following game:

(The Problem Statement): "Let there be two Gamblers (players), A and B. Player A starts the game with an initial amount of money and, Player B starts the game with an initial amount of money, such that the total amount of money available in the game is

Suppose that the probability of A winning a particular round of the game is p and in the case of B, the probability is $q=1-p$. The game stops, when any one of the players goes bankrupt, i.e., the player with all wins. Then what are the probabilities of A and B winning the Game respectively?" [3]

Thus, we have the Gambler's Ruin Problem. Here we will mainly discuss in the context of player A, and the results for player B will just follow from that.

III MODELLING THE GAMBLER'S RUIN PROBLEM THROUGH MARKOV CHAINS AND RANDOM WALKS

Both Markov Chains and Random Walks being ideas related to "Random Processes", have a very close connection with the problem stated above, facilitating the exploration of various dimensions of the problem.

RANDOM WALKS: Suppose $\{\epsilon_t\}$ is a purely random process with mean 0 and variance σ_ϵ^2 .

- The process $\{X_t\}$ is said to be a Random Walk if $X_t = X_{t-1} + \epsilon_t$.
- The above-defined process usually has an initial state of 0 at $t=0$, i.e., starts at 0, which implies, $X_0=0$.
- Further $X_1 = \epsilon_1$, $X_2 = X_1 + \epsilon_2 = \epsilon_1 + \epsilon_2$, $X_3 = X_2 + \epsilon_3 = \epsilon_1 + \epsilon_2 + \epsilon_3$, and so on.

From Gambling to Probability: Modelling Chance Games

- Hence, in general, $X_t = \sum_{i=1}^t \epsilon_i$, with a variance of $t\sigma_\epsilon^2$, as a result of which the Random Walks are instances of **Non-Stationary Processes**.
- The Gambler's Ruin Problem is a beautiful example of a **One-Dimensional Random Walk** that could be both biased as well as unbiased.
- **MARKOV CHAINS:** These are the chains that highlight the evolution of a particular system, having a more probabilistic touch to it. The ideology and concept of the Markov Chains are entirely based on the Markovian Property, which states that:

“Given information about the present state, the past and future are conditionally independent.”

A more formal framework of this Markov's assumption puts forward the formal definition of Markov Chains, i.e.,

The stochastic process $\{X_n: n = 0, 1, 2, 3, \dots\}$ is called a **Markov Chain**, if for $j, k, j_1, \dots, j_{n-1} \in N$,

$$\Rightarrow P \{X_n = k | X_{n-1} = j, X_{n-2} = j_1, \dots, X_1 = j_{n-1}\} = P \{X_n = k | X_{n-1} = j\}$$

➤ THE ABSORBED STATES, TRANSITION PROBABILITIES, AND MODELLING OF THE PROBLEM:

- Two situations are to be taken note of, when the Gambler or the player gets absorbed, i.e., the Gambler has nowhere to move further in the game, which happens only when he/she gets bankrupt having Rs. 0 or when he/she wins the game having the entire stake of money put into the game at the beginning, i.e., Rs. $(n_1 + n_2)$.

These two states are particularly known as the **Absorbed States**.

From Gambling to Probability: Modelling Chance Games

- Keeping the problem statement in mind as stated above, if we consider Gambler A, the objective of this Gambler is to reach Rs. $n_1 + n_2$, without getting bankrupt.

The Schematic Representation of different states, the Gambler A is in, along with the amount of money at different points of time with the Gambler is as follows:

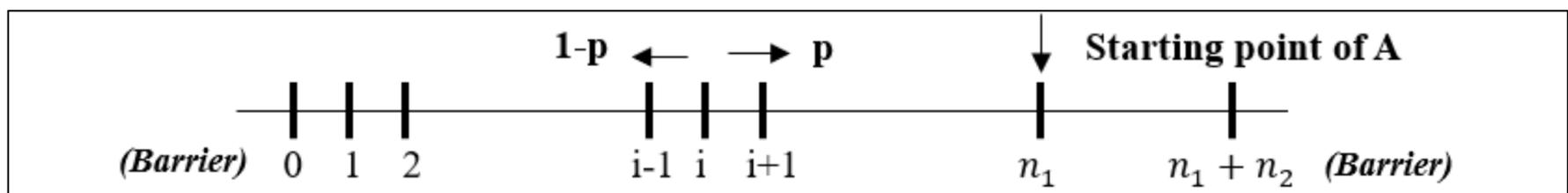


FIG 1: **Transition Diagram** of the Markov Chain under consideration

Note that, States 0 and $n_1 + n_2$ or equivalently the amount of money the Gambler has, concerns with a probability of staying at that state = 1. In other words, as we reach those two Absorbing States, the probability of staying there is unity.

Now, let X_n : the amount of money or fortune that is there with the Gambler after the n th gamble.

Then, this $\{X_n\}$ yields a **Markov Chain** on the **finite state space** $\zeta = \{0, 1, 2, \dots, n_1 + n_2\}$.

- The **Transition Probabilities** signify the chance or probability of moving or propagating from one state to the other. For the above situation, the transition probabilities are as follows:
 - $p_{i,i+1} = p$, i.e., moving from the i th state to the $(i+1)$ th state is p , or equivalently acquiring Rs. $(i+1)$ from Rs. i has a probability p , $\forall 0 < i < n_1 + n_2$.
 - $p_{i,i-1} = 1 - p = q$, i.e., moving from the i th state to the $(i-1)$ th state is q , or equivalently the Gambler loses a rupee with probability q , $\forall 0 < i < n_1 + n_2$.
 - $p_{0,0} = 1 = p_{n_1+n_2,n_1+n_2}$

From Gambling to Probability: Modelling Chance Games

- The **Probability Transition Matrix** for the above layout of the Markov Chain modeling the Gambler's Ruin Problem gives us the arrayed representation of the various transition probabilities, which is represented as follows:

(N.B: Here we consider the total number of states to be 5, i.e., from 0 to 4, or equivalently a total of Rs. 4 being put into the game. We have along the rows of the matrix, the state from which we are starting, and along the columns of the matrix, the state we are reaching to. Also, note that the sum of each row is equal to unity.)

$$T = \begin{pmatrix} \text{Different States} & \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \\ \text{State 0} & 1 & 0 & 0 & 0 & 0 \\ \text{State 1} & q & 0 & p & 0 & 0 \\ \text{State 2} & 0 & q & 0 & p & 0 \\ \text{State 3} & 0 & 0 & q & 0 & p \\ \text{State 4} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Likewise, Transition Matrices of higher order can also be constructed.

- As the game proceeds further, this Markov Chain generates a **One-Dimensional Simple Random Walk**, as follows: (Considering the case of Gambler A, where Gambler A starts with an initial fortune of Rs. n_1 .)

$$X_n = n_1 + \epsilon_1 + \epsilon_2 + \epsilon_3 + \dots + \epsilon_n, n \geq 1, X_0 = n_1$$

Here, $\{\epsilon_t\}_{t \geq 1}$ is a sequence of independently and identically distributed random variables defined as follows, $\forall t$:

$$\epsilon_t = \begin{cases} 1, & \text{with probability } p \\ -1, & \text{with probability } q = 1 - p \end{cases}$$

When $p = 0.5$, then the Random walk is said to be a **Symmetric or Unbiased Random Walk**, otherwise, it is regarded as a **Biased Random Walk**.

From Gambling to Probability: Modelling Chance Games

➤ SOLUTION TO THE GAMBLER'S RUIN PROBLEM

As in the problem stated earlier, we have taken into consideration, the Gambler's Ruin scenario w.r.t two Gamblers namely A and B.

Gambler A starts with an initial amount of ₹ n_1 and Gambler B starts with the initial amount of ₹ n_2 such that the total money is ₹ $(n_1 + n_2)$.

The probability of any Gambler winning a particular round of the Game is p and that of losing is $1 - p$. The Game stops when either of the Gambler gets broke/bankrupt, i.e., having ₹ 0 , which would imply the other Gambler has ₹ $n_1 + n_2$. On winning a round, ₹ 1 is given to the winner from the losing Gambler.

To determine P (Gambler A winning the game) = ?

Let us define the following:

P_i : the probability of winning the game for A, when A starts with an initial stake of ₹ i , i.e.,

$P_i = P(\text{A wins the game} | \text{A starts with ₹}i)$.

X : A wins the entire game.

E : A wins the first round of the game.

X_t : the amount of money with A at any time point t of the Game

To find P_i . Let us condition on the first round of the Game. Note that the results of a particular round are independent of the results of the previous round.

By Theorem of Total Probability:

$$\Rightarrow P_i = P(X \cap E | X_0 = i) + P(X \cap E' | X_0 = i) = \\ P(E | X_0 = i)P(X | X_1 = i+1) + P(E' | X_0 = i)P(X | X_1 = i-1)$$

$$\Rightarrow P_i = p P_{i+1} + (1-p) P_{i-1} \text{-----} (i), 1 \leq i \leq (n_1 + n_2 - 1)$$

From Gambling to Probability: Modelling Chance Games

Take a note of the boundary conditions: $P_0 = 0$ (A starts bankrupt) and $P_{n_1+n_2} = 1$ (B starts bankrupt) -----(ii)

Now,

Equation (i) is a *recursion* (**Second Order Linear Difference Equation**) to be solved.

We can re-write the equation (i) as:

$$x = px^2 + (1-p) \Rightarrow px^2 - x + (1-p) = 0 \Rightarrow x = \frac{1 \pm (2p-1)}{2p}$$

$$\Rightarrow x = 1, \frac{1-p}{p}$$

Hence,

$$P_i = \alpha 1^i + \beta \left\{ \frac{1-p}{p} \right\}^i, p \neq (1-p) \text{ (Since the roots are distinct)}$$

Now using the Boundary Conditions as in (ii), we get:

- $P_0 = \alpha + \beta = 0$
- $P_{n_1+n_2} = \alpha + \beta \left\{ \frac{1-p}{p} \right\}^{n_1+n_2} = 1$

Solving the above two equations simultaneously, we get the values of α and β as:

$$\Rightarrow \alpha = \frac{1}{1 - \left(\frac{1-p}{p}\right)^{n_1+n_2}} \text{ and } \beta = \frac{-1}{1 - \left(\frac{1-p}{p}\right)^{n_1+n_2}}$$

Finally, we obtain,

$$P_i = \begin{cases} \frac{1 - \left(\frac{1-p}{p}\right)^i}{1 - \left(\frac{1-p}{p}\right)^{n_1+n_2}}, & 1-p \neq p \\ \frac{i}{n_1+n_2}, & 1-p = p \Rightarrow p = 0.5 \end{cases}$$

From Gambling to Probability: Modelling Chance Games

Therefore, the probability for Gambler A winning the entire game, when initially starting with a fortune of Rs. n_1 is given by:

$$P_{n_1} = \begin{cases} \frac{1 - \left(\frac{1-p}{p}\right)^{n_1}}{1 - \left(\frac{1-p}{p}\right)^{n_1+n_2}}, & 1 - p \neq p \\ \frac{n_1}{n_1+n_2}, & p = 1/2 \end{cases}$$

Similarly, the probability for Gambler B winning the entire game, when initially starting with a fortune of Rs. n_2 is given by:

$$P_{n_2} = \begin{cases} \frac{1 - \left(\frac{1-p}{p}\right)^{n_2}}{1 - \left(\frac{1-p}{p}\right)^{n_1+n_2}}, & 1 - p \neq p \\ \frac{n_2}{n_1+n_2}, & p = 1/2 \end{cases}$$

➤ A SHORT SIMULATION OF THE GAMBLER'S RUIN PROBLEM

OBJECTIVE: To calculate the win probability of a Gambler through simulations of the game using computer software, (R - used in this case).

PROCEDURE: We have taken into account a fair game in the Gambler's Ruin Scenario, i.e., the probability for each Gambler to win a round of the game is $1/2$. The Game has been simulated for 1000 rounds (each round repeated 100 times) and we consider Gambler A to start with $n_1 = ₹1$ and Gambler B to start with $n_2 = ₹3$ such that the total sum is $n_1 + n_2 = ₹4$.

Since it is the unbiased scenario, as we increase the rounds of the Game, the probability of Gambler A winning the game should be $n_1/(n_1+n_2) = 1/4 = 0.25$, i.e., the long-run win probability of Gambler A (proved earlier in section IV).

COMPUTATION OF THE TRANSITION MATRIX: We try to formulate the Probability Transition Matrix for the above problem:

From Gambling to Probability: Modelling Chance Games

$$T = \begin{pmatrix} \text{Different States} & \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \\ \text{State 0} & 1 & 0 & 0 & 0 & 0 \\ \text{State 1} & 0.5 & 0 & 0.5 & 0 & 0 \\ \text{State 2} & 0 & 0.5 & 0 & 0.5 & 0 \\ \text{State 3} & 0 & 0 & 0.5 & 0 & 0.5 \\ \text{State 4} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

An Important Observation!

If we compute the Higher-Order Transition Matrices, that is $T^{20}, T^{30}, T^{50}, \dots$; we note that the probability of the Gambler A winning the game when initially starting with a wealth of Rs. 1 is converging to a particular value, also for Gambler B, who has started the game with an initial amount of Rs. 3, the win probability converges to a particular point.

Therefore, we have,

$$T^{20} = \begin{pmatrix} \text{Different States} & \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \\ \text{State 0} & 1 & 0 & 0 & 0 & 0 \\ \text{State 1} & 0.74951 & 0.00049 & 0 & 0.00049 & 0.24951 \\ \text{State 2} & 0.49951 & 0 & 0.00097 & 0.5 & 0.49951 \\ \text{State 3} & 0.24951 & 0.00049 & 0 & 0.00049 & 0.74951 \\ \text{State 4} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Consequently,

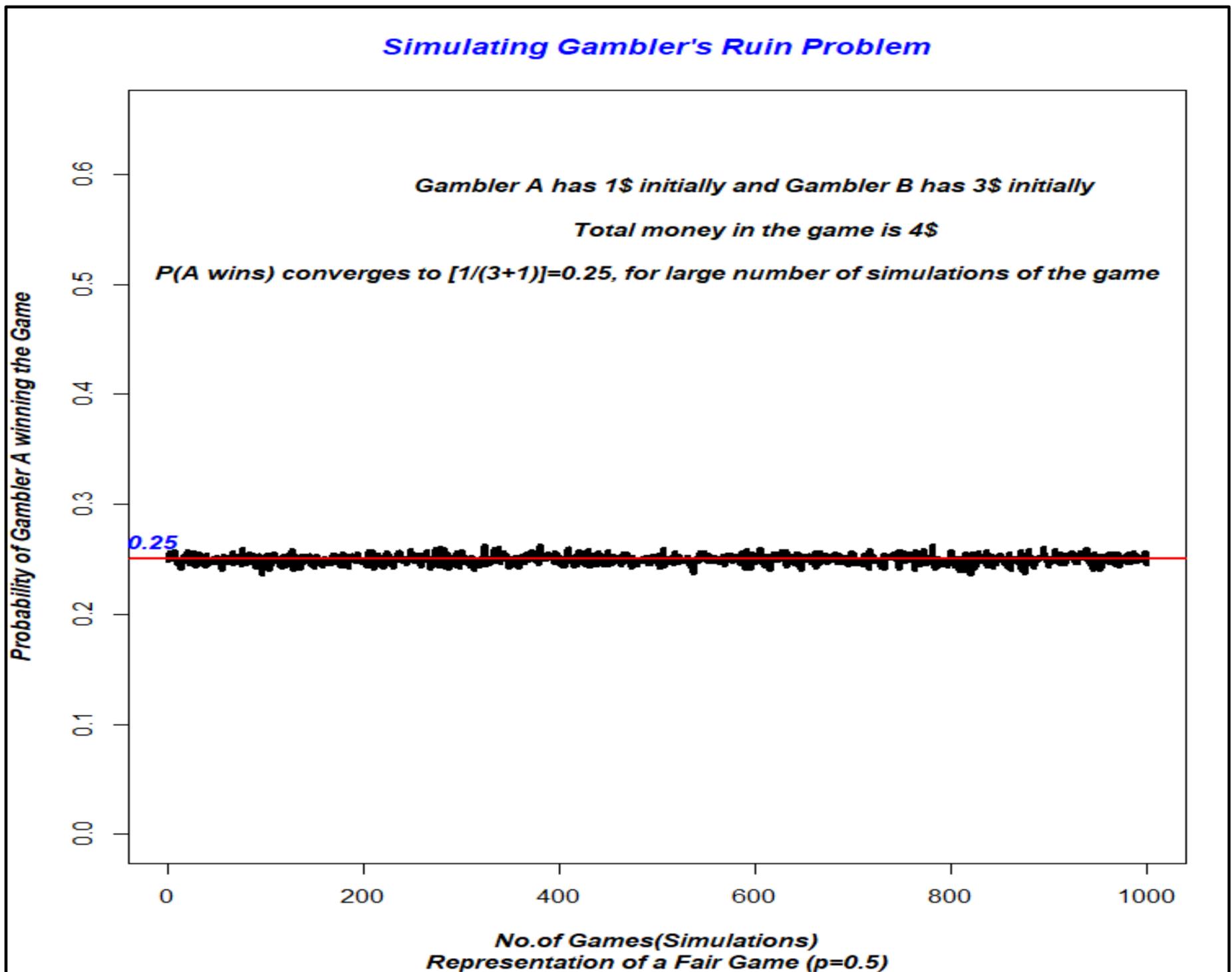
$$T^{50} = \begin{pmatrix} \text{Different States} & \text{State 0} & \text{State 1} & \text{State 2} & \text{State 3} & \text{State 4} \\ \text{State 0} & 1 & 0 & 0 & 0 & 0 \\ \text{State 1} & 0.75 & 0 & 0 & 0 & 0.25 \\ \text{State 2} & 0.50 & 0 & 0 & 0 & 0.50 \\ \text{State 3} & 0.25 & 0 & 0 & 0 & 0.75 \\ \text{State 4} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Hence, we can observe that, in long run, i.e., after 50 trials of the game stated above, the probability of Gambler A to reach State 4 from State 1 (since A started the game with Rs. 1 which is equivalent to State 1), i.e., win the game by achieving all the money converges to 0.25.

Also, in a very similar fashion, the probability of Gambler B, winning the game, i.e., reaching State 4 from State 3 (since B started with Rs. 3 which is equivalent to state 3) converges to 0.75.

From Gambling to Probability: Modelling Chance Games

A GRAPHICAL REPRESENTATION OF THE SIMULATION OF THE WIN PROBABILITY OF GAMBLER A:



VI FEW OTHER VARIANTS OF THE GAMBLER'S RUIN PROBLEM - APPLICATIONS

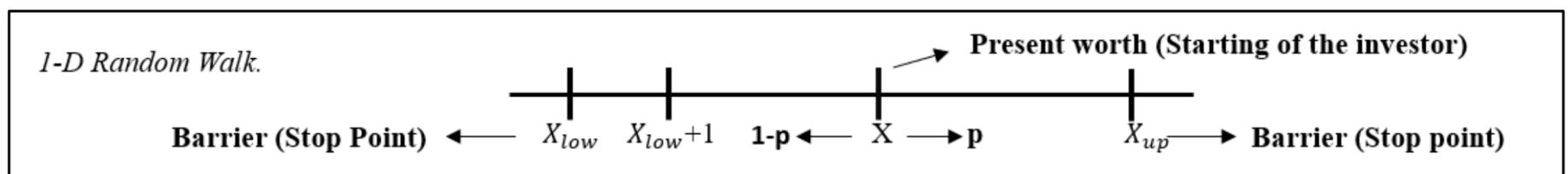
Apart from the scenario of Gambling, the problem appears in many other forms, as already mentioned earlier (Carcavi's and Huygens' structure of the problem).

From Gambling to Probability: Modelling Chance Games

➤ **Modelling the scenario of a stock-market investor using the Gambler's Ruin Problem:**

Problem Statement: Consider a hypothetical situation where a stock market investor owns shares in a particular stock whose present worth is ₹ X . The investor decides to sell off his shares if the stock price goes up to ₹ X_{up} or if it goes down to ₹ X_{low} . Now each change of stock price is either up by ₹ 1 with probability p or is down by ₹ 1 with probability $1-p$, and the successive changes in the stock prices are assumed to be independent. Then the investor needs to get an idea of the probability that he/she retires a winner. [1]

Modelling: Here the above situation can be exactly modelled through the notion of Gambler's Ruin Problem.



Note that here both the points X_{low} & X_{up} are the winning points of the investor. Finding out the probability w.r.t one point would give us the other.

The Banach's Matchbox Problem, as a case of the Gambler's Ruin Problem:

Problem Statement: Suppose a mathematician always carries 2 matchboxes initially each containing N match sticks. When a matchstick is required, the mathematician selects one of the two boxes at random and takes a stick out of it. What is the probability that one of the matchboxes will be found empty when the other contains exactly r match sticks? [2]

Reframing and Modelling: Let us consider one of the matchboxes to be Gambler A and the other to be Gambler B, starting with an initial amount of ₹ N and ₹ N respectively, such that the total money in the game is ₹ $2N$. Here the selection of any matchbox is equally likely that is with a probability of 0.5 and the matchboxes are selected independently. On selection of a matchbox, 1 matchstick is withdrawn from it implying that Gambler A loses ₹ 1. Hence, we consider the selection of a matchbox to be a loss for Gambler A, and if Gambler A wins i.e. when the box is not selected then no money is withdrawn from Gambler A. The process of

From Gambling to Probability: Modelling Chance Games

selecting match sticks stops when one of the two matchboxes is found to be empty, i.e., when Gambler A gets broke or Gambler B gets broke.

N.B: Also, in this application of the Gambler's Ruin Problem, if a Gambler wins, it is equivalent to a matchbox not getting selected by the mathematician, i.e., the Gambler does not add money to his existing stake figures, indeed his sum remains the same.

➤ Likewise, there are further forms and constructions to the problem under consideration, one of which is the extension of the problem to N (>2) players, known as the “**N-player Ruin Problem**”.

VII PRACTICALITY OF THE GAMBLER'S RUIN PROBLEM - CONCLUSION

The Gambler's Ruin Problem entails the practical scenario of the life of a Gambler, where a Gambler expects to visit a casino with an initial amount of money and return with a large sum. The problem specifically highlights the probability of the Gambler winning a particular game with a specific amount of starting stake say “ i ” or n_1 . Apart from dealing with the probability, the problem facilitates a player to build up strategies of winning betting games.

But why the term “**ruin**”?

We have already mentioned that the problem of our interest is a representation of a One-Dimensional Random Walk with two Absorbing Barriers, as illustrated by the Schematic Representation/Transition Diagram earlier. An Absorbing Barrier is a point where the game of placing bets eventually comes to an end, i.e., the game stops resulting in the victory of either of the players. In this problem considered, we speak in terms of the gambler getting broke or bankrupt (at which point the game ends). Hence the term ruin.

Now also we need to take note of the fact that: “**A Casino would not want a Gambler to win easily**”.

From Gambling to Probability: Modelling Chance Games

Speaking in favour of the casino, p if less than 0.5, makes the probability of the Gambler winning a particular round less. Similarly, if we speak against the favour of the casino, p if greater than 0.5 increases the probability of winning a round for a gambler.

This specific problem, because of its noteworthy characteristics portraying the practical applications of Statistics and Probability in the Real World holds great importance in the world of Statistics and Mathematics. The Gambler's Ruin Problem not only reflects the probabilistic insights but has various dimensions relating to other concepts such as Markov Chains and Random Walks, which are applied and used on an interdisciplinary level such as describing the evolutionary process and distinct states of the systems, including systems in the world of Physics, Chemistry, Engineering, Biology, etc.

Apart from helping the Gamblers in developing strategic policies for the Game of Gambling, it helps the casinos and gambling stations to probabilistically formulate their profit margins, by bringing in new forms and derivatives of the game, which is more lucrative.

VIII REFERENCES

[1] Random Walks: Lecture Notes, Mathematics for Computer Science December 12, 2006, Tom Leighton and Ronitt Rubinfeld, Massachusetts Institute of Technology.

[2] Gambler's Ruin and Random Variables, Statistics 110, Harvard University.

[3] 50 Challenging Problems in Probability - Frederick Mosteller.

[4] Markov's Chain & Gambler's Ruin Problem: Utilizing Linear Algebra, Statistics and R to Get Rich or At Least Not Get Ruined. Justin Zhu, December 2017.

Articles from STUDENTS

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

Shabnam Dutta (3rd Year)

We intend to observe the effectiveness of the L-moment method of parameter estimation for frequency analysis of natural extremes.

a/ Frequency Analysis:

The primary objective of frequency analysis is to relate the magnitude of extreme values to their frequency of occurrence through probability distributions. This is often a major tool to analyse natural events like floods, Tsunamis, drought, heat and cold waves, cyclones etc.[5]

b/ L-moments:

'Linear moments' (L-moments) are linear combinations of order statistics, analogous to conventional moments in their interpretation. Different data characteristics like moments, skewness and kurtosis etc., all can be measured by L-moments. These are obtained from probability weighted moments (PWM), which are probability weighted linear combinations of order statistics. L-moments can be used for summary statistics also[4].

The main utility of L-moment lies in the graduation of natural data, especially extreme events. The benefit primarily is twofold:

- i. It is a frequent realization that the commonly practiced distributions fit natural data well when their parameters are estimated by L-moments. The other methods viz. Method of Moments (MOM) and Maximum Likelihood Estimation (MLE) fall significantly behind in actual performance in the sense that when parameters are estimated by these two methods, the fit is rejected more often than L-moments by 'goodness of fit' tests. This advantage of L-moments is probably due to the fact that they use the order of the observations along with their values, unlike MOM or MLE. Thus, they contain more information about the population.
- ii. Calculation for L-moments is the simplest of the three. MLE's, especially, often cannot be found due to extreme complexity of calculations. The L-moments exist in a simple fashion even for five parameter complex distributions. The method also allows further investigation like error estimation, finding confidence intervals etc. with significant ease in almost all cases. [6]

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

c/ Common field of application:

L-moments are highly used for data of natural extreme events and natural events as well. For example:

- i. Yearly maximum flood level or water flow at a river site
- ii. Maximum or minimum rainfall in a state over years
- iii. Maximum velocity of cyclones that occur in a year
- iv. Statistical modelling of Tsunami occurrence
- v. Sea tide impact or wave height
- vi. Inflow at a dam
- vii. Heat and cold waves
- viii. Snowmelts in poles or mountains etc.

d/ Probability Weighted Moments (PWM)[1]

Let X_i denote the maximum value of a year i , $i = 1, 2, \dots, n$

Let $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ be the ordered sample observations.

• Probability Weighted Moments (PWM) in population

Probability Weighted Moments (PWM's) are defined by Greenwood et al (1979) as,

$$= \int_0^1 \{x(F)\}^r \{F(x)\}^i \{1 - F(x)\}^j dF$$

In particular, the following two moments are often considered

$$-M_{1,0,i} = \int_0^1 \{x(F)\} \{1 - F(x)\}^i dF = \alpha_i(\text{say}),$$

$$M_{1,i,0} = \int_0^1 \{x(F)\} \{F(x)\}^i dF = \beta_i(\text{say})$$

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

- Probability Weighted Moments in sample

Sample estimates a_i and b_i are unbiased for α_i and β_i respectively.

$$a_r = \frac{1}{n} \sum_{i=1}^{n-r} \frac{\binom{n-i}{r}}{\binom{n-1}{r}} x_{i:n}, \quad r=0,1,2,\dots;$$

$$b_r = \frac{1}{n} \sum_{i=r+1}^n \frac{\binom{i-1}{r}}{\binom{n-1}{r}} x_{i:n}, \quad r=0,1,2,\dots$$

Thus we have,

$$b_0 = \frac{1}{n} \sum_{i=1}^n x_{i:n} = a_0,$$

$$b_1 = \frac{1}{n(n-1)} \sum_{i=2}^n (i-1) x_{i:n},$$

$$b_2 = \frac{1}{n(n-1)(n-2)} \sum_{i=3}^n (i-1)(i-2) x_{i:n}$$

e/ L-moments:[2]

Let X_i denote the maximum gauge value of a site for year i , $i=1,2,\dots,n$

- Population L-moments (λ_r):

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(X_{r-k:n}), \quad r=1,2,\dots$$

where,

$$X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$$

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

Sample L-moments: l_r (Sample L-moments are unbiased estimator of population L-moments, but it is not true for L-moment ratio. The bias in estimating population L-moment ratios by sample L-moment ratios is very small for $n \geq 20$.)

$X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ be the ordered sample observations. Hence,

$$l_1 = \frac{1}{n} \sum_{i=1}^n X_{i:n}$$

$$l_2 = \frac{1}{n(n-1)} \sum_{i>j} (X_{i:n} - X_{j:n})$$

$$l_3 = \frac{1}{n(n-1)(n-2)} \sum_{i>j>k} 2(X_{i:n} - 2X_{j:n} + X_{k:n}) \text{ etc.}$$

and,

$$\hat{\lambda}_r = l_r, r = 1, 2, 3, \dots$$

l_1 = measure of location

$$\hat{\tau} = t = \frac{l_2}{l_1} \text{ (measure of dispersion)}$$

$$\hat{\tau}_3 = t_3 = \frac{l_3}{l_2} \text{ (measure of skewness)}$$

$$\hat{\tau}_4 = t_4 = \frac{l_4}{l_2} \text{ (measure of kurtosis)}$$

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

- L-moments in terms of PWM:

$$l_1 = b_0 = a_0$$

$$l_2 = 2b_1 - b_0 = a_0 - 2a_1$$

$$l_3 = 6b_2 - 6b_1 + b_0 = a_0 - 6a_1 + 6a_2$$

$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 = a_0 - 12a_1 + 30a_2 - 20a_3$$

Chosen Probability Distributions for food frequency analysis and estimate of parameters through L-moments are as given by J.R.M Hosking.

f/ Some common distributions and their parameter estimates by L-moment method:[3]

1/ Normal distribution:

$$\text{P.D.F} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty$$

L-MOMENTS:

$$\hat{\mu} = l_1, \quad \hat{\sigma} = \sqrt{\pi} l_2$$

2/ Two parameter log normal (LN2) distribution:

$$\text{P.D.F} = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{\ln x - \mu}{\sigma}\right]^2}, 0 < x < \infty$$

L-MOMENTS:

$$\hat{\mu} = \ln l_1 - \frac{\hat{\sigma}^2}{2}, \quad \hat{\sigma} = 2 \operatorname{erf}^{-1}\left(\frac{l_2}{l_1}\right), \text{ where } \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$$

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

3/ Gumbel (GEV1) distribution:

$$P.D.F = \frac{1}{\alpha} e^{-\left(\frac{x-\mu}{\alpha}\right)} e^{-e^{-\left(\frac{x-\mu}{\alpha}\right)}}, \quad -\infty < x < \infty$$

L-MOMENTS:

$$\hat{\alpha} = l_2 / \ln 2, \quad \hat{\mu} = l_1 - 0.5772157 \hat{\alpha}$$

4/ Generalized extreme value (GEV2) distribution:

$$P.D.F = \frac{1}{\alpha} \left(1 - k \left(\frac{x-\mu}{\alpha}\right)\right)^{\frac{1}{k}-1} e^{-\left(1 - k \left(\frac{x-\mu}{\alpha}\right)\right)^{\frac{1}{k}}}, \quad \begin{cases} x < \mu + \frac{\alpha}{k}, \text{ if } k > 0 \\ x > \mu + \frac{\alpha}{k}, \text{ if } k < 0 \end{cases}$$

μ is the location parameter, α is the parameter and k is the shape parameter.

L-MOMENTS:

$$\text{Let, } C = \frac{2}{3+t_3} - \frac{\ln 2}{\ln 3},$$

$$\hat{k} = 7.859C + 2.9554C^2, \quad \hat{\alpha} = \frac{l_2 \hat{k}}{\Gamma\left\{\left(1+\hat{k}\right)\left(1-2^{-\hat{k}}\right)\right\}}, \quad \hat{\mu} = l_1 + \frac{\hat{\alpha}}{\hat{k}} \left[\Gamma\left(1+\hat{k}\right) - 1\right]$$

5/ Generalized Logistic distribution:

$$P.D.F = \frac{1}{\alpha} \left[1 - k \left(\frac{x-\varepsilon}{\alpha}\right)\right]^{\left(\frac{1}{k}-1\right)} \left[1 + \left\{1 - k \left(\frac{x-\varepsilon}{\alpha}\right)\right\}^{\frac{1}{k}}\right]^{-2} \quad \begin{cases} x \leq \varepsilon + \frac{\alpha}{k}, \text{ if } k > 0 \\ x \geq \varepsilon + \frac{\alpha}{k}, \text{ if } k < 0 \end{cases}$$

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

L-MOMENTS:

$$\hat{k} = -t_3, \quad \hat{\alpha} = \frac{l_2}{\sqrt{|1+\hat{k}||1-\hat{k}|}}, \quad \hat{\varepsilon} = l_2 + \frac{l_2 - \hat{\alpha}}{\hat{k}}$$

6/ Generalized Pareto distribution

$$P.D.F = \frac{1}{\alpha} \left[1 - \frac{k}{\alpha} (x - \varepsilon) \right]^{\frac{1}{k} - 1}, \quad \begin{cases} x \leq \varepsilon + \frac{\alpha}{k}, & \text{if } k > 0 \\ x \geq \varepsilon, & \text{if } k \leq 0 \end{cases}$$

L-MOMENTS:

$$\hat{k} = \frac{1 - 3t_3}{1 + t_3}, \quad \hat{\alpha} = l_2 (1 + \hat{k})(2 + \hat{k}), \quad \hat{\varepsilon} = l_1 - l_2 (2 + \hat{k})$$

g/ Our main objective is fitting a distribution by estimation of labelling parameters from observed data by method of L-moments. Now, how to check the goodness of fit?

There is no uniquely best method available for this purpose. We mention two common methods here.

a) For upper (or lower) tail-based fit one may consider the upper (or lower) tail-based method of D-Index.

Let, $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the ordered sample observations.

The initial estimate of probability of non-exceedance (plotting position)[7] for all distributions is taken (as given by Weibull) to be $p_i = \frac{i}{n+1}$. Let, $F(o_i) = p_i$

$$\text{Then, D-Index upper} = \frac{\sum_{i=n-5}^n |o_i - x_{(i)}|}{\bar{x}}$$

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

$$\text{and, D-Index lower} = \frac{\sum_{i=1}^6 |o_i - x_{(i)}|}{\bar{x}}$$

b) In the case of our problem, we consider the most common method for ranking the good fit distributions. **This is root mean square error (RMSE)**, where,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (o_i - x_{(i)})^2}{n}}$$

c) Test for goodness of fit: Pearsonian frequency χ^2 test :

We test, H: Data fits the given distribution.

Against, K: Data does not fit the given distribution, at $100\alpha\%$ level of significance

The test statistic is

$$\chi_w^2 = \frac{\sum_{i=1}^k (f_i - e_i)^2}{e_i} \quad \square \quad \chi^2$$

distribution with $(k-m-1)$ d.f. under H.

Where, k = number of classes,

f_i = observed frequency in the i^{th} class,

e_i = expected frequency in the i^{th} class

m = number of parameters estimated in the distribution

The critical region is

$\omega: \{\chi_w^2 > \chi_{\alpha, k-m-1}^2\}$, where, $\chi_{\alpha, k-m-1}^2$ is the upper $100\alpha\%$ cut off point.

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

h/ Advantages:

- i. L-moments are simple and more robust than conventional moments.
- ii. L-moments are far more meaningful when dealing with outliers in data than conventional moments.
- iii. L-moments are used as summary statistics in extreme value theory (EVT). Though they are not resistant statistics and a single extreme can throw them off, nevertheless, as they are linear and properly weighted, they are much less affected by extreme values than conventional moments.
- iv. Existence of L-moments only requires the R.V. to have finite mean. Hence the L-moments exist even if higher conventional moments do not exist, e.g. Student's t-distribution with lower d.f.. A finite variance is required in addition in order for s.e. of estimates to exist.
- v. Trimmed L-moments are generalization of L-moments that give zero weight to extreme observations. They are more robust to the presence of outliers.
- vi. a) Fitting a distribution to data from nature more often fails when parameters are estimated by MOM or MLE in comparison to L-moments. It has been a surprising observation though the estimators may not be unbiased.
b) It is very frequently observed that RMSE and D-index have lower values for L-moment method corresponding to the different plotting point positions given in different texts.

i/ Disadvantage:

As per literature, one disadvantage of L-moment ratio for estimation is their typically smaller sensitivity although lesser sensitivity to modelling errors can sometimes be a potential advantage.

On Linear Moments (By J.R.M. Hosking) in the Context of Frequency Analysis

j/ References:

- 1) Hosking, J. R. M. (1986). The theory of probability weighted moments. *Research Report RC12210*, IBM Research Division, Yorktown Heights, N.Y.
- 2) Hosking, J. R. M. (1989). [Some theoretical results concerning L-moments](#). *Research Report RC14492*, IBM Research Division, Yorktown Heights, N.Y.
- 3) Hosking, J. R. M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B*, **52**, 105-124.
- 4) Hosking, J. R. M. (1992). Moments or L-moments? An example comparing two measures of distributional shape. *The American Statistician*, **46**, 186-189.
- 5) Hosking, J. R. M., and Wallis, J. R. (1993). Some statistics useful in regional frequency analysis. *Water Resources Research*, **29**, 271-281. Correction: *Water Resources Research*, **31** (1995), 251.
- 6) Hosking, J. R. M. (1995). The use of L-moments in the analysis of censored data. In *Recent advances in life-testing and reliability*, ed. N. Balakrishnan, 545-564. Boca Raton, Fla.: CRC Press.
- 7) Hosking, J. R. M., and Wallis, J. R. (1995). A comparison of unbiased and plotting-position estimators of L-moments. *Water Resources Research*, **31**, 2019-2025.
- 8) Hosking, J. R. M. (1998). L-moments. In *Encyclopedia of statistical sciences, update vol. 2*, ed. S. Kotz, C. Read and D. L. Banks, pp. 357-362. New York: Wiley.
- 9) Hosking, J., Bonti, G., and Siegel, D. (2000). Beyond the lognormal. *Risk*, **13:5** (May 2000), 59-62.

Articles from STUDENTS

Randomness

Subharanjan Mandal (3rd Year)

What does it mean to be random? Can anything really be random? What is the most random thing ever? If something is unpredictable and contains no recognizable pattern, we call it random. Whatever will happen tomorrow is not random. Some things will happen for sure, like the sun rising in the east. Even human behavior is predictable to a certain extent. Let us take, as an example, a coin. Coin flips and rolling dice are not exactly random; they are random because we are lazy. If we know the initial conditions, the exact force applied for a particular flip or roll, we can theoretically determine the outcome before it happens. Researchers have developed coin flipping robots that can precisely control the flip and get the result that they desire “one hundred percent” of the time.

So the question is, is there anything that we cannot predict even if we know everything? How can you be sure that there isn't any pattern in the process that we are looking at? It is difficult to identify a random process. It is easier to be certain that a process is not random than identifying a truly random process. Despite this trick, nowadays, we call apparent non-random things random, like randomly running into a friend at a restaurant or calling unknown guys at a party “randos”. They aren't “randos” in the mathematical sense; they knew about the party and were in a perfect mood to go into the party, which seems very much predictable. In the 1970s, a student of MIT published a research paper that popularized the word “random” for simply meaning ‘strange’. But being strange does not mean that it does not have any identifiable pattern. Sometimes, we call certain predictable things random just because we are too lazy to find a reasonable cause. Coins and dice are sensitive to their initial conditions and may exert randomness to a certain extent but making a coin or a die perfectly unbiased is extremely difficult. A biased coin means that it is predictable to some extent. In order to get a nearly unbiased outcome, we can catch the coin after flipping. This is because letting the coin fall on the ground means more spins. Researchers have found that as a coin spins, larger biases come into play like its mass, center of gravity, and the shape of its edge. So if the outcome of flipping a coin can be predicted by calculating the initial conditions, why don't we do it often? That is because it is extremely difficult to determine every single condition. Even an extremely small change in the initial conditions will add on and lead to chaos, which will

Randomness

result in a completely different outcome.

In order to find a perfectly random event, we need a system that is determined by nothing. But where can we find such an event? Well, luckily there is an answer and the answer is "Quantum". Quantum mechanics describe the properties of quantum-sized things as "probabilities", not because we do not know enough to predict, but there is nothing to predict. Whether or not a particular atom will decay, or, the spin of an electron, can only be known once we observe them. They are determined by "a deep-seeded randomness woven into the universe itself". This was quite unbelievable for Einstein. He said, "God played dice with the universe". Let us take, for example, quantum entanglement. It says that two particles, separated from one another, will show similar outcomes even when they are placed at extremely large distances. However, experiments have found that what one machine will find in one particle determines what the other machine will find in the other particle. So all of it lies in the measurement, that is, when we look, the chances are determined. Experiments have found that the chance of seeing a particular quantum quality does not exist beforehand; they exist when we observe them.

So, if you are feeling low or think yourself to be too predictable, just remember that you are made up of billions of such "dice". The chances of you, coming into consciousness, were astronomically small and still, you are here. The particles, constituting you, travelled a long length of time, met each other by "chance" to enjoy consciousness for a brief period just to disintegrate and continue their journey into the void till the end of time.

Articles from STUDENTS

Data Visualization : Glimpses of Today's World in the Light of Data

Adrija Saha, Shrayan Roy (2nd Year)

Introduction :

Pictures are storytellers. They can convey a thousand words. To many people numbers are mundane and tedious. But as students of statistics, we present the story of these numbers using pictures. John W. Tukey once said,

“ The greatest value of a picture is when it forces us to notice what we never expected to see”.

To understand the data, we need to hear what it tries to tell us. The beauty of data visualization lies in the fact that it beautifully depicts the whole untold story in front of our eyes and often it becomes a vital step in producing a fruitful decision.

Here, we will try to show you the story of slow human progress using data and data visualization tools. It will be interesting to know the changes in the world since 1970 till date. We, the common human beings, are very much in disagreement to agree to what is actually happening, instead we prefer to celebrate our pre-conceived idea than to upgrade it. We will see how these things contribute to create a difference between what we think about the world and how the world really is.

Here, we will use the motivating dataset of R 'Gapminder', which is produced by compiling thousands of spreadsheets from 'Gapminder Foundation'.

Our dataset :

It is under library "dslabs". We can get access to it writing some lines of code and we can also see the structure of this dataset. We have 9 variables in our dataset and we have data from the year 1960 to 2010.

Data Visualization: Glimpses of Today's World in The Light of Data

```
> library(tidyverse)
> library(dslabs)
> data(gapminder)
> str(gapminder)
'data.frame':  10545 obs. of  9 variables:
 $ country      : Factor w/ 185 levels "Albania"
 $ year        : int  1960 1960 1960 1960 1960
 $ infant_mortality: num  115.4 148.2 208 NA 59.9 .
 $ life_expectancy: num  62.9 47.5 36 63 65.4 ...
 $ fertility    : num  6.19 7.65 7.32 4.43 3.11
 $ population   : num  1636054 11124892 5270844
 $ gdp         : num  NA 1.38e+10 NA NA 1.08e+1
 $ continent   : Factor w/ 5 levels "Africa","A
 $ region      : Factor w/ 22 levels "Australia
```

(Code Snippet in R)

Tools :

We will use R software as our analysis tool and also the 'Gapminder' App for further visual analysis. The "ggplot2" package will be helpful in our analysis.

Our Survey :

We have conducted a small survey and asked some questions to know what people around us think about the world. We have circulated those questions via google form mostly among college going students. We have collected 237 responses. We made some questions of our own that are very much relevant to our topic and also got references from some books, advices from our respected professors.

Some pre-requisites :

1. Pictures are story tellers: In the picture(Diagram : O1) below. You can understand how the life style has improved as we move from Economic level 1 to level 4. Level 1 people have to struggle every day for food and shelter.

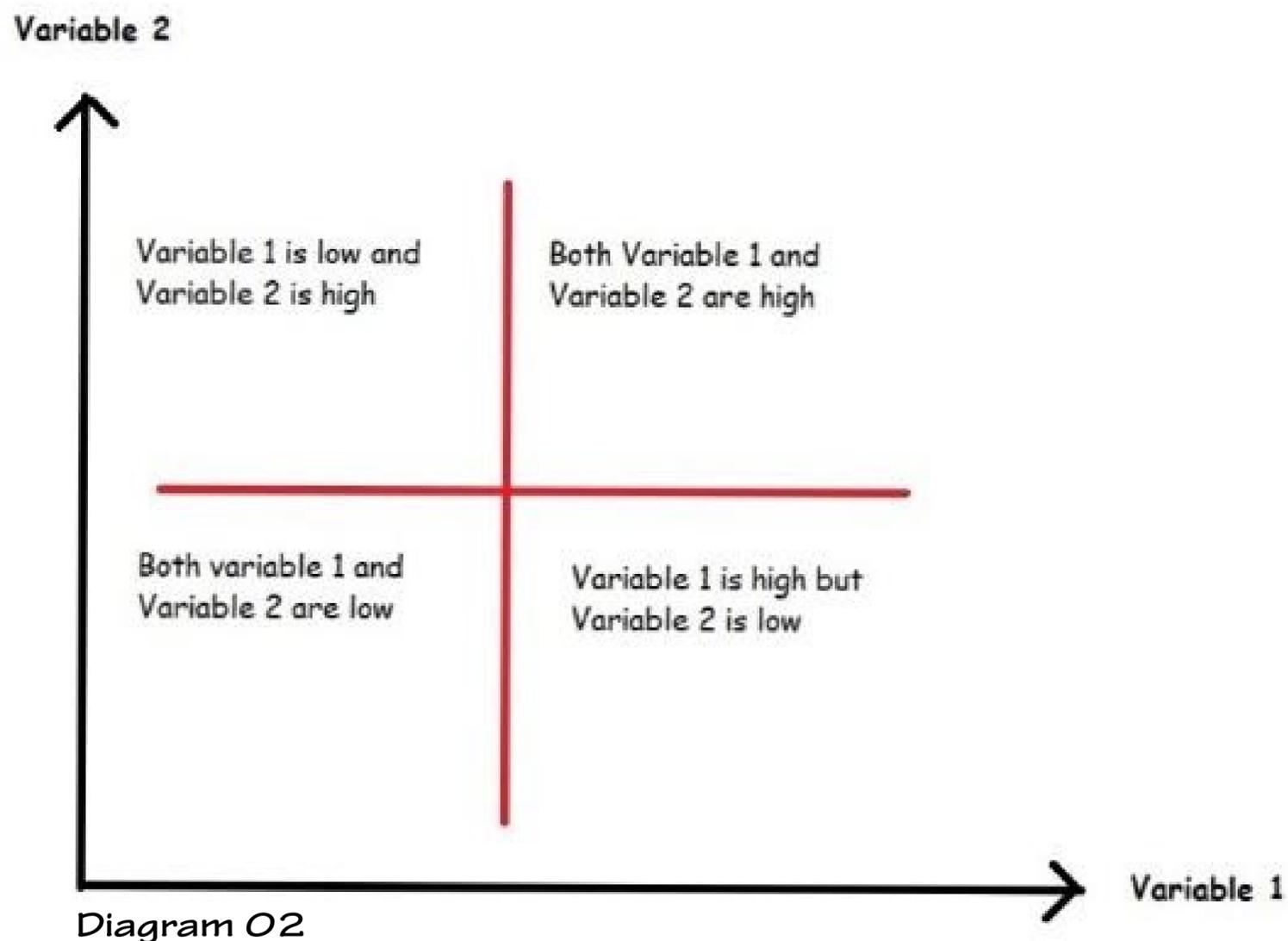
Different Economic levels



Diagram O1 (Source : Dollar Street, Gapminder)

Data Visualization: Glimpses of Today's World in The Light of Data

Understanding plots: In the picture(Diagram: O2) here. You can understand the implementations of points in different portions of the plots. Keep it in mind, it will be useful for later discussion.



Discussions :

Now, we will start our main discussion by analysing the questions we have asked in our survey one by one.

Question Number 1 :

Where do you think the majority of the Population of the World live in?

a. Low-income country **b.** Middle income country **c.** High income country

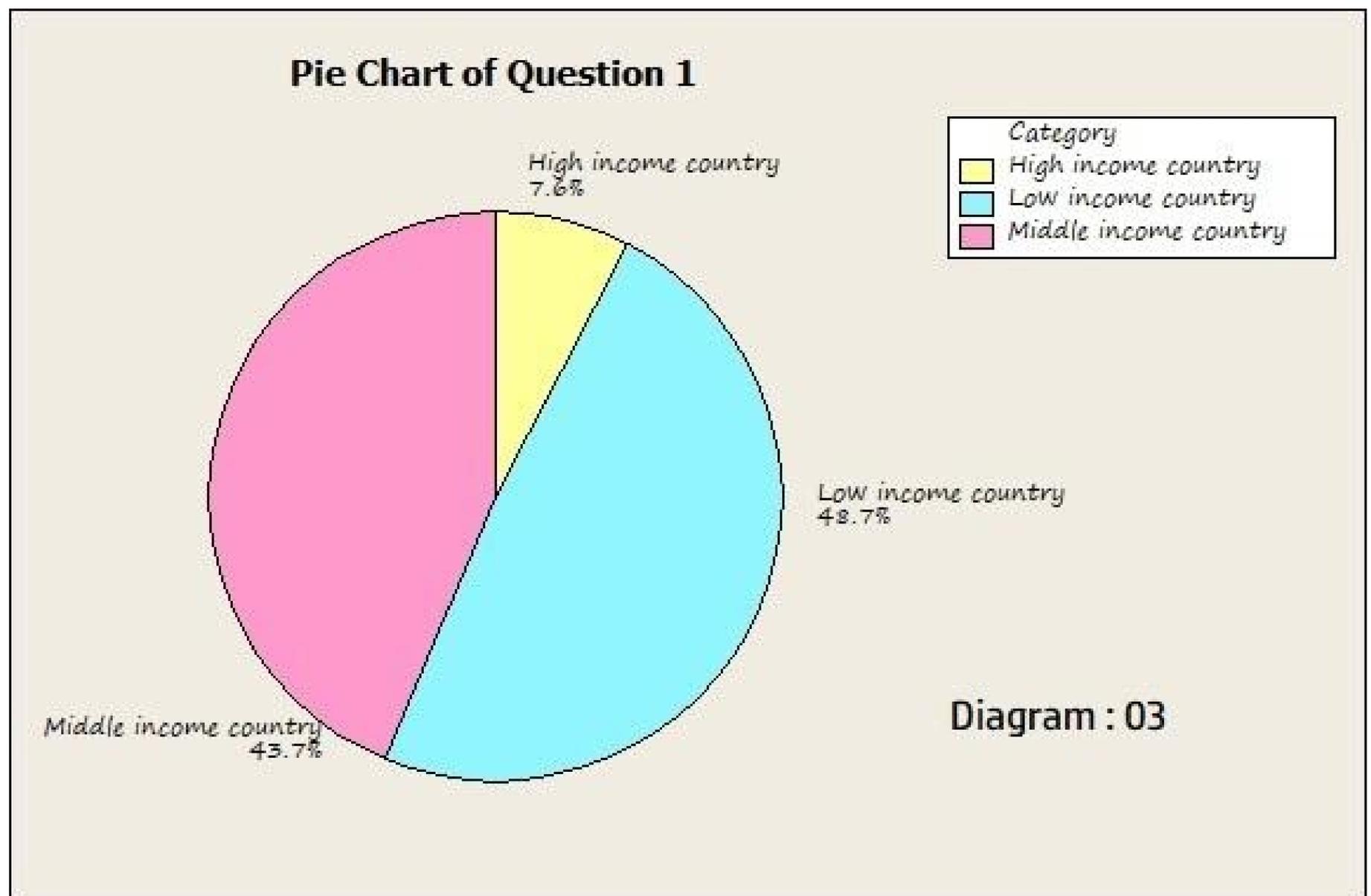
Pie Chart of the Survey : Below we represent the results of this survey using Pie chart.(Diagram : O3)

As we have discussed earlier, we have a negative instinct that all things around the world are getting worse. But this is wrong. (Well, we must say that all things are not improving.) This above question is proposed by Gapminder Foundation [1].

Data Visualization: Glimpses of Today's World in The Light of Data

The Correct answer is 'Middle Income Country'. The majority of people live neither in low-income countries nor in high income countries, but in middle income countries. But when you see the percentage who gave correct answer, it is only 43.7%, which is quite low. Actually, many of us think that two types of people live in the world - Poor and Rich. This old mindset leads us to think that things are getting worse and worse. There is no way to make these things normal(good). But Why are we not aware of the fact that the poverty is being eliminated in the last 70 years at a consistent rate? Let Data tell the story.

We all have a pre conceived notion that the world is divided into two parts. The 'Developing Country' (Asian countries, Sub-Saharan African countries etc) and the 'Developed country'(American countries, European Countries). But the concept of 'Developing country' and 'Developed Country' is no longer valid. This concept was perhaps valid 50-60 years ago, but not today. Let's try to understand this using Graphs. Good health and economy lead a country to be better. So, let's make a scatter plot of 'Life Expectancy' and 'GDP per capita' for the years 1970 and 2010.



Data Visualization: Glimpses of Today's World in The Light of Data

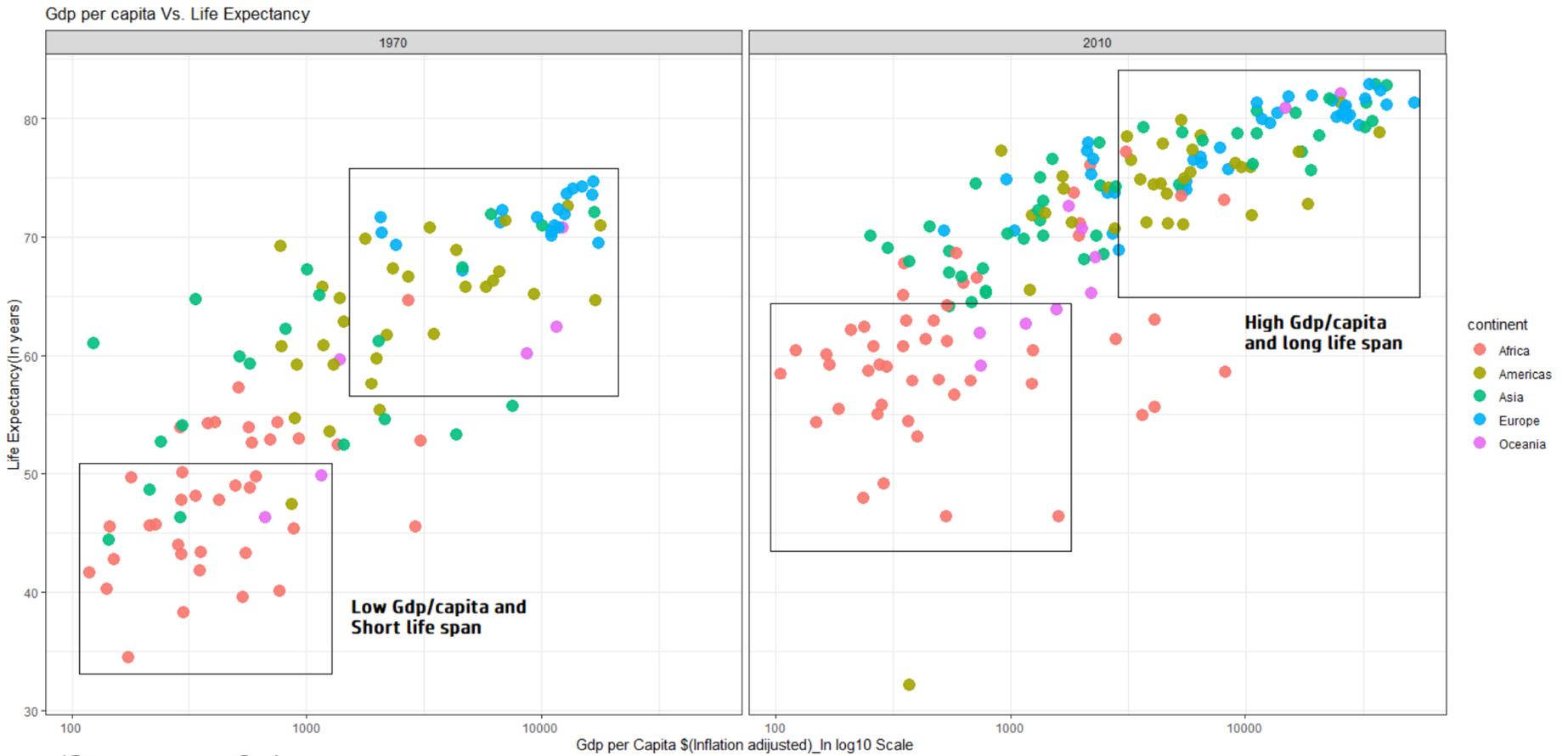


Diagram 04

From the scatterplots (Diagram : 04), we see that in 1970 there was a clear division in the world, this division was 'Developed' and 'Developing' country. But in 2010 there was no division in the world any more. Most of the countries have succeeded in achieving high life expectancy and high income. Though there are some countries in the lower left portion of the graph meaning low life expectancy and low income.

So, the human kind is succeeding in achieving a better world and this silent progress was unknown by many of us. We make decisions on the basis of old mindset which was valid 50-60 years ago. But the world has changed very much since then. The data is telling us the story of this huge advancement of human beings.

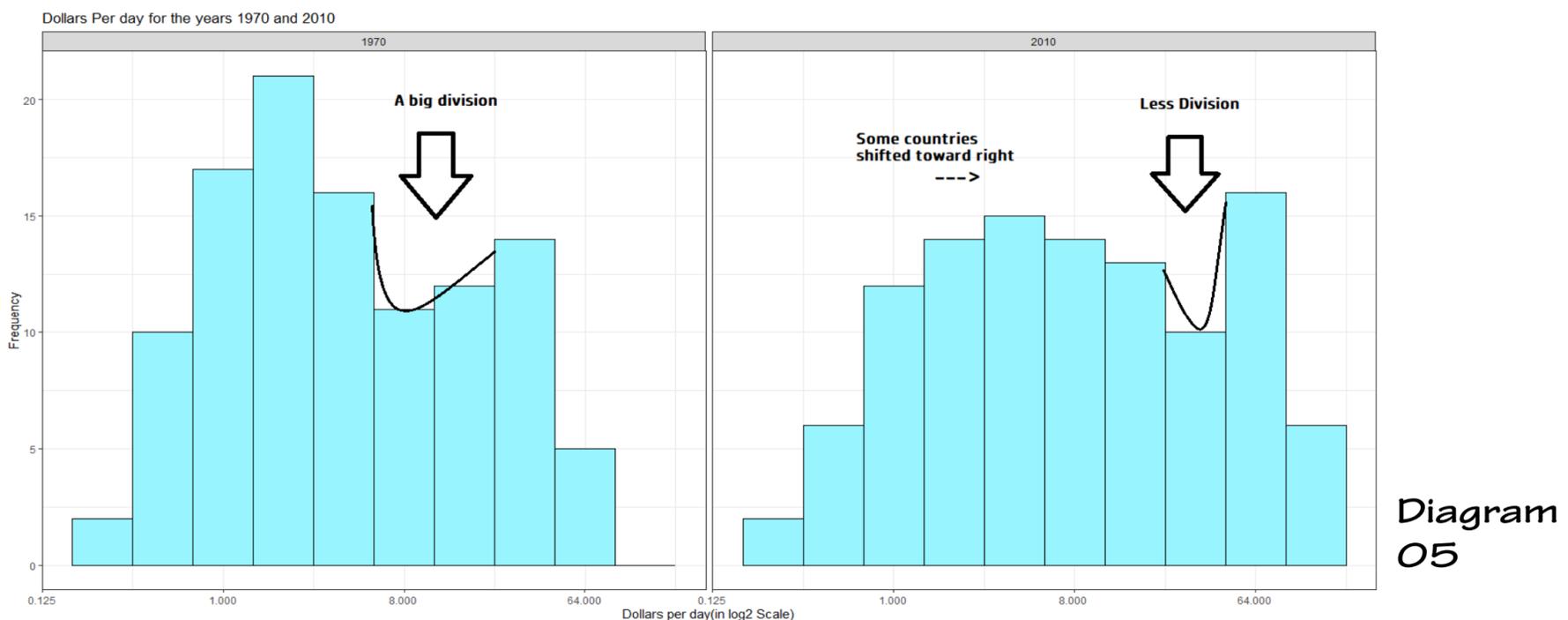


Diagram 05

Data Visualization: Glimpses of Today's World in The Light of Data

Let's continue with our discussion. We have created a new variable called 'Dollars per day' in our dataset and we have created the histogram of the dollars per day for the years 1970 and 2010 (for those countries whose data is available in both years). We can see that there was a hump in the graph (Diagram : 05) of 1970. Why? As we can see in the scatterplot that the world was very much divided into two parts in the year 1970. After that, independence of many Asian as well as African countries lead to five-year planning etc, which had a deep impact on the health system as well as income of the countries. Let's try to understand this change in different ways. We create another histogram (Diagram : 06)

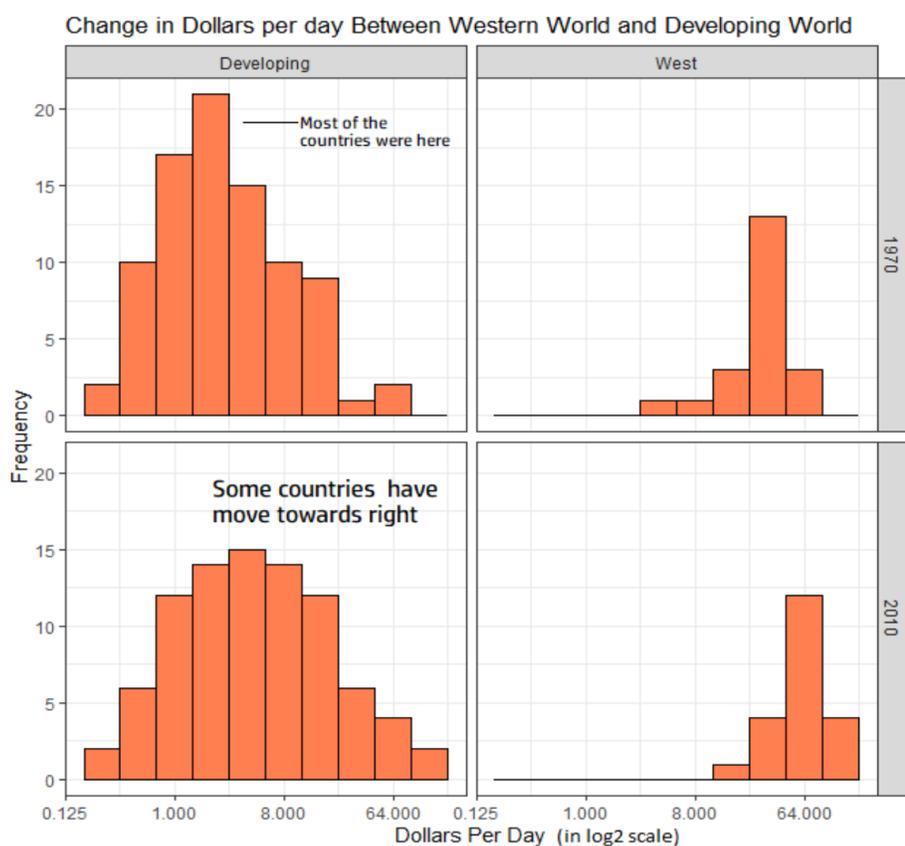


Diagram 06

We see that most of the countries in developing world have shifted towards right meaning those countries have improved very much in these 40 years. Also, we see that the countries in western world (Regions In our dataset: "Western Europe", "Northern Europe", "Southern Europe", "Northern America", "Australia and New Zealand") have also shifted towards right, but the change is not much as compared to developing countries.

Now, this graph is not enough, because it does not reflect which region is doing better. So, to understand the change in the region level we make a box plot of fold increase of dollars per day between 1970 and 2010 (Diagram : 07).

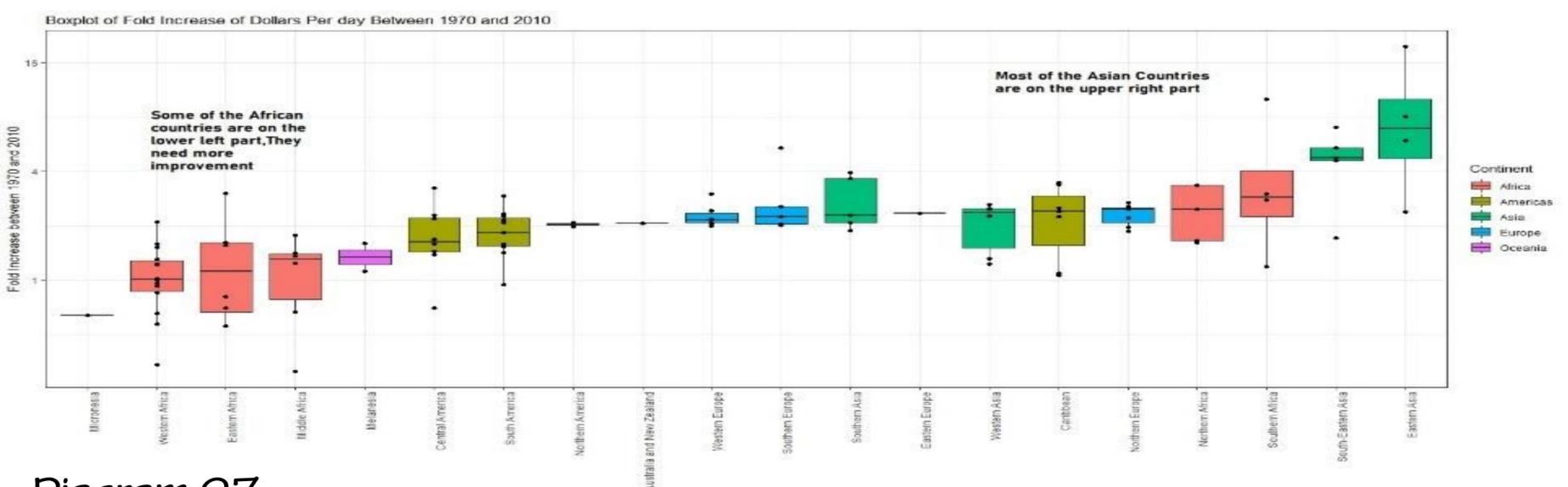


Diagram 07

Data Visualization: Glimpses of Today's World in The Light of Data

We see that most of the Asian regions are in the upper portion and most of the European regions are in the middle portion. Notice that Eastern Asia (Japan, China, North & South Korea etc) have achieved the biggest improvements. Also, while Southern Africa, Northern Africa are doing well, Western Africa and Eastern Africa are still on the lower side. It means that the fold increase of dollars per day between 1970 and 2010 is relatively high for the countries of so-called 'Developing world' and we need more improvement for African countries. We can also see the spread of points within a region in the boxplot.

So, through different graphs we see that it is not justified to use the words 'Rich' and 'Poor' or 'Developed' and 'Developing' countries or 'We' and 'They' any more.

The people living in middle income countries don't exist in the divided mind set but they definitely exist. About 75% of humanity lives in this category. These people earn about 10 dollars per day. Combining middle (level 2 and 3) and high-income countries (level 4), that makes 91% of humanity and the rest fall in low-income countries (about 9%). These people living in low-income countries (level 1) are tremendous strugglers. They fight every day for a living. Many of them are in extreme poverty (it is a hypothetical line drawn by the economist. They earn about 1 dollar per day. There is an uncertainty whether they will get food or not). They know how much effort they need to go to the next level (middle income).

Today the vast majority of the people are spread out in the middle-income countries, with the same range of standard of living as people had in Western Europe and North America in 1950s. You should know that human history started with everyone on level 1. For 100,000 years nobody went to level 2 and most children died before they became parents. In 1800, estimates say that most of the people were on level 1 (Extreme poverty).

Imagine how beautiful is this change! Another interesting thing is that United States is always on the money side. But things have changed, and they understand that they should invest more on health, which have increased the life expectancy of USA. The catchup of Asian countries like Japan, China, India etc and of African countries are inevitable.

Data Visualization: Glimpses of Today's World in The Light of Data

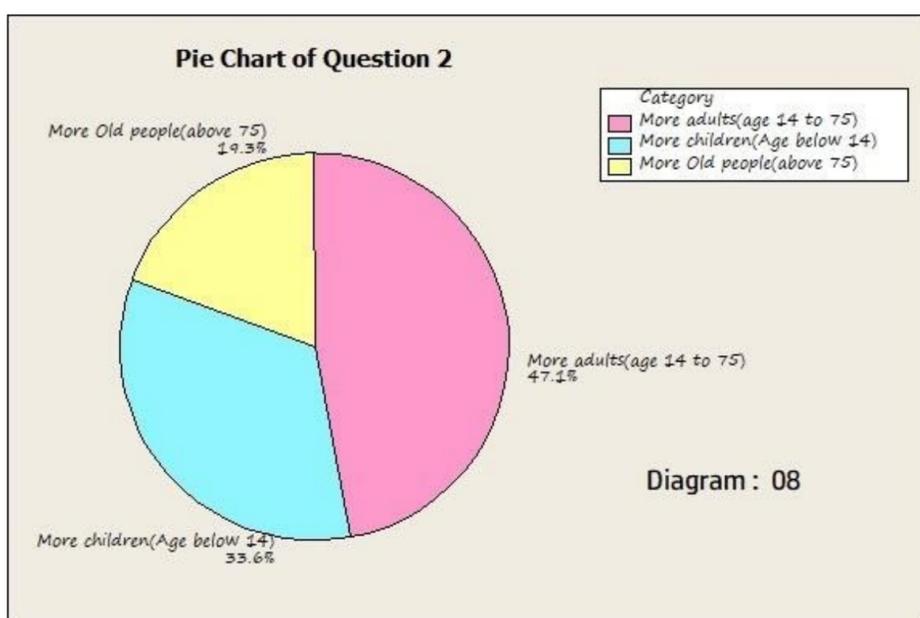
Question Number 2 :

It is predicted that by 2100, the world population will have increased by 4 billion

people. Which of the following in your opinion will be the major contributor?

- a.** More children(Age below 14) **b.** More adults(age 14 to 75) **c.** More Old People(above 75)

Pie Chart of the Survey : We represent the results of this survey using Pie chart.(Diagram : 08)



This question was also proposed by Gapminder Foundation. It was basically placed to know what actually common people think about the very much discussed hot topic 'Population growth of world in near future'. The correct answer would be—According to UN prediction, the increase in population of world by 2100 (by

more 4 billion people around) would be mainly for MORE ADULTS(age 15-74 years). In our survey 47.1% have given us the correct answer. Now, to understand the actual reason behind this growth, let us first visualize the UN forecast about World's Population Growth using a graph.

From this graph(Diagram: 09), it is quite clear that the increase in population growth in this one century would not be for more children, but for more adults. Look at the children line in the graph (i.e., "5-14 years old" and "Under 5s" in the graph) it already got flat i.e., the number of children(in a year) already became marginal.

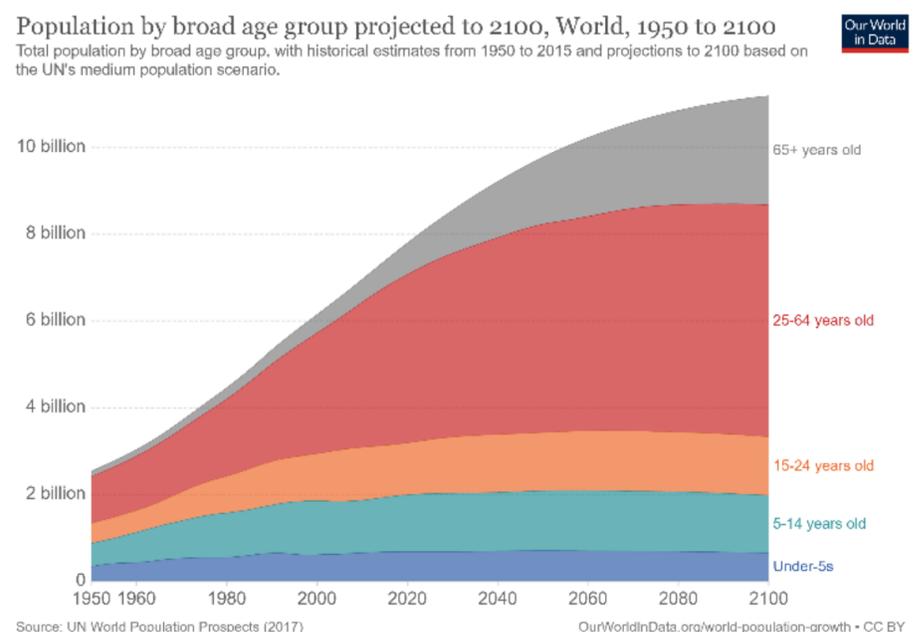


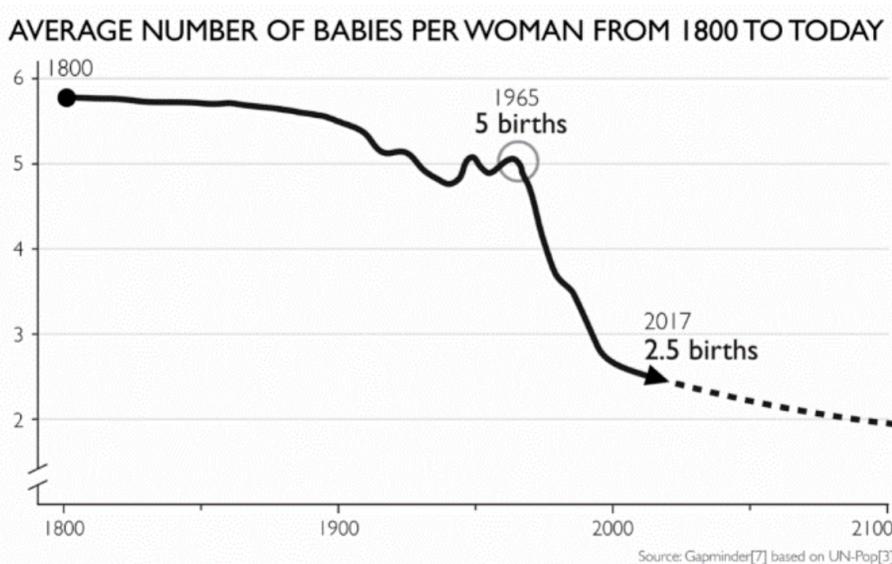
Diagram 09 (Source: Our World in Data)

Experts panic about the drastic growth in world population and they say that Population growth must stop right now! We all know the increasing world population is creating problem both economically and by damaging Natural Environment.

Data Visualization: Glimpses of Today's World in The Light of Data

But, let us try to focus what Data says. Is it really feasible to stop the increase in World Population completely right today?

If we look about the annual reports of number of births, we will surely appreciate that the annual number of births have already stopped increasing and experts say that world has already reached 'Peak child'(i.e., Number of children will continue to be at 2 billion for rest of the years) at around 2000(A very significant year to population scientist). It is all due to the falling fertility rate.

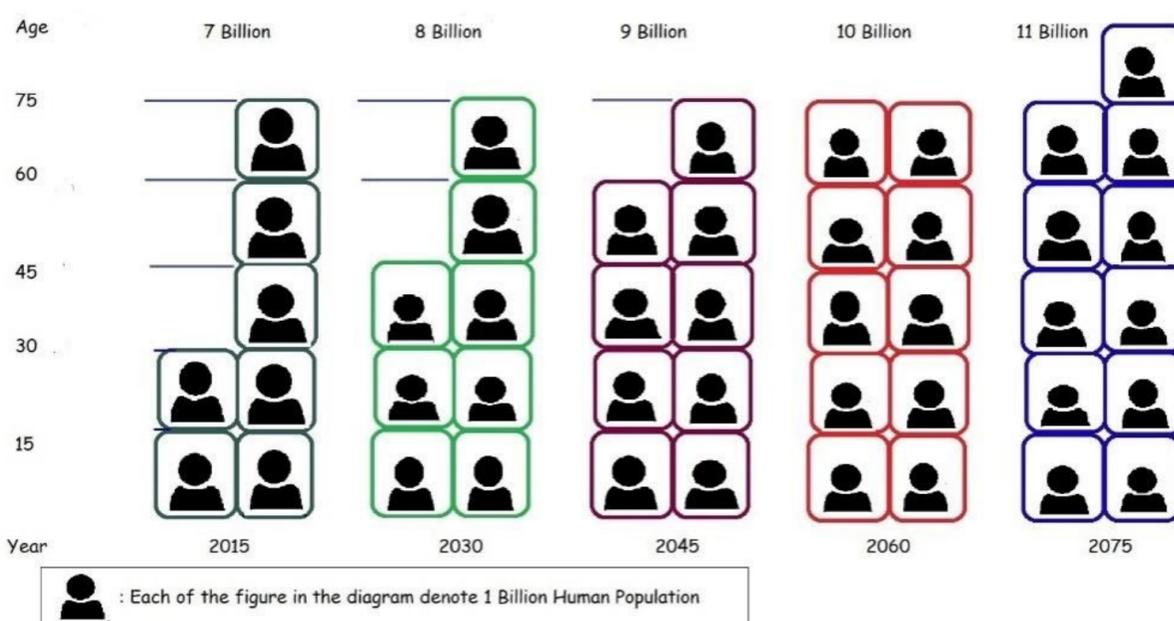


In the adjacent line diagram (Diagram : 10), we can see that at around 1965 the average number of babies per woman was around 5, probably because high infant mortality rate(You can say as an insurance to infant mortality). As time passed with the upgradation of people's economic level, education,

Diagram 10 (Source: Gapminder)

improvement in medical science, this child mortality rate drops down and parents decide to have a smaller number of children. Experts say, at around 2000 world has already achieved its 'peak child', It is also evident from the tendency of the Number of Babies per woman in the graph becoming marginal.

Now, we will straight away try to answer the question that we were looking for. Let's make a diagram for better understanding.



Each block in the diagram (Diagram: 11) represents 1 billion people. In the above diagram, we can see that in 2015, 2 billion people are of age group 0-15 years, 2 billion people are of age group 15-30

Diagram 11

Data Visualization: Glimpses of Today's World in The Light of Data

years old and 1 billion people are each in 30-45 years, 45-60 years, 60-75 years age group.

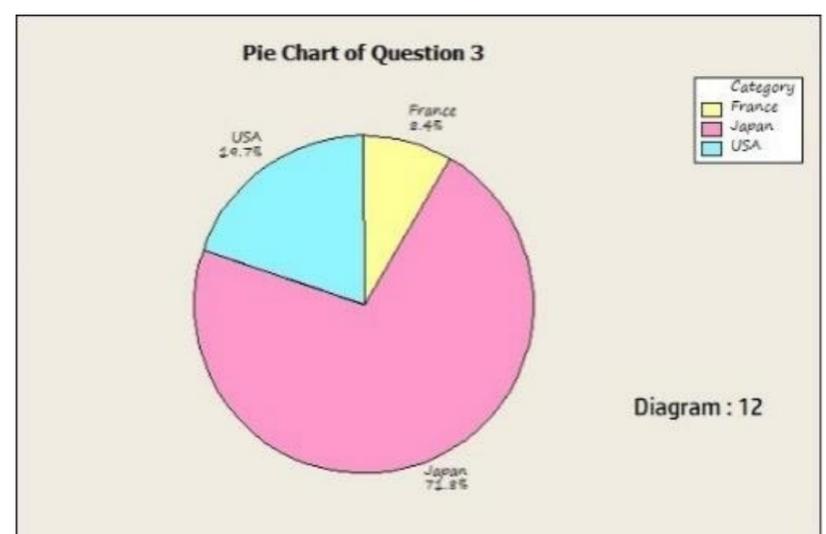
Now as the world has already reached 'peak child', we may assume that there will be 2 billion new born (aged 0 to 15 years old) in 2030 also. On the other hand, today's 0-15 years old people will grow up to 15-30 years, today's 15-30 years old people will grow up to 30-45 years old. Now compare it to the diagram of 2015. Without increasing number of new born, we have already got 1 billion more people in the age group of 30-45 years old! So, from where will these 1 billion more people come in 2030? . It will come from today's young people getting older then. Same pattern will repeat itself for next three generations. The 2 billion people of age group 30-45 (in 2030) will grow up to 45-60 years old in 2045 increasing one billion more adult in totality. Similarly, in 2060, these 45-60 years old people will become 60-75 years and will add 1 billion more in the adult group. But after 2060, you will understand looking at the diagram that each generation of two billion people will be replaced by two billion other people stopping the growth of population further.(you can say we will get the fruit of reaching 'peak child' in 2000 from 2060 onwards in practical).

So, it is clear that basically the large population increase that is going to happen in future is neither because of more children nor for increase in life expectancy of human beings, but will take place for a greater number of adults and this increase is in fact unavoidable! This process of increase in the number of adults by existing young aged people that is responsible for increase in population, can be called as 'Fill up effect'. Though in fact, UN experts predict that by 2100, life expectancy of people will also increase roughly by 11 years, adding one more billion people to the total population and thus the total population will be around 11 billion then.(This is a rough picture for understanding in a simpler way. Note that the figures are not exact but rounded off for better depiction.)

Question Number 3 :

The world's life expectancy is about 72 years. Now, which country in your opinion has the highest life expectancy among the following?

a. USA **b.** Japan **c.** France



Data Visualization: Glimpses of Today's World in The Light of Data

Pie Chart of the Survey: We represent the results of this survey using Pie chart.(Diagram : 12)

The objective of placing this question is to see the generalization Instinct. We as a human being make generalizations of some facts. If you see the question, it is about the life expectancies of different countries. See the options, USA(United States of America) is an American country, France is a European country and Japan is an Asian country. As we have discussed in explaining Question 1, that the Asian and African countries made a huge progress in last 50-60 years and this change is inevitable. The change for Asian countries is huge. (Though many countries like-Afghanistan are in level 1.) The Answer is Japan. Around 71.8% gave the correct answer, which is very high. We will try to explain why so many people gave correct answer.

First, let us discuss why the countries United States, France have lower life expectancy compared to Japan. Most of us think that the American countries and European countries were characterized by 'long life expectancy' and 'small families', while the Asian countries, African countries were characterized by 'low life expectancy' and 'large families'. But things are changing. For this, consider the scatterplot of 'Life Expectancy' and 'Fertility Rate' for the years 1970 and 2010 (Diagram : 13).

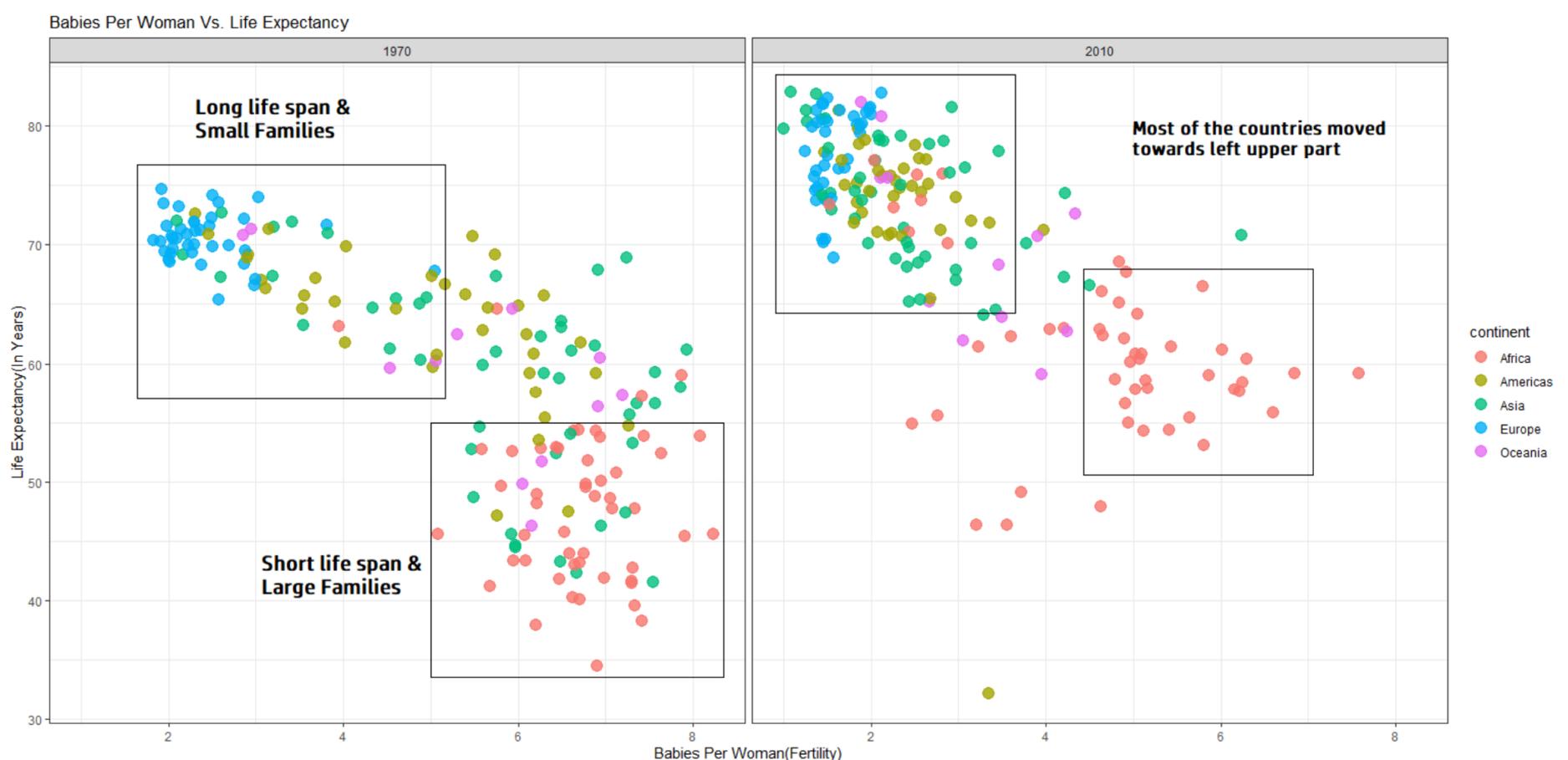


Diagram 13

Data Visualization: Glimpses of Today's World in The Light of Data

We see that most of the Asian countries and African countries in 1970 were on the upper left portion of the graph and most of the European and American countries were on the lower right portion of the graph. That concept of division of the world on the basis of life expectancy and family size was valid in 1970. But see in 2010. Most of the Asian countries, even some of the African countries have moved towards lower right part of the graph from the upper left part of the graph. Isn't it interesting? So, say "Good Bye" to those old concepts, they are no longer valid. These new trends of world economy and health system improvements many of us didn't know about. But they have materialized. Japan is one of the brightest examples of this. It has become a veritable name among Asia's fastest grown countries. But how did this change happen?

Most of the Asian countries were under foreign domination in 1800. Being under foreign domination, the development was not possible for those countries. For example, India was under British dominance for almost 200 years. There was no chance for economic as well as social development for India. But Japan was exception. Japan maintained its national sovereignty. That's why the speedy development of Japan was inevitable. While encountering the first question we have already shown Eastern Asia has highest improvement between 1970 and 2010 and Japan is an Eastern Asian country. Now, Using Time series graphs you can see the change more clearly.(Diagram 14 and 15)

Change In Life Expectancy Over Years 1960-2010

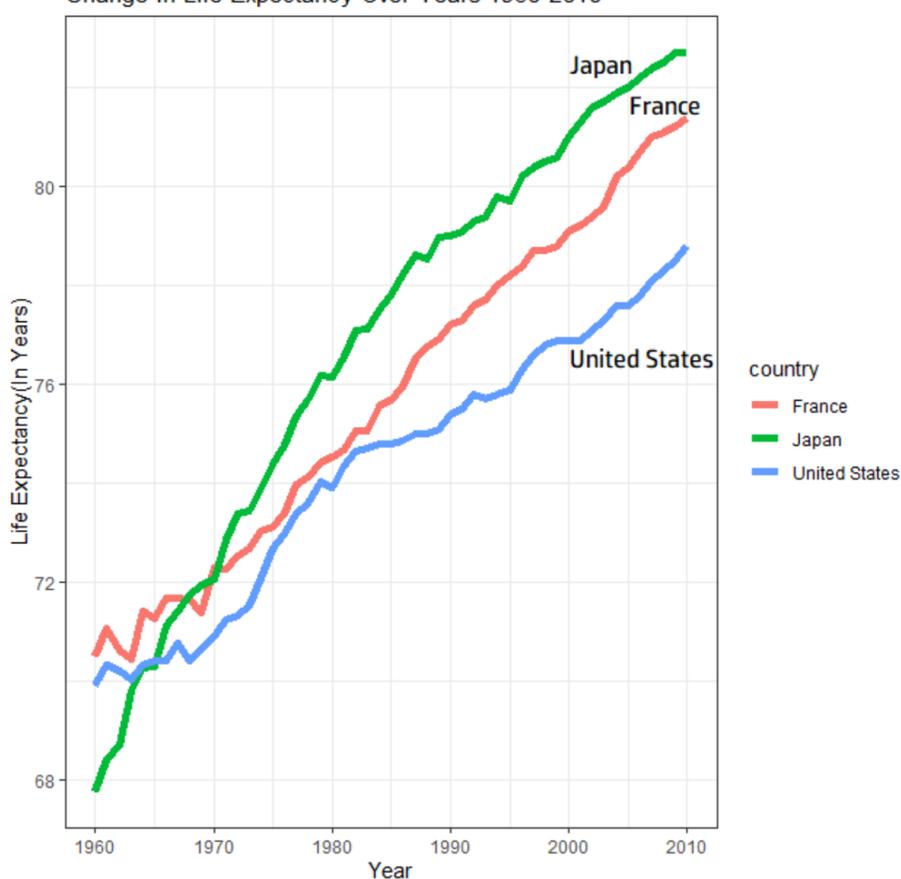


Diagram 14

Change In Life Expectancy and Gdp per Capita Of Japan Over years 1960-2010

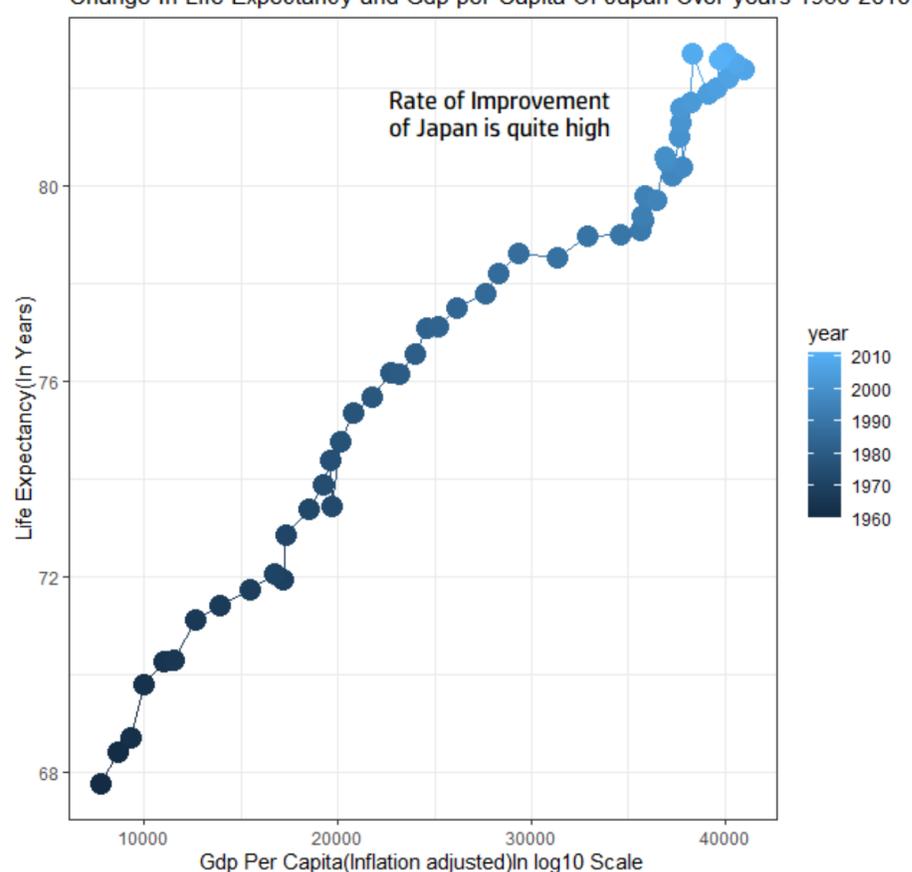


Diagram 15

Data Visualization: Glimpses of Today's World in The Light of Data

From the graph(Diagram: 14)we can see that how Japan has increased its Life Expectancy over the years. In 1960, life expectancy of Japan was much less than that of USA and France. Though each of them has increased its life expectancy over the years, But the rate of increase in life expectancy for Japan is very much significant.

"Doesn't matter if the cat is black or white as long as it catches mice"

In Diagram 15, we have plotted GDP per capita and life expectancy of Japan over the years (using colour scale), you can see how rapidly the points move towards upper right portion of the graph.

Now, let us discuss another beautiful aspect, the life expectancy of a country. You often hear *"Numbers can speak"*. Yes, they can speak. People find statistics boring, as it contains boring numbers, "which has no meaning". But let us show you the most beautiful thing. The life expectancy of a country is not just a number, it shows how the country has improved its health facility, food nutrition, economy etc, how it has decreased its violence. This measure takes the temperature of whole society. A small number represents so many things. That's why we take you through the history of development of Japan. In 1945 there was a certain fall in life expectancy of Japan due to "Hiroshima Nagasaki bombing".

Always remember *"Numbers are Boring, people are interesting"*. We will see more of it in the next question.

Now, many people gave correct answer to this question. Let us try to explain this. In 2020 the outbreak of the pandemic 'Covid-19' made many people realize that the news frontline countries like- Unites States, France, Britain, etc failed to manage the outbreak. That resulted into death of many people in these countries. Whereas, countries like Japan, Singapore succeeded to avoid the dangerousness of the pandemic. This news was shown many times . It may be a reason of giving correct answer. (Note that, since, we have data till 2010 in "gapminder" dataset. That's why we are using time series plots up to year 2010, though in 2020 also, Japan has high life expectancy than France and USA).

Question Number 4 :

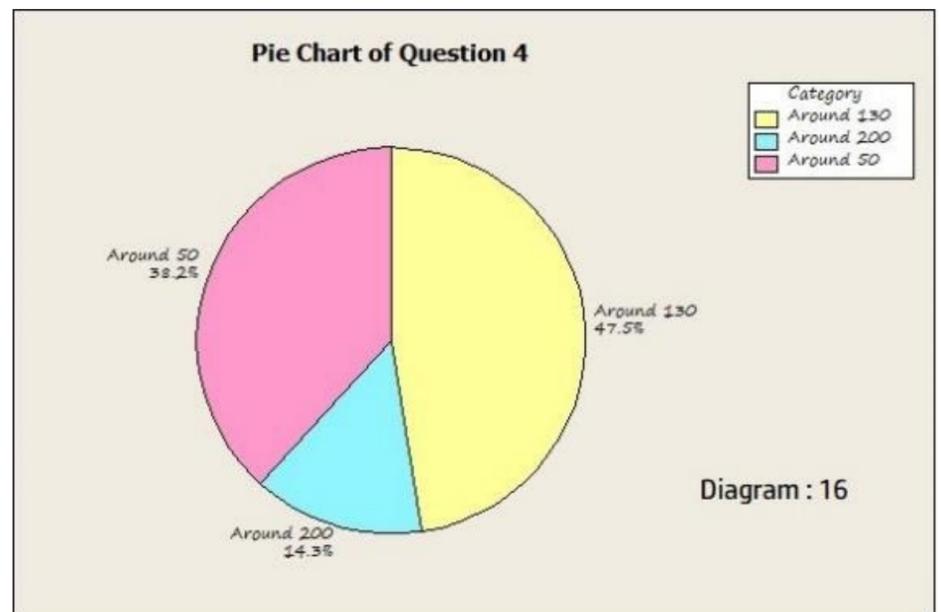
Ethiopia is a country in sub-Saharan Africa. In 1960 the child(0-5years old)mortality(per 1000 birth) was about 276. What do you think about its present child mortality rate? (per 1000 birth)

a. Around 50 **b.** Around 130 **c.** Around 200

Data Visualization: Glimpses of Today's World in The Light of Data

Pie Chart of the Survey : We represent the results of this survey using Pie chart.(Diagram : 16).

The 4th question in our survey is about the present child mortality rate of Sub Saharan African country Ethiopia. The correct answer is- its current child mortality rate is around 50 per



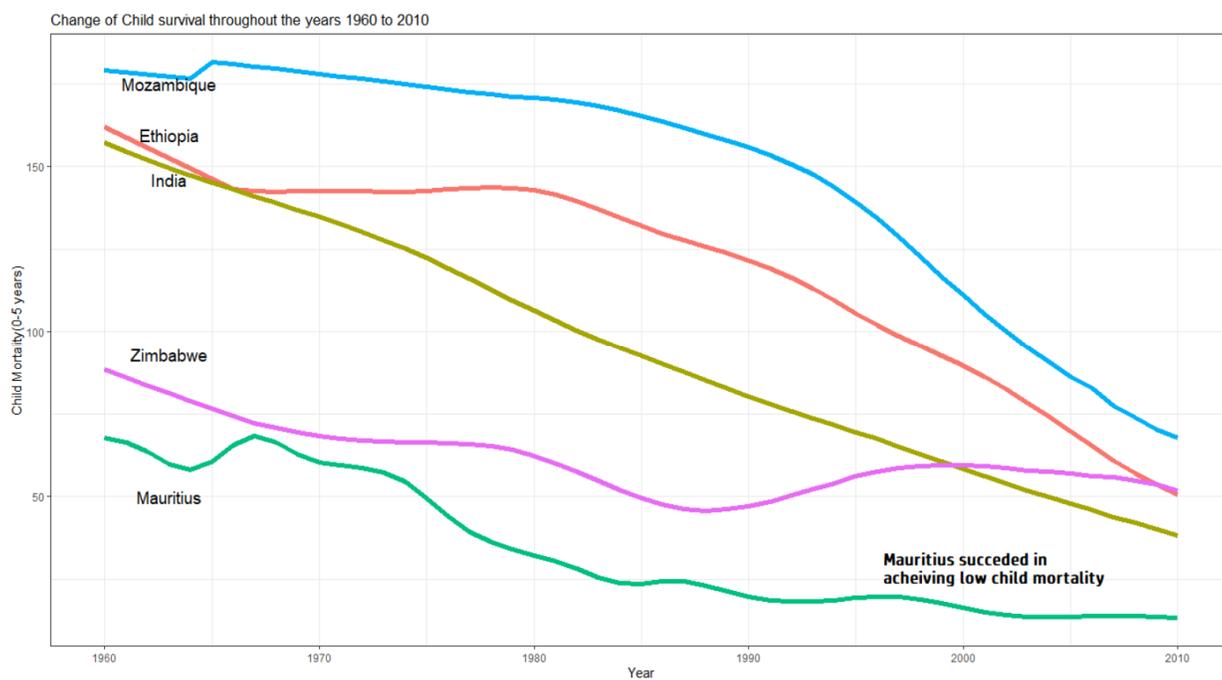
1000 birth. Surprisingly only 38.2% of the respondent have given the correct answer and 47.5% have chosen the option 'Around 130' and even 14.3% have chosen the option 'Around 200'. Is it not reflecting our pre-conceived mindset about a typical Sub Saharan African country and also our ignorance to their power of improving day by day?

In previous questions we have seen that how the old mindset about the world leads us to think all things negatively. 50-60 years ago, most of African countries were Level -1 countries and only a few were in level - 2 (like Mauritius, Algeria). From our previous discussion, we have understood that people living in Level 1 countries had to fight every day for food. So it was not always possible to give their children proper food, care, treatment, nutrition and nourishment to keep them healthy and hearty. These all things are represented by a single number, 'Child Mortality'. (Again, we found "Numbers can speak"!)

But things have changed! Today, we can see that many African countries have been succeeded to fight against extreme poverty and upgrade themselves from Level-1 to Level-2. Ethiopia is also one of them. The child mortality rate of Ethiopia has decreased from 276 to around 50. To witness this change let us draw the time series plot of child mortality rate(0 - 5 years) over the years 1960-2010 for the countries: Ethiopia, Mauritius, Mozambique, Zimbabwe(All these being Sub Saharan African Countries) and India(Asian Country).

From the graph(Diagram : 17) we can see how these Sub Saharan African and Asian Countries have decreased their child mortality rate drastically within last 50-60 years. An important observation is the child mortality rate of Mauritius was always less than India over the years.

Data Visualization: Glimpses of Today's World in The Light of Data



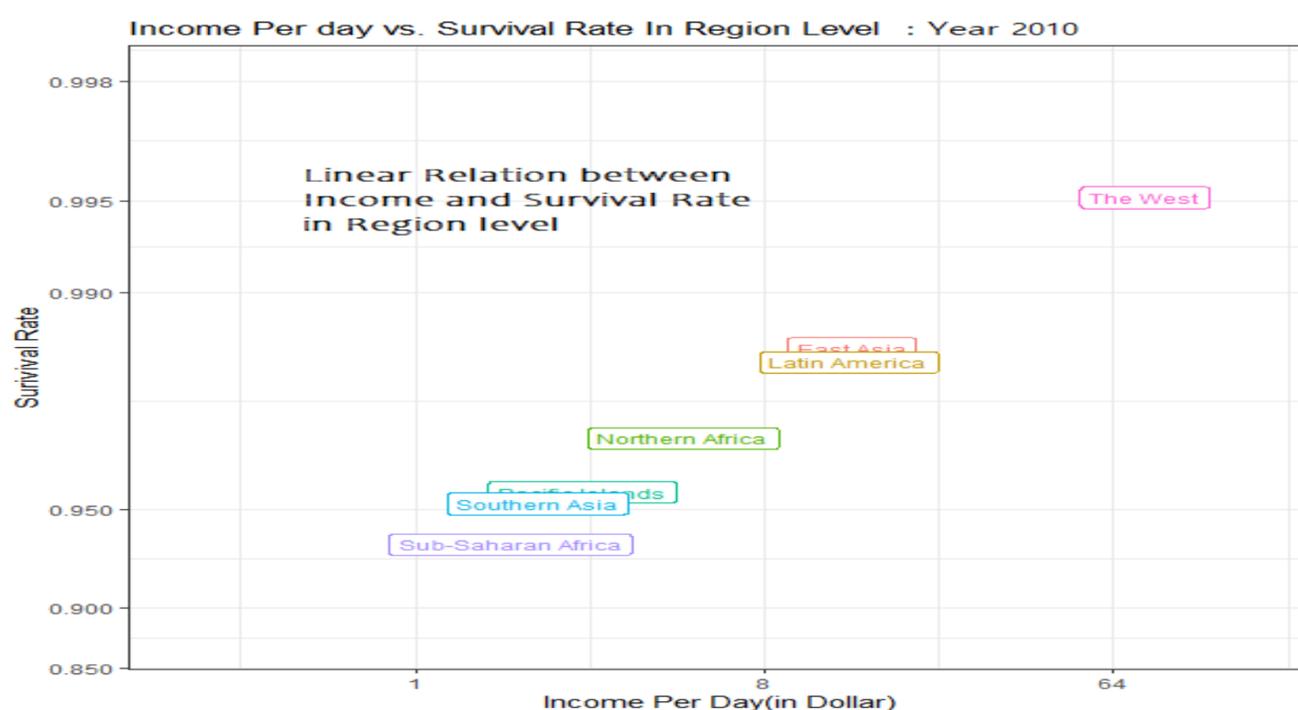
Ethiopia and other countries have also decreased their child mortality rate significantly. So, it is clear that African countries have improved its economic condition in last

Diagram 17

50-60 years, some African countries are even doing well than many Asian countries, the thing is that they need support from the outer world to boost up their economy, health care system, education and consequently they will be able to increase their life expectancies, decrease the number of babies per woman and child mortality. Countries like Mauritius, Ethiopia, Egypt etc are the brightest examples of this.

So where lies the problem within ourselves? Actually, we unknowingly divided the world into some zones like - European, American, Asian, Latin American, Sub Saharan African etc. whenever a question comes, we, in our subconscious mind, use our preconceived notion about that zone and conclude.

Let's draw Income Per day vs. Survival Rate in Region level (Year 2010).(Diagram: 18)



Observe that, there is a linear relation between income and survival rate, i.e. as income per day increases survival rate also increases. This is what we actually think in our minds.

Diagram 18

Data Visualization: Glimpses of Today's World in The Light of Data

As the question comes from Sub Saharan African countries, using this logic, many of us conclude that it's present child mortality rate would be high. Now, let's draw the same graph in country level.(Diagram : 19)

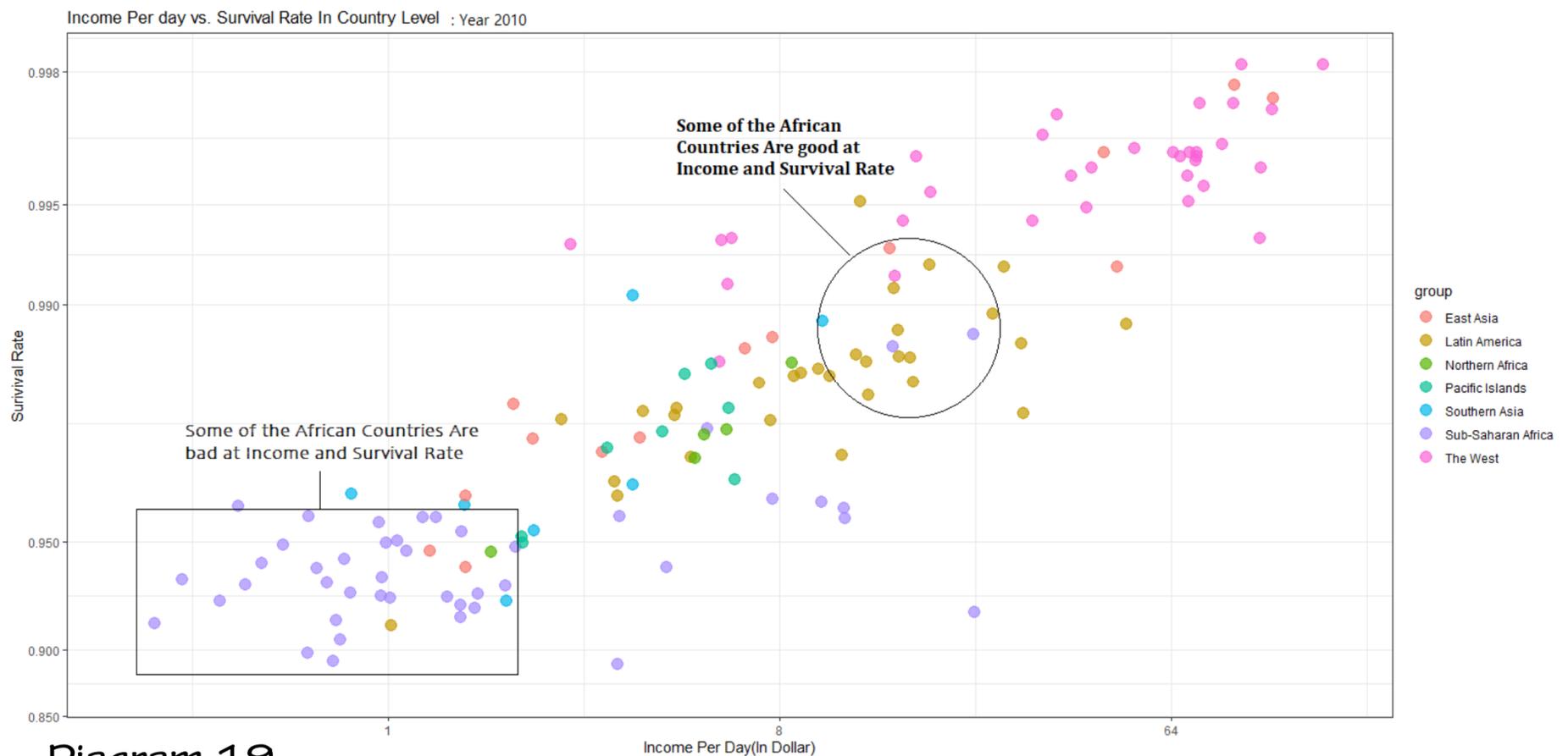


Diagram 19

When the big picture is taken into account, it is really something different. Some of African countries are actually good at both income and survival rate. The variability within a region is large and countries having same income can have different survival rates, which is not captured by the average figures. Jumping to some conclusion only looking at average figures is called 'Ecological Fallacy', probably, which is intending people to choose the wrong answer.(We use logit transformation for survival rate to show the changes significantly)

(Note that, since, we have data till 2010 in "gapminder" dataset. That's why we are using time series plots up to year 2010, though in 2020 also, Child mortality of Ethiopia is about 50 per 1000 birth)

Conclusion :

From the above discussion probably, we have been able to show you the glimpses of the world using data and data visualization. So, in many cases there has been a clear-cut differentiation between our pre-conceived mindset about the world and its reality.

In some cases, it also reflects our ignorance to the happening changes in the world. World is changing very fast ! Scenario of the world is not very eternal, what is true for today may not be true tomorrow.

Data Visualization: Glimpses of Today's World in The Light of Data

So only thing we can do is to walk in a rhythmic manner with them and keep upgrading ourselves with the happening changes.

So, don't lose hope! Mankind is doing well and we can assure that the world is better than many of you think.

Reference :

[1] FACTFULNESS by Hans Rosling, Ola Rosling and Anna Rosling Rönnlund

For Further Readings :

- 1.<https://www.weforum.org/agenda/2019/09/5-reasons-why-ethiopia-could-be-the-next-global-economy-to-watch/>
- 2.<https://ourworldindata.org/>
- 3.<https://en.wikipedia.org/wiki/Logit>
- 4.<https://www.worldbank.org/en/country/vietnam/overview>

Articles from STUDENTS

Determination of Gender in Dinosaurs: Statistics Concludes

Shantanu Nayek (2nd Year)

Nature has its way of distinguishing male species from females through various physical and behavioural patterns. Some of the most apparent instances where male species are physically distinguishable from females are -

- Peacocks have 6 feet long feathers, whereas peahens have significantly shorter feathers.
- Lions bear majestic manes whereas lionesses don't.
- Female hawks are on an average 30% bigger than male ones.

These differences make it significantly easier for biologists and naturalists to identify the gender of a species. However, studying these shreds of evidence through fossils becomes extremely difficult, leading to frequent inaccuracies. This is one of the major problems that palaeontologists face during the determination of the gender of various species of dinosaurs. Not until a few years ago, through various experimental research and analysis, it has been found that statistics and certain statistical experiments provide a considerably efficient solution to the above-mentioned problem.

According to the palaeontologists, the only fool-proof method was to look for fossilized eggs within the dinosaur's skeleton, following which they would label it as a female. Any skeleton lacking fossilized eggs was classified as males. However, the presence of dinosaur skeletons that are available and fit for studying is very limited. Therefore, this process of identification was discarded. Further, there were no conclusive results that ensured that dinosaurs bigger in size could be classified as a certain specific gender. As a result, gender identification became a crucial and difficult hurdle to cross.

In order to study sexual dimorphism, palaeontologists came up with a new method. They started taking the help of a statistical process known as 'significance testing'. Through this method, they collected certain data points on a particular feature and calculated the probability that those characteristics resulted from pure chance rather than an actual cause.

Determination of Gender in Dinosaurs: Statistics Concludes

This technique is analogous to the one where physicians determine whether a new medicine is more efficient than a placebo. According to Evan Saitta [1], this process actually works well for big datasets. Although, since fossil specimens are very limited, this method results in false negatives in several cases.

Today, scientists and palaeontologists have joined together in their quest for a new method to obtain accurate results using smaller datasets. Along with his colleagues, Saitta started experimenting with other Statistical tools which work well for smaller datasets, called 'effect size statistics'. Effect size statistics helps to determine the degree of sexual differences along with calculating the extent of uncertainty in the obtained estimated results. The importance of this method is that it takes natural variations into account excluding various effects of dimorphism. Various codes for this line of research, especially the simulations, were written by Max Stockdale, the University of Bristol, while the measurements of the limited number of fossils, which include the estimates of the body mass dimorphism, were done by Saitta and his colleagues. The error bars obtained in the estimate were not possible to visualise in the significance testing method.

According to Saitta, there is a lot of research yet to be done for this study, but he is quite satisfied with the work that has been done. This is because the statistical simulations led to significantly consistent results, in spite of the limited number of resources in fossils data.

REFERENCES

[1] Evan T Saitta, Maximilian T Stockdale, Nicholas R Longrich, Vincent Bonhomme, Michael J Benton, Innes C Cuthill, Peter J Makovicky. **An effect size statistical framework for investigating sexual dimorphism in non-avian dinosaurs and other extinct taxa**. Biological Journal of the Linnean Society, 2020; DOI: [10.1093/biolinnean/blaa105](https://doi.org/10.1093/biolinnean/blaa105)

Articles from STUDENTS

Statistics: Lies, Damned Lies?

Saptarshi Chowdhury, Utsyo Chakraborty (1st Year)

A very renowned phrase, popularized by Samuel Langhorne Clemens aka Mark Twain, has been in touch for quite a long time to describe the persuasive strength of Statistics to support weak arguments:

"There are three kinds of lies: lies, damned lies, and statistics."

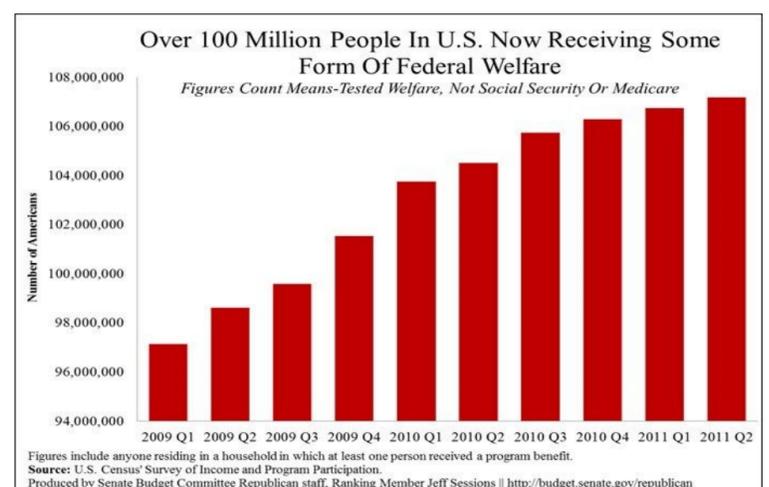
~Benjamin Disraeli, 1st Earl of Beaconsfield

If a person says that he/she has not been influenced by Statistics, he/she would be either lying or living under the rock. Statistics, as a domain, has become a part and parcel of our daily life, and this is the very reason why few entities use it negatively or misleadingly to delude the mass.

Few examples where major organizations have manipulated data/visualization of data to mislead people are:

1. On August 8, 2012, USA Today (by Daniel Harper) published the following graph concluding that "Over 100 Million People in the U.S. now receiving some form of Federal Welfare".

Now, at one glance, it seems that the welfare problem is just spiraling out of control for the Government to handle but if you notice carefully, the graph is faulty (which will be discussed in the next section).



2. A few months back, Argentinean TV channel C5N manipulated the y-axis to hide the terrible number of COVID-19 test as can be observed from the below graph:



(Source: C5N)

At a single glance, the above graph suggests that Argentina tests almost as many people as the USA (EEUU)! However, Argentina tests 330 people per million, while the USA (EEUU) tests 7000. That is almost 20 times more, but the USA's bar is only 1.2 times higher, suggesting a clear manipulation.

Statistics: Lies, Damned Lies?

of data visualization.

Some of the major sources of Misleading Statistics are:

- 1) **Biased sampling:** Given the limited amount of resources in the form of time, cost, and manpower constrained upon statisticians who are interested in carrying out a survey or an experiment, selecting a sample from the population is always the more practical way out. However, collecting data from a biased sample that does not truly represent the population can skew the results obtained from the collected data, for better or for worse.
- 2) **Heavily worded and opinionated questionnaires:** A study conducted by Georgina Gous and Jacqueline Wheatcroft has been comprehensively able to prove that phrasing a question will always influence how responses will be generated. Questions may be constructed in such a way as to reinforce pre-existing opinions. This leads to biased and thus faulty data.
- 3) **“Correlation does not imply causation.”** The reason for the almost chant-like repetition of this phrase is very justified, as several organizations falsely claim causality based on a strong correlation obtained between two seemingly independent events. There is always some hidden "lurking" factor causing that correlation.
- 4) **Manipulated Graphs and Charts:** Instead of being beguiled by hardcore numerical facts and figures, looking at data through graphical means can help us conclude in a much more relaxed manner. However, this has turned out to be the most misleading source of statistics in recent times. Sampling and confirmation bias can easily be manipulated graphically and thus can be used to deceive audiences.

To perform a live study of how people could be misled by data and their faulty visualizations, we decided to circulate a questionnaire through a Google Form to our mutual contacts. The questionnaire had 4 questions with attached graphical representations of data. We received 43 responses from students of various institutions, namely St. Xavier's College Kolkata, ISI Kolkata, IIT Bombay, Jadavpur University, IISER Kolkata, RKM Narendrapur, Delhi University, Calcutta University, Christ University,

Statistics: Lies, Damned Lies?

Bethune College, and Jorhat Engineering College.

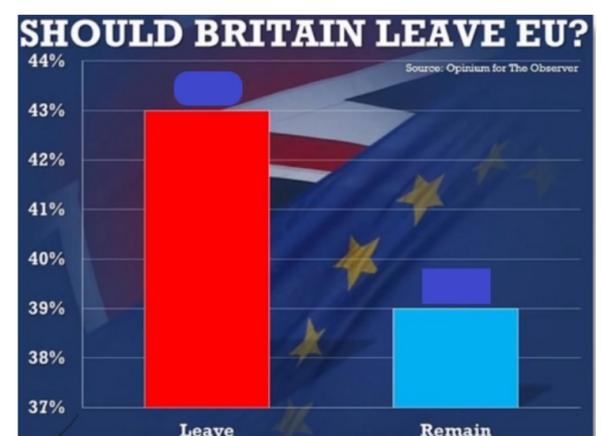
The questions asked were as follows:

Q1) In 2016, a poll was conducted by "The Observer" which is a renowned British Newspaper under the "Guardian Media Group". The poll posed a question to the readers: "Should Britain Leave EU?" and asked them to vote within 24 hours. After the voting was completed, a bar chart was plotted to visualize the response. Answer the following question related to the image below:

If Red: Leave and Light Blue: Remain, then the proportion of people wanting Britain to leave the EU is actually much larger than that of people wanting Britain to remain.

A: TRUE

B: FALSE



(Source: Opinion for The Observer)

RESULTS:

OPTION	NUMBER OF RESPONSES	PERCENTAGE OF RESPONSES
A	25	58.1%
B	18	41.9%

From the above graph, we can notice that the y axis starts from 37% and not from 0%, which should have been the ideal starting point. Hence, the data which concludes that 43% of people who want Britain to leave the EU as compared to 39% of people who want Britain to remain in the EU is **not significantly large** ($\approx 4\%$). Due to faulty indexing, the first glance of this graph tells us a completely different story, **as our study testifies**.

Q2) Consider that a data collection is done where a certain University wants to know the number of students getting above 80% over the years in the Final Semester Examination in Statistics. Based on the line graph below, visualized by the HOD:

What can you interpret about the graph from the image below?

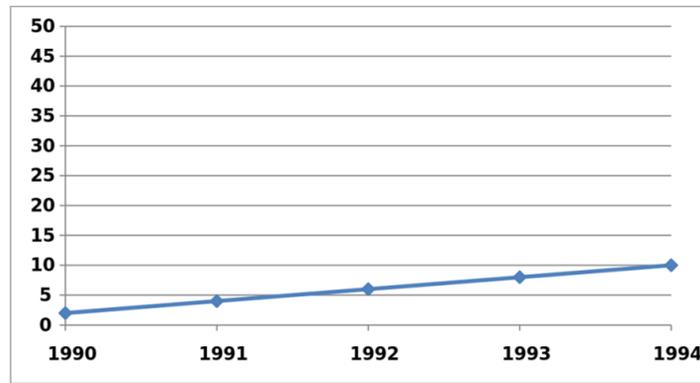
A: Less volatility

B: More volatility

C: Less Growth

D: Both A and C

Statistics: Lies, Damned Lies?

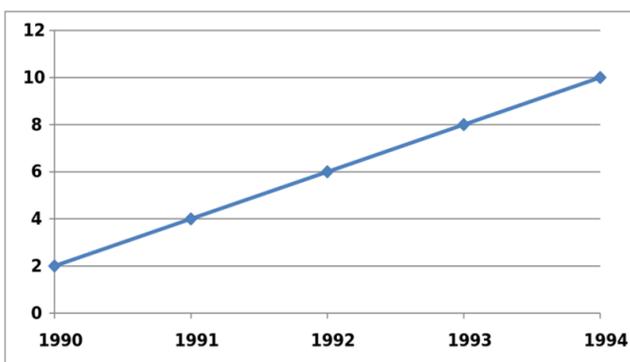


(Source: Wikipedia)

Results:

OPTION	NUMBER OF RESPONSES	PERCENTAGE OF RESPONSES
A	13	30.2%
B	7	16.3%
C	3	7%
D	20	46.5%

The line diagram provided by us was in actuality a manipulation of the line diagram shown below:



(Source: Wikipedia)

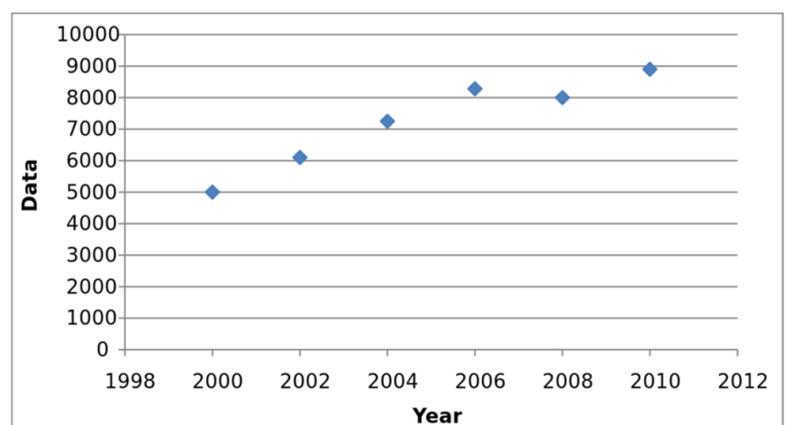
Looking at this line diagram, there seems to be no volatility whatsoever and a linear growth in the number of students who have received above 80%. Since the **y-axis maximum has been inflated** in the diagram provided by us, there appears to be a small amount of growth, thus the slope of the line is less steep as is testified by the responses obtained.

Q3) A study was conducted by the "Bombay Stock Exchange" to learn whether they were actually making any profit in the Economic Crisis (between 1998 and 2010) that the world was going through. On the basis of it, a scatter plot was drawn by their Chief Analyst. Answer the question based on the image below:

If Y-Axis: Profit and X-Axis: Year, can you conclude that the growth appears to be more or less linear with less variation?

A: Yes B: No

C: Insufficient Information to conclude anything



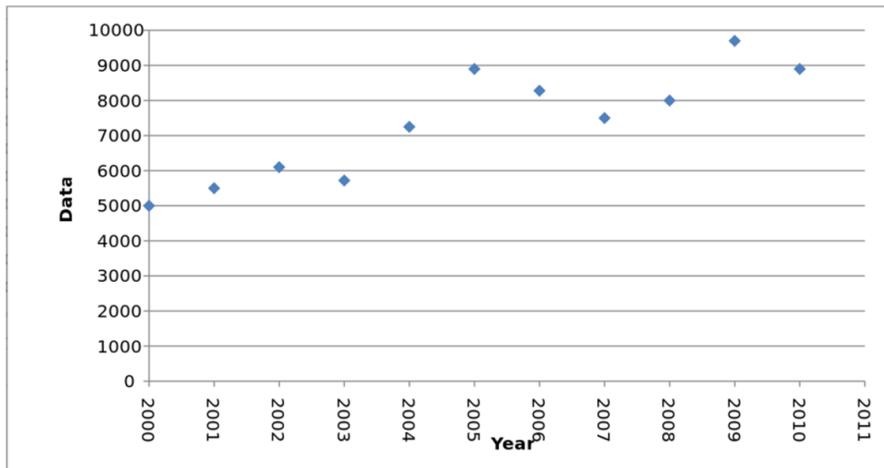
(Source: Wikipedia)

Statistics: Lies, Damned Lies?

Results:

OPTION	NUMBER OF RESPONSES	PERCENTAGE OF RESPONSES
A	32	74.4%
B	5	11.6%
C	6	14.0%

The scatter plot provided by us was in actuality a manipulation of the scatter plot shown below:



(Source: Wikipedia)

The scatter plot provided by us had conveniently **omitted observations** from the odd years between 1998 and 2010. This resulted in a linear-looking scatter plot whereas, in actuality, it **isn't linear at all**.

Hence, the respondents were misled into believing that profit was linearly increasing over the years.

Q4) A certain Russian news channel conducted a study and collected data on the number of COVID-19 cases from March 5, 2020, to March 31, 2020. After visualizing the data, a bar chart was plotted by them. Answer the question below the image. Here, Y-Axis: Number of cases per day; X-Axis: Day of the month.

After March 26th, the growth seems to slow down, and thus the Russian Government has done a good job to flatten the curve.

- A: True B: False
- C: Insufficient Information to conclude anything



(Source: Russia Today)

Results:

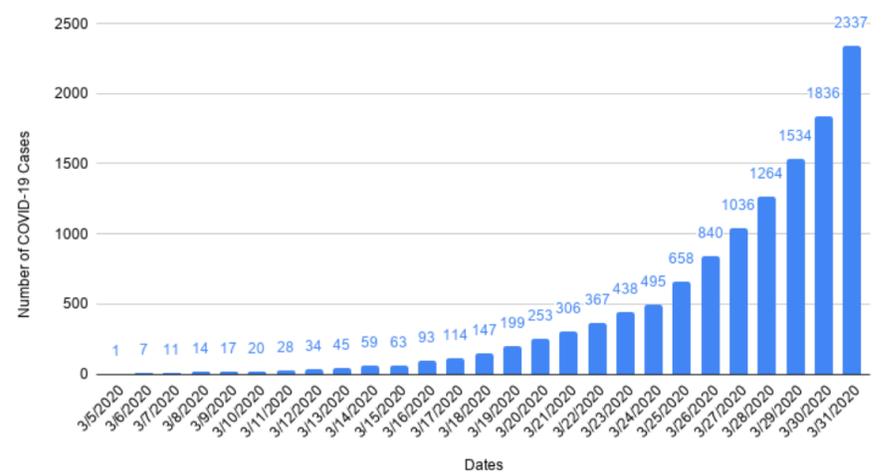
OPTION	NUMBER OF RESPONSES	PERCENTAGE OF RESPONSES
A	11	25.6%
B	26	60.4%
C	6	14.0%

In our given diagram, the bar heights correspond to the number of COVID-19 cases quite precisely till the 26th of March. However, after this date, the bar heights seem to be arbitrarily decided. An increase by 500 cases on March 31st does not correspond to the visual increase, which shows something like a 50-case increment. **In actuality, the graph should have been:**

Statistics: Lies, Damned Lies?

Even though the graph provided by us is misleading and it can be interpreted from it that the Russian Government has done a good job in flattening the curve, a majority of our respondents were wise enough to figure out the anomaly and responded False.

Number of COVID-19 Cases in Russia from March 5 to March 31



(Source: Nikita Kotsehub, Towards Data Science)

CONCLUSION:

Many individuals around the globe have been continuously misleading people using faulty techniques of collection and visualization of data. To cross-check this statement, we conducted a study (a very small scale one!) involving people having a background in statistics. Their responses show us that despite having proper and educated knowledge on how to interpret visualizations of data accurately, they have fallen prey to the dangers of data corruption. So how do we escape from this rather inevitable maze? The answer is pretty simple and does not require a Ph.D. to appreciate: "keep your eyes open!"

References:

- 1) Kotsehub, N. (2020, June 24). Stopping COVID-19 with misleading graphs. Retrieved February 14, 2021, from <https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9>
- 2) Misleading graphs: Real life examples. (2021, January 11). Retrieved February 14, 2021, from: <https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/>
- 3) Hessney, S. (2020, June 08). 'Air travel Surges 123%': A lesson about misleading statistics. Retrieved February 14, 2021, from <https://www.nytimes.com/2020/06/08/learning/air-travel-surges-123-a-lesson-about-misleading-statistics.html>
- 4) How accurate data was turned into misleading articles by the government ...and the press. (n.d.). Retrieved February 14, 2021, from <https://www.baekdal.com/trends/how-accurate-data-was-turned-into-misleading-articles-by-the-government-and-the-press/>
- 5) Huff, D. (1993). How to lie with statistics. WW Norton.

Articles from STUDENTS

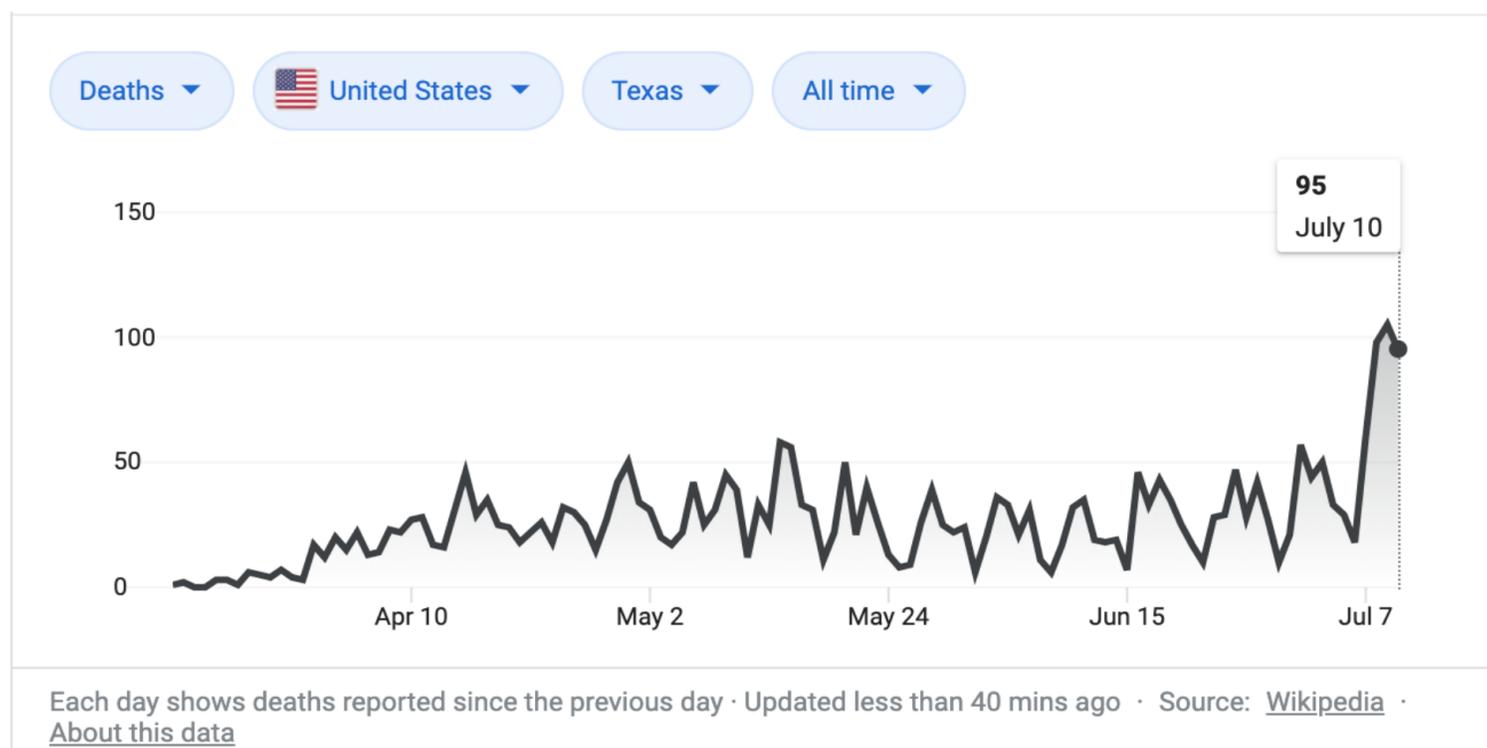
Misleading Statistics of COVID-19

Soham Chatterjee, Atreyee Roy (1st Year)

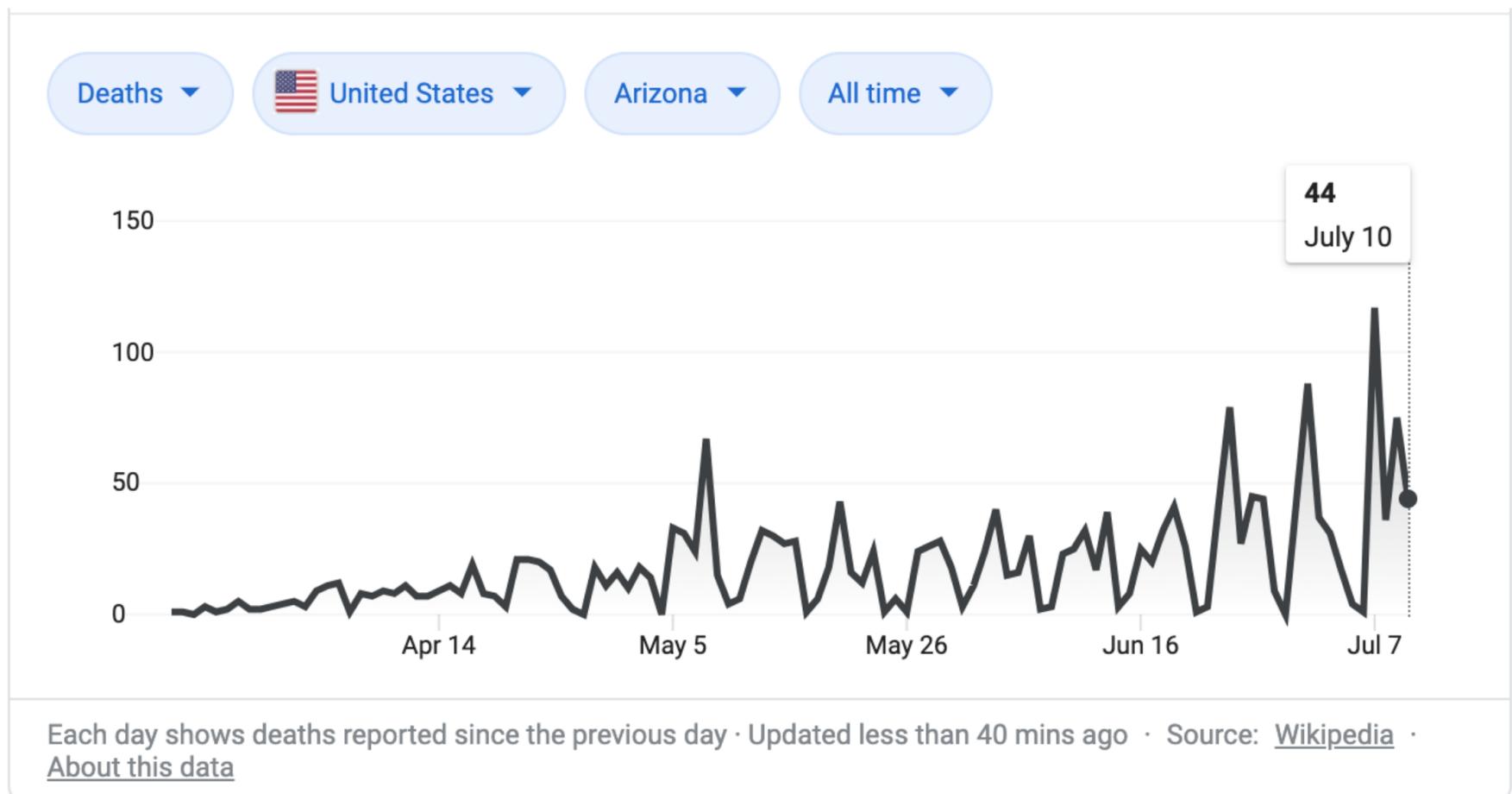
The 2019-20 coronavirus pandemic has affected millions of people, many of whom have succumbed to this disease [3]. Fortunately, these COVID-19 cases are decreasing in number. If we look at the data of the United States, the downward trend in deaths is clearly visible.



Before being optimistic and starting to celebrate our victory against this deadly virus, let us peek into the data of Arizona and Texas (taken from a trusted website, GitHub).



Misleading Statistics OF COVID-19

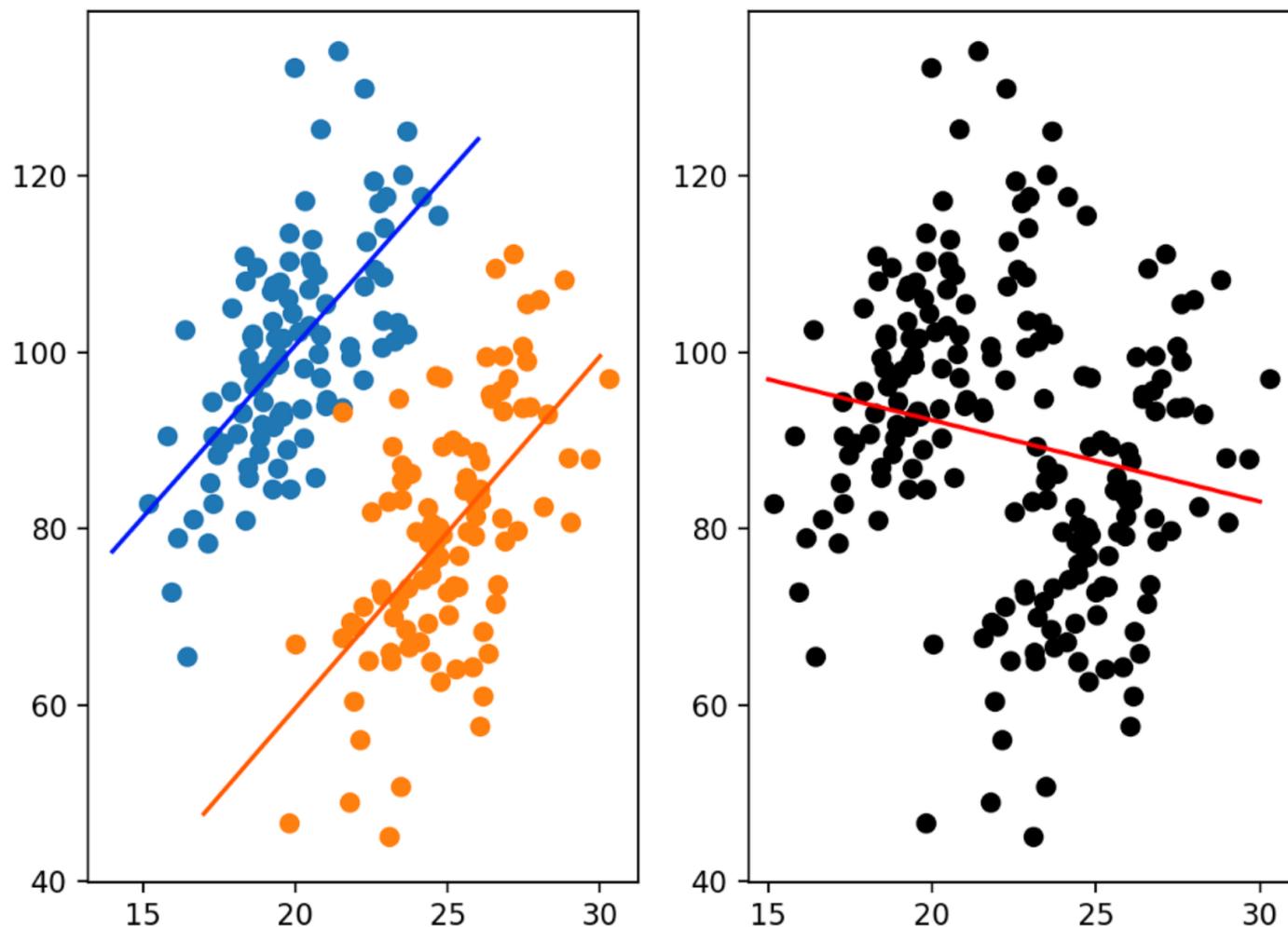


Both the graphs of Arizona and Texas show an increase in deaths in late-June and July which is completely opposite as shown by the United States Graph. What is going on? If the multiple states show an upward trend in the deaths, why does the number of deaths decrease nationally? Has GitHub got the calculations wrong?

No, it is because we have unknowingly entered the world of Simpson's Paradox.

Simpson's Paradox occurs when two or more groups of data individually show a specific correlation, however, this trend reverses or nullifies when these groups of data are aggregated.

Misleading Statistics OF COVID-19



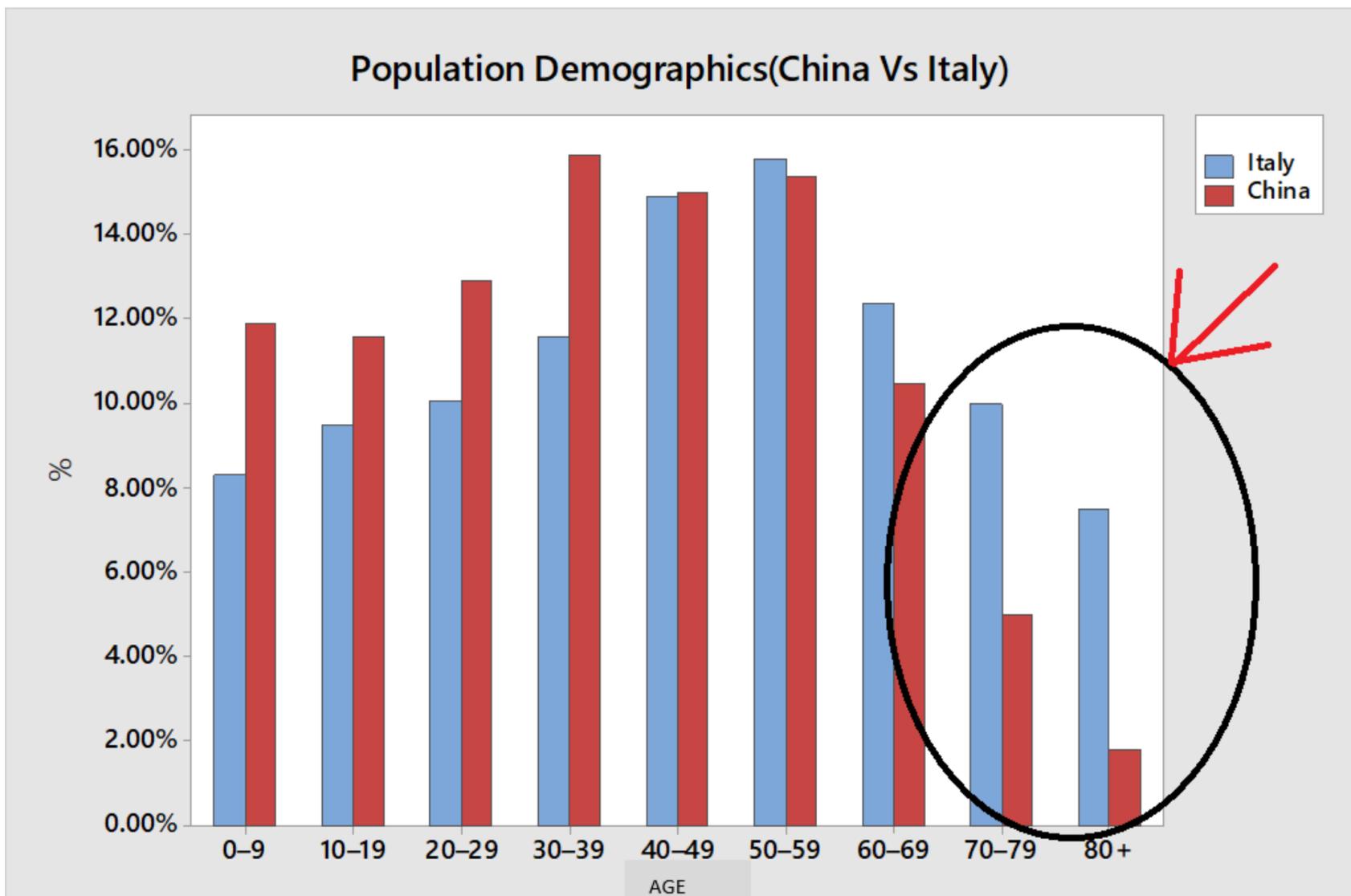
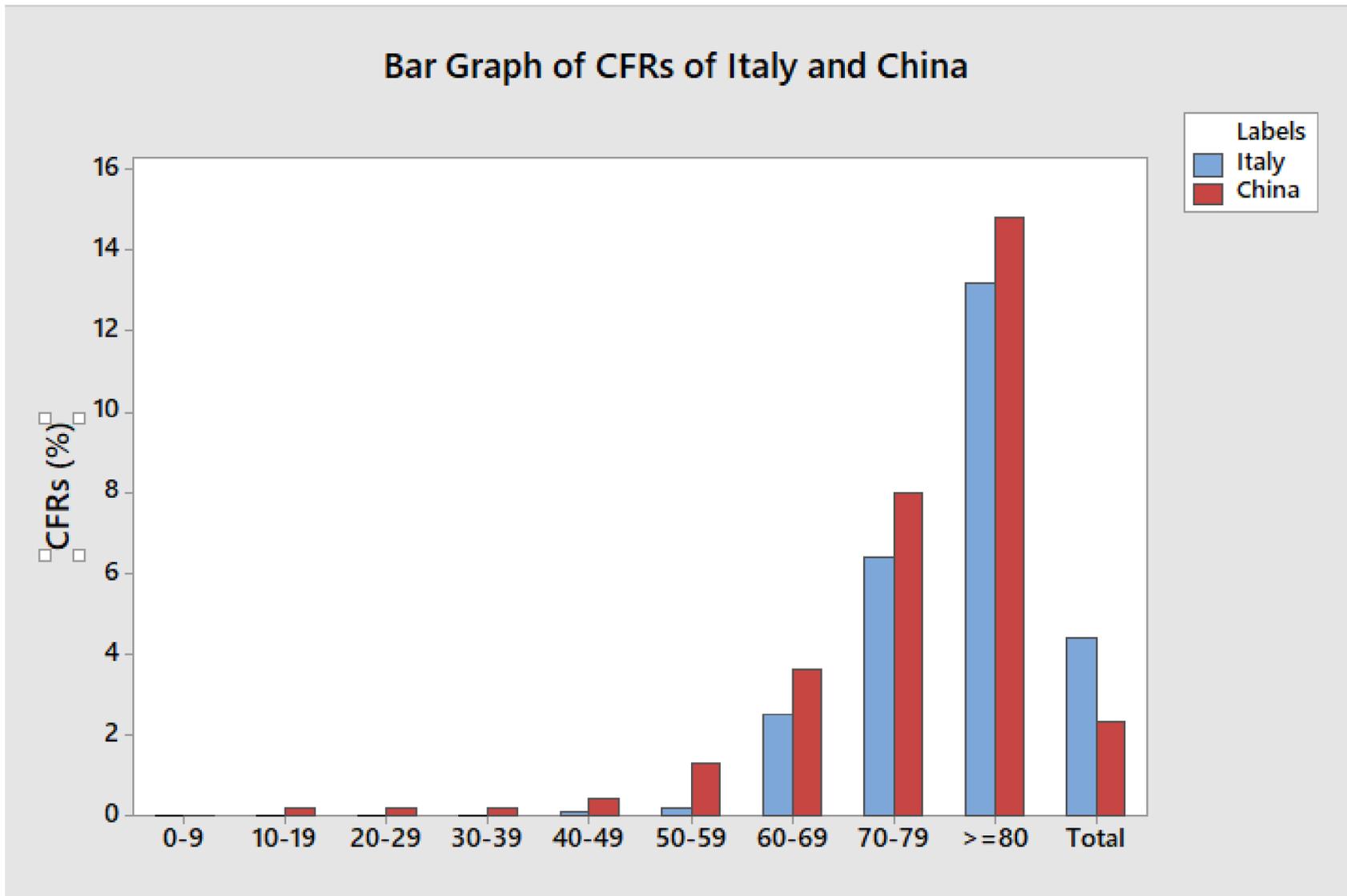
As shown in the figure, Data group 1 (marked by blue) and Data group 2 (marked by yellow) both show an increasing trend or a positive linear relationship. However, when we combine these two data sets, we get a reverse trend or negative linear relationship.

Simpson Paradox has found its application in many fields. As the topic suggests, we shall discuss how the Paradox explains some misconclusions of Covid-19 statistics.

Simpson's paradox in Covid-19 case fatality rates:

On comparing data of 44,672 cases from China with early reports from Italy, we see an exemplification of Simpson's paradox in Covid-19 Case Fatality Rates (CFRs) [2]. Before starting, let us make it clear what CFRs are. CFR is the proportion of people who succumb to a specific disease among all individuals affected with the disease over a stipulated period of time. Hence, we can say, one of the most important indicators of COVID-19 is CFR.

Misleading Statistics OF COVID-19



Misleading Statistics OF COVID-19

As shown in the figure, Data group 1 (marked by blue) and Data group 2 (marked by yellow) both show an increasing trend or a positive linear relationship. However, when we combine these two data sets, we get a reverse trend or negative linear relationship.

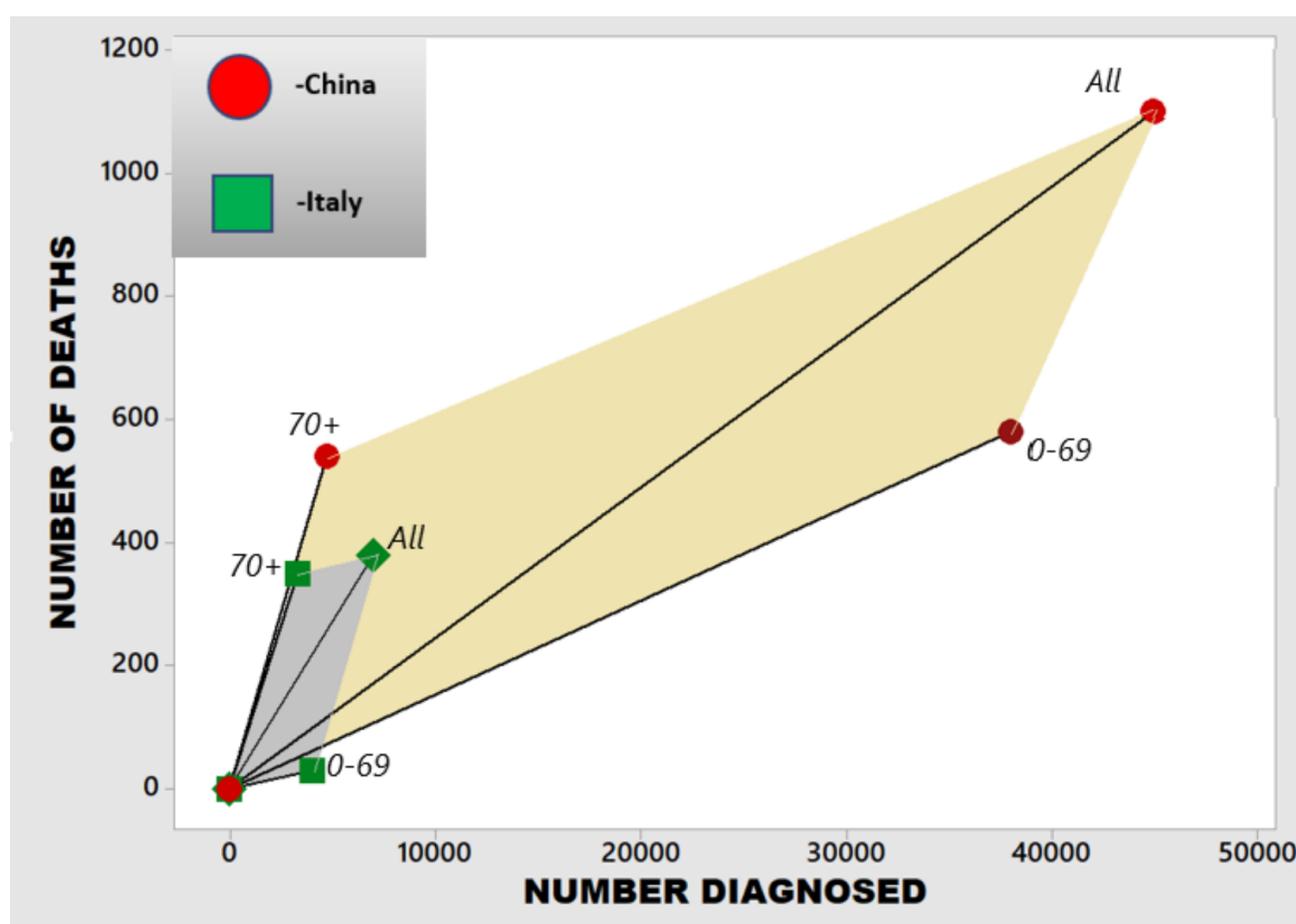
Simpson Paradox has found its application in many fields. As the topic suggests, we shall discuss how the Paradox explains some misconclusions of Covid-19 statistics.

Simpson's paradox in Covid-19 case fatality rates:

On comparing data of 44,672 cases from China with early reports from Italy, we see an exemplification of Simpson's paradox in Covid-19 Case Fatality Rates (CFRs) [2]. Before starting, let us make it clear what CFRs are. CFR is the proportion of people who succumb to a specific disease among all individuals affected with the disease over a stipulated period of time. Hence, we can say, one of the most important indicators of COVID-19 is CFR.

Vectorial interpretation:

This phenomenon can be further illustrated via a two-dimensional vector space where the vertical axis represents the number of people who have succumbed to COVID-19 and the horizontal axis represents the number of people diagnosed with COVID-19.



Misleading Statistics OF COVID-19

Consider the 0-69 group Italy vector. Its slope determines the group's CFR; its magnitude being proportional to the group size. In the same way for the 70+ group and combined group ("All" i.e., the sum of the previous mentioned vectors whose slope represents the total CFR in Italy) we get the vectors as shown.

Now when we compare the vectors for Italy with that of China, we notice how for both the subgroups Below 70 and Above 70, China's vector has a seemingly greater slope. However, for the 'All' group Italy's vector has a greater slope. Why does this happen? This is simply because the component vectors for each subgroup have different magnitudes, also implying that the larger vectors dominate the whole point of comparison, finally implying how age demographics play a pivotal role.

Causal model: Age as a mediator

Consider:

C: the country where a positive case is reported.

A: Age group

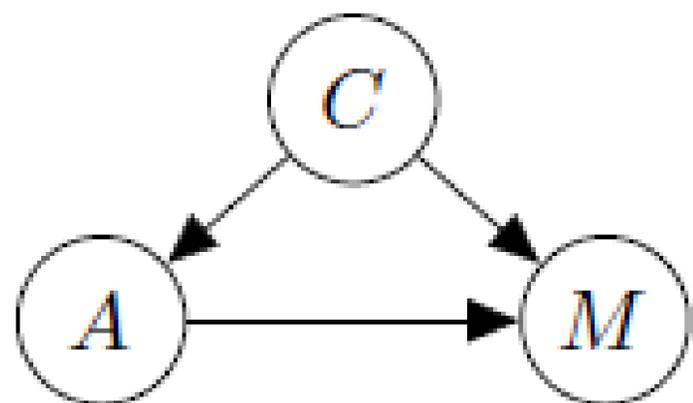
M: Mortality (M=0, 1; 0: patient has died; 1: Patient has not died).

Now,

(C → A): reflects that age demographics depends on the country.

(A → M): reflects that disease is more fatal for the old.

(C → M): reflects factors other than age like lockdown rules, testing, health infrastructure, etc.



One may ask "If you are, say around 62 years old, is it safer to be in China during the pandemic?". As the causal map suggests age being a mediator, one needs to carefully analyse mediation without jumping to conclusions.

Which one is the reality?

In statistics, Simpson's paradox poses a spirit of inquiry, but the COVID-19 statistics shows that it delivers a rudimentary question [4]. Which depicts the reality correctly?

Misleading Statistics OF COVID-19

What level of aggregation do we want to focus on? If we put everything into one parcel, we may miss out information on what the reality is, in our example the case of China vs Italy. In this case, we have to consider the data generation process, the causal model that is responsible for the data. Only if we fathom the technique that is generating the data, we can prevent ourselves from hypothesising flawed inferences from numbers. That is why, it is rightly said,

“Simpson’s Paradox is an interesting statistical phenomenon but it also demonstrates the best shield against manipulation is the ability to think rationally and ask why.”

Now answer,

We know, India has come up with various working COVID-19 vaccines. But now one may ask which vaccine is suitable.

Suppose you are given this data.

Overall:

	Proposed COVID-19 Vaccine 1	Proposed COVID-19 Vaccine 2
Not Immune	29,750	45,025
Immune	25,250	5,475

For Mild Cases:

NORTH AMERICA		
	Proposed COVID-19 Vaccine 1	Proposed COVID-19 Vaccine 2
Not Immune	4,750	45,000
Immune	250	5,000

For Severe Cases:

EUROPE		
	Proposed COVID-19 Vaccine 1	Proposed COVID-19 Vaccine 2
Not Immune	25,000	25
Immune	25,000	475

Misleading Statistics OF COVID-19

The data shows that the Vaccine 2 provides better immunity in both mild and severe cases but however overall Vaccine 1 is better (45.9% efficient) [1]. Given a bet of Rs.100 that a randomly selected COVID patient will be immune to the virus by taking Vaccine 2, will you take the bet?

References:

[1]<https://www.capgemini.com/gb-en/2020/11/what-is-simpsons-paradox-why-correlation-is-not-always-straightforward/>

[2] Wang, Z., Rousseau, R. COVID-19, the Yule-Simpson paradox and research evaluation. *Scientometrics* (2021).

<https://doi.org/10.1007/s11192-020-03830-w>

[3] <https://towardsdatascience.com/what-is-simpsons-paradox-4a53cd4e9ee2>

[4]<https://towardsdatascience.com/simpsons-paradox-how-to-prove-two-opposite-arguments-using-one-dataset-1c9c917f5ff9>

Articles from STUDENTS

Interpretation of Statistics in Different Fields of Science

Samapan Kar (1st Year)

“If your experiment needs a statistician, you need a better experiment” –

Ernest Rutherford

Although every time we can't change the whole experiment, we can modify that with the help of statisticians. Statistics is the subject that attempts to clarify those fundamental queries about experimental design and inference. Statistics became an inherent part of most of the scientific experiments to get more accurate results and conclusions using its very own sophisticated approach. Let's take a glance at those two important uses.

In Quantum Mechanics:

The statistical interpretation of Quantum mechanics was established to interpret and modify the theories with minimum assumptions. In the early 20th century when Quantum mechanics was introduced, several arguments were put in place to consider the quantum state description, to apply only to an entity of similarly prepared systems and not to any individual physical system. Most of the problems are related to these arguments. So, the introduction of the hidden variables to see the end result of individual events is totally compatible with the statistical predictions of quantum theory.

At the very beginning of 20th century when people were trying to get a good answer about “What is Electron” and “What is an atom made of”, de Broglie made his groundbreaking hypothesis saying that electron also possess some wave nature which was verified by Davison-Germer and G. P. Thompson individually and that further lead to a question...

How can we conceptualize an electron which shows both particle and wave nature? We know that waves do not stay at a point like particles. Then how can we find an electron in a particular position of 3-dimensional space at a particular time, if we consider it as a wave?

Interpretation of Statistics in Different Fields of Science

At this point W. Heisenberg was prominent in the world of Quantum mechanics. After a lot of experiments, he said that we cannot find the position of an electron exactly at any space time. What we can do is, to calculate the PROBABILITY of finding the electron at that point and based on this he formulated his paper known as “Matrix Mechanics”.

A Mathematical form of Heisenberg’s Uncertainty Principle can be written as:

$$\sigma_x \sigma_p > \frac{\hbar}{2}$$

Here x is the position of electron and p is the momentum of electron, σ_x is the standard deviation in the measurement of x , σ_p is the standard deviation in measurement of p and \hbar is the reduced Planck’s constant with value 1.054×10^{-34} Js. So standard deviation plays an important role to measure the uncertainties in calculation of observable quantum mechanical variables.

After Heisenberg’s formulation Erwin Schrodinger in his paper on Quantum Theory (known as “Wave Mechanics”) said that we can express any quantum mechanical system (such as electron) with a Wave Function $|\psi\rangle$. But this wave function didn’t have any physical interpretation.

Here comes the Genius Max Born who interpreted wave function in a different manner. He said that if we multiply $|\psi\rangle$ with its complex conjugate $|\psi^*\rangle$ then we will get the probability density function for finding a quantum mechanical particle (such as an electron) in space. On the other hand, Schrodinger considered this quantity multiplied with charge as the charge density of an electron or any other charged QM particle in space. But Born’s interpretation was successfully verified and also called the “Copenhagen Interpretation of Wave Function”.

Suppose a QM particle is constrained to move on X axis then the probability of finding the particle between 2 points x_1 and x_2 is,

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \psi(x) \psi^*(x) dx$$

Now if the particle is restricted to move between 2 points on X axis (say 0 to a) within an infinite potential well, then we see that the total probability of finding the electron between these 2 points will be 1 (considering wave function $|\psi\rangle$ to be normalized)

Interpretation of Statistics in Different Fields of Science

$$\int_0^a \psi^*(x) \psi(x) dx = 1$$

For 3D space we can say that probability of finding the QM particle throughout the whole space is 1 (considering $|\psi\rangle$ to be normalized)

$$\iiint dx dy dz \psi^*(r,t) \psi(r,t) = 1$$

It is convenient to abbreviate this equation by letting $dx \cdot dy \cdot dz = d$ and write,

$$\int_{-\infty}^{\infty} d\tau \psi^*(r,t) \psi(r,t) = 1$$

(Where d is the infinitesimal volume element)

The above equations were referred to from Quantum Chemistry by Donald A. McQuarrie [2].

There were lots of arguments to disprove the Copenhagen Interpretation. The E-P-R paradox is one of the strongest arguments, although it was resolved later.

After this interpretation of wave function and due to Heisenberg's uncertainty in measurement criteria, probability has become an integral part of Quantum Mechanics, where everything is based on chance, where we don't have much experimental evidence but have the strong power of Probability to predict, depending on the limited observations.

In Phylogenetic studies and Evolution theory:

The scientific study of phylogeny is known as Phylogenetics. Phylogeny is related to the history of evolution of a taxonomic group of organisms. It is basically involved with the connection of an organism to different organisms as indicated by evolutionary similarities and differences [1]. In this study, if we compare a chicken with a crocodile, it would be difficult to imagine that they are closely related. We will also be surprised by the fact that birds are even more closely related to dinosaurs.

But how are the biologists able to state such astounding facts? How can we confirm such outrageous claims? The answer is nothing but evolutionary regression studies or more specifically the regression of biological variables across the species.

Interpretation of Statistics in Different Fields of Science

Nowadays, Evolutionary regression is the most used statistical procedure to contemplate the relations of evolution between variables that represent biological attributes and to examine the hypothetical relations about adaptation of variables related to the environment. Regression models are more advantageous for the connections related to phylogenetic attributes between the connected species. Also, sometimes the statistical presumptions are opposed with the biological perceptions which later alter the examination.

For instance, one of the most common universal assumptions is, the residuals whipped up due to regression of phylogenetic attributes advance as Brownian motion is brought up for scientific benefit. It is conflicting with the evolution related to adaptation. Numerous ways to manage the error, that arises due to analytical bias in regression of evolution, depend on statistical procedures that viably accept the fact that there is no biological mistake in the fitted model.

However, estimation errors are not genuine concerns for regression analysis. In comparative analysis, standard estimation error of the means of variables of a small sample of individuals is utilized as estimation error, as opposed to the error we get from non-random sampling of individual species. These errors ought to be absorbed for better and precise results and to lessen the bias in regression slope. However, for simplicity, we should be cautious about the observation error and the true biological mistakes.

- Regression model:-

In this model different biological traits are utilized as variables of the model. Furthermore, those biological traits vary from species to species. There are various reasons for such distinctions. Now we think about a biological trait, Y . Certain natural traits will affect Y . But the most appropriate value will rely upon different natural components which shift between various species in various examples. Although, determination of Y seems to be impacted by further practically related characters because of associated preferences.

Interpretation of Statistics in Different Fields of Science

Numerous sources of indirect selection following up on Y will be there because of connections with attributes. The distinctions in traits Y could likewise appear because of genetic drift.

Let, Y denote $f(X_1, X_2, \dots, X_m)$, for some function f , where X_i 's are the accurate conditions of applicable factors in these species. Now this is not an exact prediction because we can't assume the impacts of these conditions. So, we can use the test of influence for evolutionary regression on the variables, whose conditions are known. Now if we take X_1 as that type of variable, the regression model will be like

$$Y = \beta_0 + \beta_1 X_1 + r(X_1, X_2, \dots, X_m)$$

here $r(X_1, X_2, \dots, X_m)$ denotes the residuals related to biological attributes which is also equal to $f(X_1, X_2, \dots, X_m) - (\beta_0 + \beta_1 X_1)$ [3]. Due to different conditions of x variables the residuals are distinctive. It seems appropriate to assume that there will be countless such variables with mostly little differences between species.

Also, the central limit theorem then recommends that those residuals might follow a Normal distribution which is also proposed by the standard linear regression model.

REFERENCES :

[1] Evolution - Douglas J. Futuyama, Mark Kirkpatrick

[2] Quantum Chemistry - Donald A. McQuarrie

[3] Thomas F. Hansen, Krzysztof Bartoszek, Interpreting the Evolutionary Regression: The Interplay Between Observational and Biological Errors in Phylogenetic Comparative Studies, *Systematic Biology*, Volume 61, Issue 3, May 2012,

Pages 413-425, <https://doi.org/10.1093/sysbio/syr122>

Articles from STUDENTS

Lady With the Data

Abhay Ashok Kansal (1st Year)

When people in the world talk about war and how the medical team helped the injured soldiers and took equal steps with the soldiers, one always remembers Florence Nightingale. The Crimean War introduced the world with, whom we today appreciate as, The Lady with the Lamp. But it is known to very few that this lady was a Statistician.

Florence Nightingale not only introduced the world with the softness of heart but what she also brought forward was coxcomb (Figure 1). Hundreds of years before different software hit the market, and graphs and charts were something we all needed, Nightingale made data beautiful and visually appealing. When she knocked at the main British base, she found that the hospital sat on top of a large pool, contaminating the water and the hospital building. Patients lay on in their excrement on stretchers covering throughout the hallways. The visitors to the hospital included rodents and bugs. With an increase in the number of ill and injured, the scarcity of soaps and other basic ailments for cure and help increased. It was growing difficult to manage so much with no basic help. Even the water available was not clean and that too had to be rationed [1]. The state had never been so helpless before. More soldiers were dying from infectious diseases caused due to unhygienic and unsterile conditions than from injuries incurred in battle. It was during this time that the government had to bring forward data on what was to be looked into, "Lady with the Lamp" used her wit and brought forward the Rose diagram. Today these diagrams may be of less significance to you and me but during those days it was one of the most significant and eye-catching diagrams used to present data to the world.

Nightingale plays an inspiring role and is a source of inspiration for women in the field of statistics. She paved the road in so many ways. Today we know career options in statistics are growing in nearly every type of industry, but Florence Nightingale probably didn't have many female statisticians to serve as role models in the 1850s.

Lady with The Data

This chart has been referenced by several different names: Polar Area, Wind Rose, Rose, Coxcomb, or even 'Consultants' chart [2]. A polar chart shows values for poly quantitative measures in the same display. It uses a circular layout comprising several equally angled sectors as if representing the slices of a pizza, one for each quantity under the measure. In contrast to the radar chart (which uses position along a scale), the polar chart is seen to have variation in the size of the sector areas which represent the quantitative values. So, the chart can be seen as a mixture of a bar chart (using length as a pre-attentive feature) and a pie chart (radial in nature, segments that are wider at the top, thinner in the centre, i.e., polar coordinates). This helps us understand the form of the chart as a means for representing data.

Florence Nightingale kept all the records of the death toll at the hospital. In the months before she took over from the military, the death rate at the hospital was high enough and usually varied between 32 and 42 percent annually. From her coxcomb analysis, Nightingale was able to prove to the world that the maximum deaths were caused due to "preventable diseases" and that more people died because of low-quality medical treatment than compared to those who were martyred in the war.

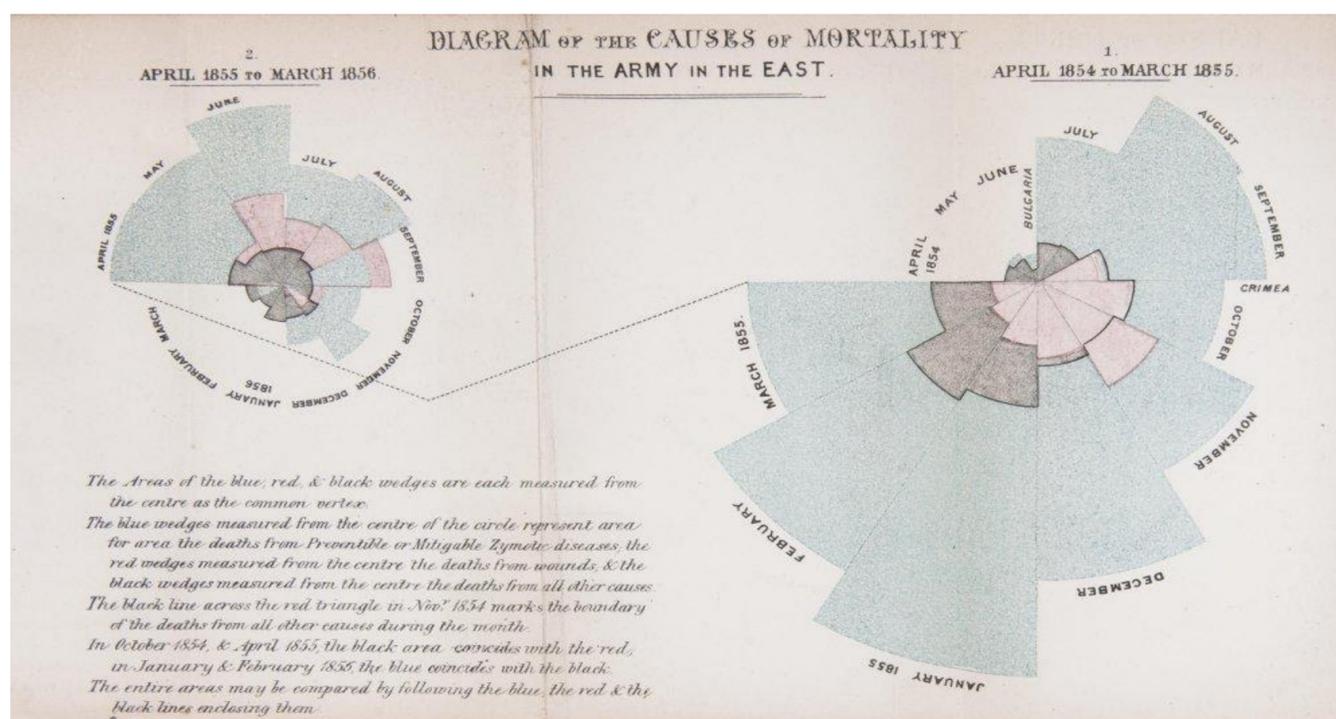


Figure 1: Diagram depicting cause of mortality prepared by Florence Nightingale [2]

Lady with The Data

Now let us try to understand the maths behind the coxcomb diagram presented by Florence. Let us assume that there were X patients who were admitted on average in the hospital. Now let us consider that Y patients out of the total died because of the disease and unhygienic conditions prevalent in the hospital. So, Nightingale assumed that in a month Y out of X died because of bad conditions. So, in every 1000 patients who are admitted to the hospital, $1000 * Y/X$ die because of the former. Then we can further see that in a year $12000 * Y/X$ patients die. So, in the coxcomb diagram (Figure 1) which you are seeing, in each month around $1000 * Y/X$ patients out of every thousand admitted die in the hospital. So, Florence calculated the “annual rate of mortality per 1000”. This is the blue area (Figure 1) which can be seen in the diagram. Now let us see how the data calculated can be represented in a coxcomb.

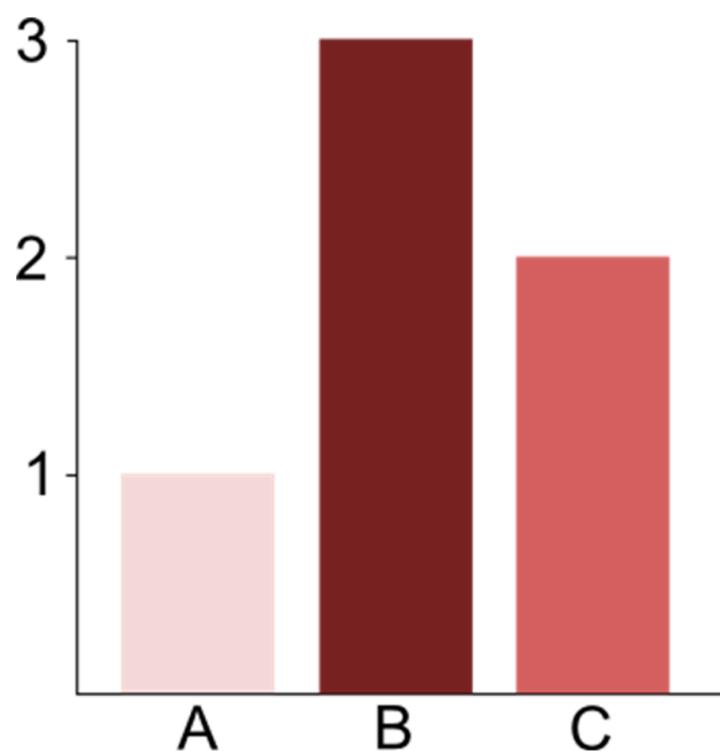
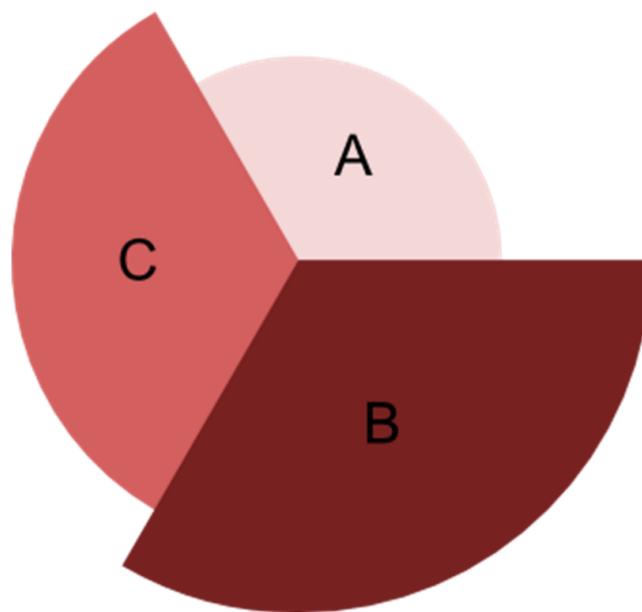


Figure 2: A simple bar chart

In the above diagram (Figure 2) you can see that we are representing three categories A, B and C in the bar chart. One can clearly see that the value of A is 1 unit, B is 3 units and C is 2 units. Now we are going to represent the same using a coxcomb



Now, we see that each sector of the coxcomb represents a category under study. (Figure 3) [3]. The angle which each sector projects is equal, i.e.

Angle projected by the sector is = $360/\text{number of categories under study}$.

For the above data, the angle projected by each sector is $360/3 = 120$ degrees while if we look at the coxcomb by Florence, the angle projected by each month is equal to $360/12=30$ degrees [3].

Now what differentiates sectors from each other, is the area which each sector covers. The area covered by sector A is based on the radius of each sector.

Let us assume that the area of sector A is a.

So, we calculate area as

$$\text{Area of sector A} = \pi a^2 = 1$$

Similarly, the area of sector B and C is

$$\text{Area of sector B} = \pi b^2 = 3$$

$$\text{Area of sector C} = \pi c^2 = 2$$

We can solve these equations to determine A, B and C enabling us to plot the coxcomb.

Lady with The Data

If we study, we find that the coxcomb represents a pie chart, but here the angle of each sector, representing the different categories, are equal but the area of the sector differs as the value of the sector changes, which represents a bar chart.

Nightingale's "coxcomb" diagram was a modern and creative approach to statistics and graphical presentation. It offered statistics as a predictive tool for medical care and treatment and brought forward what can be termed as a statistical revolution. Her work became the foundation of modern bioinformatics - the use of statistics and highly extensive data processing and bringing a system to diagnose, predict, and recommend medical treatment options worldwide. When Nightingale departed for Scutari, she would have ever imagined that her work would lay the foundation for a new technology, that her experiences would result in something exceptional and would change how people practically dealt with such situations; yet that is what happened. The separate courses of events of her life amalgamated into an extraordinary achievement in the field of Statistics in the life of a single person.

Nightingale left behind for people a mind-set to bring revolution in the world of Statistics. From coxcombs to pie charts and more, the world has been introduced to better and easier methods. Yet Nightingale has been an inspiration for all statisticians who are working to make understanding of data easier and bringing forth answers to questions that earlier muted the higher authorities.

References:

[1]https://en.wikipedia.org/wiki/Florence_Nightingale

[2]<https://towardsdatascience.com/a-history-of-polar-area-coxcomb-rose-charts-how-to-make-them-in-rs-ggplot2-a1b1ce2defd1>

[3] <http://understandinguncertainty.org/node/214>

Articles from STUDENTS

Strength lies in numbers, does wisdom?

Tathagata Benerjee (1st Year)

"With most things, the average is mediocrity. With decision making, it's often excellence."

~James Surowiecki (2004)

In our world, it is often perceived that decision-making is best left in the hands of experts and policymakers. It is a common, seemingly obvious assumption that a small group of trained individuals will always produce better results than the crowd, a mass of average people, right? Some intriguing real-life experiments, such as the Parable of the Ox and the Jelly-Bean Jar Experiment show that it isn't always necessarily true, like most other perceptions.

The Parable of the Ox

In 1907, Francis Galton, an English statistician, observed a competition of guessing the weight of an ox in a village fair [1]. He got hold of the 787 tickets, containing the guessed weights by the people who participated, and obtained the average guess, which was taken to be the median of the observed values (1207 lb.), remarkably close to the actual weight of the ox (1198 lb.). That was surprisingly accurate, even closer to the guesses made by professional cattle experts. Galton mentioned it in a letter to the famed scientific magazine *Nature*, titled 'Vox Populi' (voice of the crowds), bringing into light the concept of "Wisdom of Crowds" for the first time.

It is rumored that the average of the guesses was so accurate that the owners of the fair often used the crowd estimates if the weights broke down!

Jelly-bean Jar Experiment

In a famous experiment conducted by statistician David Spiegelhalter [2] and mathematics communicator James Grime, a picture of a jar containing numerous jelly beans was uploaded on YouTube and the viewers were asked to guess the number of jelly beans in the jar. The 915 observations recorded ranged from 219 to 31,337. The data distribution was observed to be highly positively skewed, causing the arithmetic mean (2,408) to be slightly unfit as a representative value. The mode of 10,000 showed a tendency of people to choose rounded values, but the median (1,775) proved to be quite a good guess. The true

Strength Lies in Numbers, Does Wisdom?

number of beans was 1,616, lying in the 45th percentile. While not astounding a result as Galton's ox, this result was merely a 9.84% overestimate of the true value. Only about 1 in 10 people guessed as accurately, showing that the crowd's assumption was fairly good even in this experiment.

So, are these mere coincidences? According to the hypothesis of 'The Wisdom of Crowds', they aren't. Instead, these are the mere results of statistical analysis that demonstrate the 'wisdom' possessed by crowds of individuals, who quite often provide better answers or opinions when compared to experts. The answer of the 'crowd' is generally obtained by finding the average (a suitable measure of central tendency) of all the answers provided by the individual members of the crowd.

What is 'Wisdom of Crowds'?

Wisdom of Crowds is the ideology that large groups of people are collectively smarter and more accurate than individual experts regarding problem-solving, decision making, etc [3]. Although popularized by James Surowiecki's book, "The Wisdom of Crowds", this concept dates back to ancient Athens and was also observed by Francis Galton in the event mentioned earlier. The essence of the idea is that the average judgement of a mass converges upon the right solution.

Although intriguing, the idea doesn't work well for all variations of crowds. For a crowd to be 'wise', it is a necessity that it be diverse, as well as independent. If it doesn't, the experiment would very likely lead to an undesirable bias. Also, the crowd must be chosen via random sampling, for it to

correctly represent the mass, and to produce the best results.

The statistical secret behind crowd wisdom

Over the years, statisticians have wondered what the real reason was behind these eerily accurate results, was this a miracle or did it have some proper scientific justification?

It has been found that the factor that mattered most regarding this was the diversity and independence of the crowd; experts tend to think alike, often herding together, away from the actual values. This can be due to the influence of thoughts of other experts, or overconfidence.

Strength Lies in Numbers, Does Wisdom?

At the same time, the diverse thought processes of different individuals of a crowd allowed them to make varied choices, ranging both near and far from the real value. In practice, the variations due to the choices about the actual value usually cancel each other out mostly, causing the representative value of the crowd (usually a measure of central tendency) to land very close to accuracy. This is usually improved with more diversification of the crowd, by the inclusion of people of different ages and social backgrounds. This is even applicable for most decision-making procedures, the varied thought processes of a large number of individuals usually producing a better choice than most experts ever would.

Use of crowd wisdom

The phenomenon of crowd wisdom, although seemingly unlikely, isn't as uncommon as it might seem. It isn't simply restricted to certain statistical experiments or particular situations; it is pretty effective in real life as well [3].

Customer feedback surveys are a very common example, where companies seek responses about their products/services from a wide range of customers and implement the changes commonly suggested by the people. In this way, decisions of a mass can serve as improvements over those initially made by the company's experts.

Public opinion polls are another important example of this phenomenon, ranging from the headline-causing elections to trending social media polls about which movie is the greatest of all time. The choices of an enormous population are taken, and the majority of the same is declared as the winner of the same. Participatory decision-making is another important aspect of the same, where the decision of the majority of a group is selected for solving any problem and achieving development goals. And who doesn't know, majority is another name for mode, just one measure of the average?

Is group intelligence a solution to all problems?

The answer is absolutely no. While the wisdom of crowds is a fascinating idea that produces reasonably accurate results in many cases, there are several situations where it falls short of individual prowess as well. [5]

Strength Lies in Numbers, Does Wisdom?

When the individual members of a crowd are influenced by each other's guesses, i.e, their thinking isn't independent, this is a common occurrence. Humans like to go with the flow, hence their choices end up very close to each other, often away from the accurate value, defeating the entire purpose of the wisdom of crowds. If we simply wished for close, similar answers, a group of experts would always be preferable.

In cases where the crowd is not very diverse, a similar problem occurs. The like-minded crowd, driven by common thoughts and bias, often end up with similar yet undesirable results, away from the actual value.

Another common type of situation where even a perfect, wise crowd often fails when the problem is a complicated one, involving a puzzle that requires a particular skill or knowledge to solve, or when the root cause of the problem is to be diagnosed, often referred to as the 'silver bullet'. Experts fare way better in these kinds of cases. That's why, in an airplane at 30,000 feet, we usually trust the pilot rather than an opinion poll among the passengers regarding how to fly a plane.

Conclusion

Wisdom of crowds is a fascinating phenomenon, a reality despite its elusive traits [4]. From ancient days till date, humans have operated and survived in crowds. In doing so, we observe that a decision of the mass often proves to be superior to that of the few.

However, like the individuals constituting them, crowds are often imperfect. If criteria like diversity, independence, etc. are maintained, they can act with intelligence, even wisdom, but if they're removed, crowds can become flawed, foolish, or even lethal.

The crowd is constantly evolving, striving towards perfection, like humanity itself. In today's ever-changing world, it is viewed that crowd intelligence, combined with individual prowess, is the best path to a golden future. The combined implementation of the two is the key we hold towards progress.

Strength Lies in Numbers, Does Wisdom?

References:

- [1] Collins, Rod, *The Wisdom of Crowds: Myth or Reality?* Retrieved from <https://optimity.advisors.com/insights/blog/wisdom-crowds-myth-or-reality> on February 6,2020
- [2] Spiegelhalter, David(2019). *The Art of Statistics: Learning from Data*. Great Britain. Pelican Books.
- [3] Surowiecki, James(2004).*The Wisdom of Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. USA. Doubleday Anchor.
- [4] Yong,Ed(January 31,2013). *The Real Wisdom of the Crowds*. Retrieved from <https://www.nationalgeographic.com/science/phenomena/2013/01/31/the-real-wisdom-of-the-crowds/> on February 6,2020
- [5] Hopkins, Mark(September 16,2008).*Wisdom of Crowds Isn't the Answer for Everything*. Retrieved from <https://mashable.com/2008/09/15/wisdom-of-the-crowds-isnt-the-answer-for-everything> on February 7,2020

Our PROFESSORS



From left: Prof. Debjit Sengupta, Dr. Ayan Chandra, Dr. Durba Bhattacharya, Prof. Pallabi Ghosh, Prof. Madhura Das Gupta, Dr. Surabhi Dasgupta and Dr. Surupa Chakraborty

Our STUDENTS



Third Years



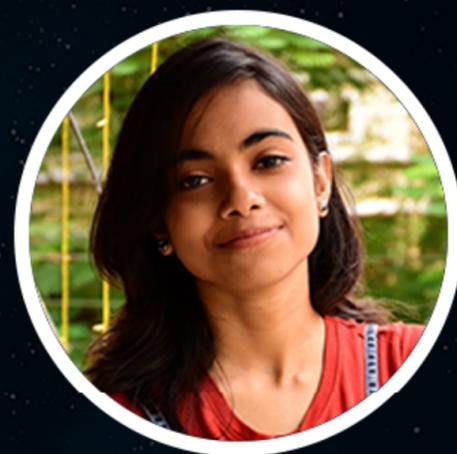
Second Years

EPSILON DELTA 2021

Student Committee



Soham Biswas
Student Convenor



Amrita Bhattacharjee
Student Co-convenor



Srijan Sen
Student Editor,
Prakarsho Vol XIII



Rajnandini Kar
Student Associate Editor,
Prakarsho Vol XIII



Supratim Pal
Student Cultural Head



Somjit Roy
Student Event Head