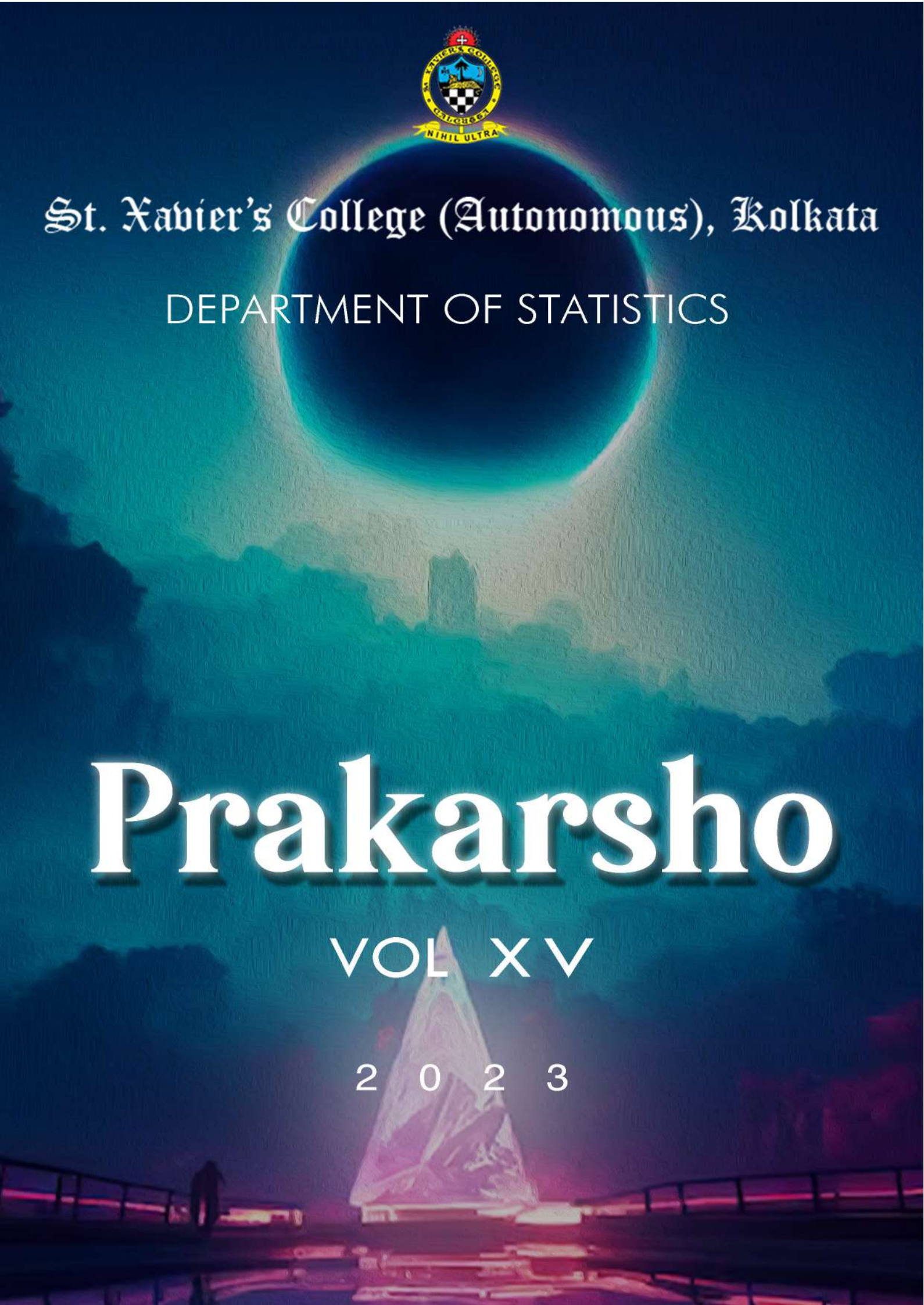

St. Xavier's College (Autonomous), Kolkata
DEPARTMENT OF STATISTICS
Prakarsho
VOL XV
2023

Department of Statistics

Faculty

"A frame full of pillars of this department, inspirations for the students, examples of the resilience. They are beyond just professors; they are ones who hold this department stronger. We are grateful for a lifetime to these mountain-like human beings, who always stand with their students, their own family."

St. Xavier's College (Autonomous), Kolkata

DEPARTMENT OF STATISTICS

Prakarsho

VOL XV

2023

Prakarsho

VOL XV

2023

SCAN
TO GRAB YOUR
E-PRAKARSHO

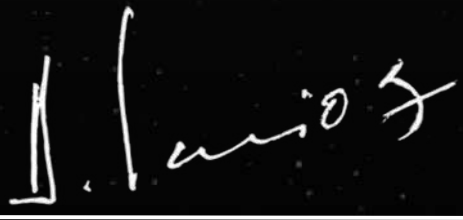
Message from the Principal

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

"I am pleased to learn that the Department of Statistics of our college is successfully publishing the 15th edition of its annual departmental magazine, "PRAKARSHO".

Since its initiation, the magazine has served as a platform for budding statisticians and data scientists to showcase their passion and genius in the form of articles that are published in it. With the addition of the postgraduation in Data Science, the merit of the magazine only grows stronger.

My heartiest congratulations to all faculty members and students, and I wish them the best for this issue as well as for their future ventures. God Bless you all! Nihil Ultra!"



Principal

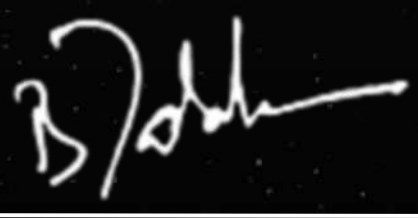
Message from the Vice-Principal

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

"Like every year, this time again the Department of Statistics St. Xavier's College(Autonomous), Kolkata has proved its role in the development of research and innovation by showcasing its commitment and fervour in the latest edition of its magazine, PRAKARSHO Vol XV.

Always striving for excellence, it brings forward riveting statistical and data-related ideas from the students as well as renowned personalities. Their hard work and zeal towards the betterment of the globe with academic persistence is truly commendable.

My heartiest congratulations to all faculty members and students. I wish them success."



Vice Principal

Message from the Dean of Science

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

"The Department of Statistics of St. Xavier's College(Autonomous), Kolkata, being an integral part of the college has filled me with extensive pride with the publication of the 15th edition of its annual departmental magazine PRAKARSHO.

The students and the faculty members continue to bring forward the best of both worlds, i.e., Statistics and Data Science and have once again showcased the strive for excellence of the department. I would like to extend my sincere congratulations to the faculty members, the editorial committee and the students.

It is gratifying to see yet another batch preserving the legacy of the esteemed department by stepping outside the bounds of curriculum in search of knowledge and I feel assured of the success of PRAKARSHO XV. Good luck!"

Dean of Science

Message from the Head , Department of Statistics

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 15th edition of the Departmental Magazine, PRAKARSHO. Yet again, we have been able to reflect the spirit of our department in the magazine, which would not have been possible without the relentless determination and untiring efforts of the students.

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 Editorial Board and the 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."

Durba Bhattacharya

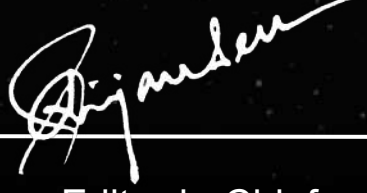
Head of the Department

Message from the Editor's Desk

"With time, we are realizing the significance of being able to predict the future. We, as statisticians or data scientists, are responsible for extracting the knowledge out of the pre-existing and continuously occurring events. But with growing diversity in the universe and fast changing world, it is difficult to use all conventional processes of analysis. We are, definitely, in need of modern research and research methodologies to cope up with the challenges coming. And in no way this magazine is an exception from seeking the help of new research from young minds.

Even during the pandemic our magazine did not stop from its responsibility of delivering modern thoughts from the new-age learners. After battling with the deadly virus for more than a couple of years, it is our immense pleasure to bring to you the fifteenth edition of our generational platform of research, thinking, applications and exchange of knowledge, the PRAKARSHO, 2023.

Nihil Ultra!



Editor-in-Chief

Editorial Board

Editorial & Designing Board

Editorial

Srinjoy Chaudhuri
Ranit Sarkar
Swapnanil Basu
Subhajit Karmakar
Soham Choudhury
Priyadarshini Ghosal
Aritra Bhattacharya
Bishwayan Ghosh
Devika K V
Ishan Datta
Abhay Ashok Kansal
Tamasha Dutta

Sunanda Maity
Ayan Saha
Shieladitya Basu
Tulip Baid
Parthiv Pratim Das
Reetika Choudhury
Manav William
Koena Dey
Anushka Bose
Sreeja Choudhury
Pratya Bhukta

Designing

Shamie Dasgupta
Sambit Ghosh

Arka Roy
Sreyasi Dey

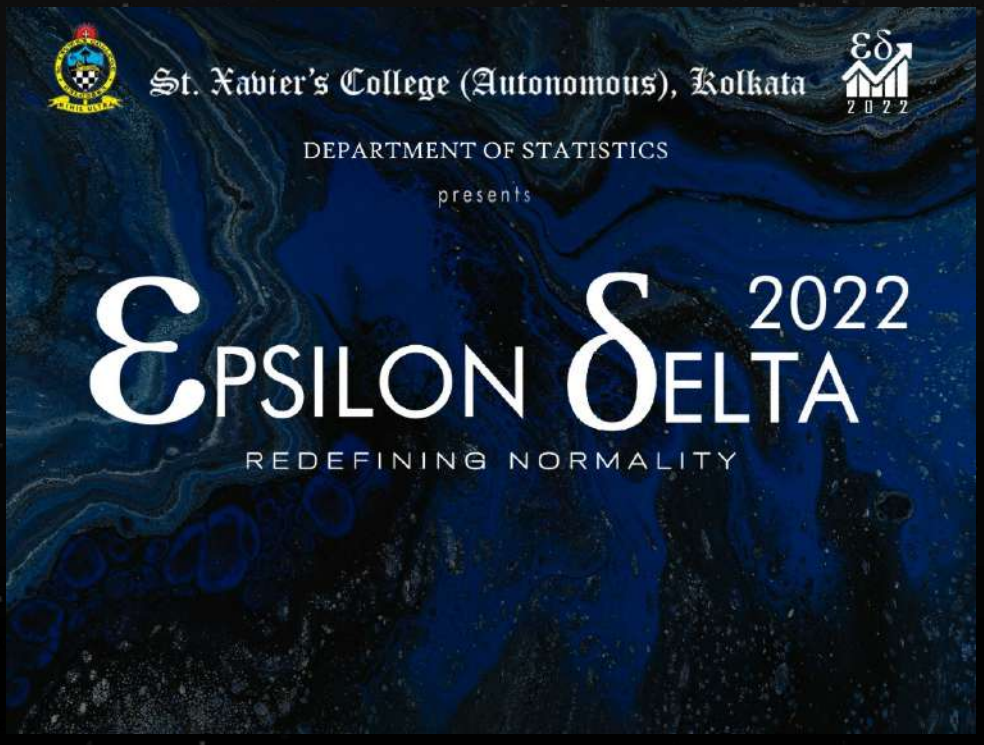
Departmental Report

2022-2023

Departmental Activities

- **Epsilon Delta 2022**

The Department organized its annual departmental event, Epsilon Delta on 14th March, 2022 on the virtual platform MS Teams. The program commenced with the launch of the 14th edition of the Departmental Magazine 'Prakarsho'. The department organized 'PROECTURA' – an inter- college paper presentation event where students from colleges all over India participated and presented their research papers and ideas. This event was followed by the events 'X-QUIZZIT' and 'CHECKMATE'. The event concluded with a short cultural presentation, performed and compiled by the students of the Department of Statistics. Over 200+ students virtually attended the event.



Departmental Report

2022-2023

Departmental Activities

- **Inauguration of Data Science Lab**

After completing more than sixty glorious years of undergraduate teaching, the Department of Statistics has started offering MSc in Data Science from the academic year 2022-23. The Data Science Laboratory was inaugurated on 4th July 2022.

Departmental Report

2022-2023

Departmental Activities

- **Webinar on "Artificial Intelligence in Healthcare"**

Dr. Shibasish Dasgupta, Associate Director of Quantitative Data Science, Pfizer, Adjunct Professor in Statistics and Data Science, Chennai Mathematical Institute (CMI), was the invited speaker of the day who presented an enlightening and informative talk on 'Artificial Intelligence in Healthcare'.

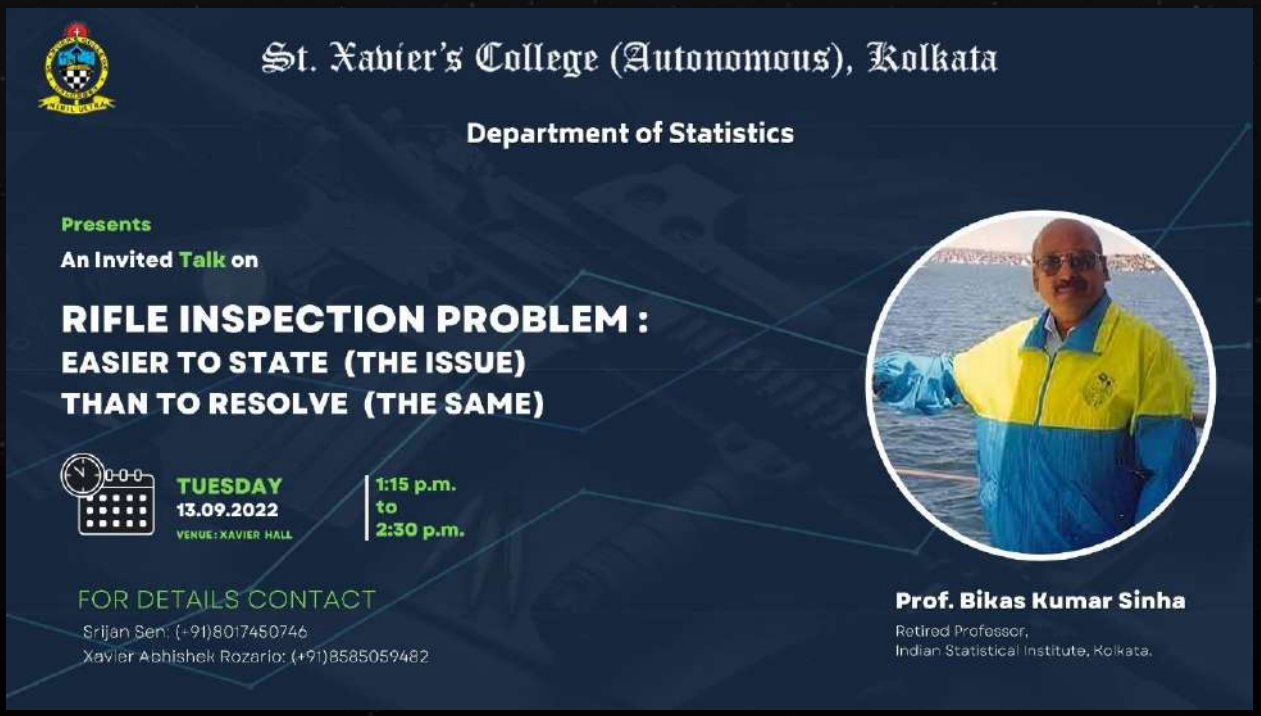
Departmental Report

2022-2023

Departmental Activities

- **Talks**

Prof. Bikas Kumar Sinha, Retired Professor, Indian Statistical Institute delivered an invited talk on "FANCY LIFE OF A SMART GAMBLER" and "OH CAPTAIN! MY CAPTAIN" and "RIFLE INSPECTION PROBLEM: EASIER TO STATE (THE ISSUE) THAN TO RESOLVE (THE SAME)" on 13th September,2022.



Departmental Report

2022-2023

Departmental Activities

- **Talks**

Prof. Soutir Bandyopadhyay, Director of Graduate Studies, Department of Applied Mathematics and Statistics, Colorado School of Mines, delivered an invited talk on “Wendland Meets Markov: Kriging For Large Spatial Data” on 15 th July,2022.

SMART INVENTORY MANAGEMENT USING COMPUTER VISION

Somedip Karmakar,
Staff Data Scientist,
Walmart Global TechIndia

❖ **Abstract:**

In any retail business, it is very important that the customers are able to find the required items on the shelf. Out-of-stock scenarios can be a major cause of customer walk-offs. It is a critical part of assortment and replenishment domains to ensure that products are always available, and proactively determine situations of out-of-stock even before it happens, so that the items can be restocked. However, for very large stores with millions of items, it becomes very difficult to manually keep track of all the items on the shelf, and hence a computer vision-based solution can automate the task of determination of potential out-of-stock and an integrated system can be developed to replenish the items from the back room or raise alerts to the replenishment managers to deliver the next batch of products. The process can also be utilized to identify cases where wrong items are placed on the shelf, leading to the unavailability of the products, and thereby reducing shrinkage. The store shelf images, gathered from shelf-scanning robots, drones and cameras are stitched together to get a view of the current status of shelf inventory, and the planogram provides information on the actual number of items required to be present. We have discussed an innovative use of image augmentation, unsupervised image processing and semi-supervised deep learning-based localized void detection algorithm to overcome the challenges of the requirement of labelled data.

The algorithm can detect actual voids and partial voids present in shelves. The solution is highly scalable and accurate and can be implemented for a wide variety of products without any retraining. The paper covers the various challenges which we overcame, and also showcases the model performance on sample shelf layouts.

❖ **Keywords:**

Computer vision, unsupervised models, machine learning, deep learning, semi-supervised, void detection, image similarity, image embeddings, replenishment and inventory management.

❖ **Introduction:**

In any retail business, it is very important that the customers are able to find the required items on the shelf. At retail stores, out-of-stock shelves inevitably reduce sales and customers cannot be considered a temporary loss. Some survey results claim that in case of out-of-stock, 31% of the customers would purchase in a different store, and 9% of customers do not purchase any products. Since checking stock is such an important task, stores generally increase the frequency of checking, but this also increases the time spent by staff. The challenge for store owners is how to create an efficient process for checking the shelves. For very large stores with millions of items, it becomes very difficult to manually keep track of all the items on the shelf, and hence a computer vision-based solution can automate the task of determination of potential out-of-stock and an integrated system can be developed to replenish the items from the back room or raise alerts to the replenishment managers to deliver the next batch of products. In some situations, the customers can decide they no longer need a product and can put it in a different store location than the original. This is a serious problem, since the product becomes unaccounted for, and might be lost or stolen. Usually, the back room of the stores would maintain the inventory required to replenish the empty shelves, and such restocking usually happens across the stores during off-hours for all categories. There can be some high-moving items which can get over in a short period of time, and customers would not be able to find the required product, even when it is available in the back room. As a retailer, the end goal is to increase sales, improve overall category profitability and become a store that caters for its customers so that they'll keep coming back. By ensuring that the shelves are well stocked, the retailer can create a culture of accountability in their stores and increase customer satisfaction by manifolds. With the advances in technology, computer vision and artificial intelligence can come to the rescue of retailers. There are robots which can scan the shelves, drones which fly overhead, and even static cameras which can take images at regular intervals. The images from these devices are passed to a central database, where they are processed, and stitched together to recreate the actual implementation on the store shelves.

Column of ALUMNUS

Our system utilizes computer vision to get a real-time feed of the store shelves and can determine beforehand if a particular product is out-of-stock or is tending towards out-of-stock, and then generates an alert to the store associates to replenish the required products. In some cases, the replenishment process could also be automated using drones or robots in the store. Usually, the store-specific planogram details like product name, and horizontal and vertical facing quantities, would be available beforehand in the central data storage. Now our system can utilize the pre-defined information, and stitched store shelf images, to determine exact or partial out-of-stock scenarios in real-time. There has been some work in the domain of out-of-stock detection, however, all such work is heavily dependent on supervised learning, with tagged data, and none of them considers partial out-of-stocks. It is very difficult and too expensive for a large retailer to tag millions of items and build a supervised learning framework. That is where our novel technique and system are very crucial.

The rest of this paper has been divided into the following sections: 1. Introduction 2. Overview of the system, 3. Ensembled Out of Stock Detection Framework, 4. Modelling & Inferencing Framework, 5. Advantages of the system, 6. Results and Model Accuracy. Each section covers the technical aspects, implementation, and benefits.

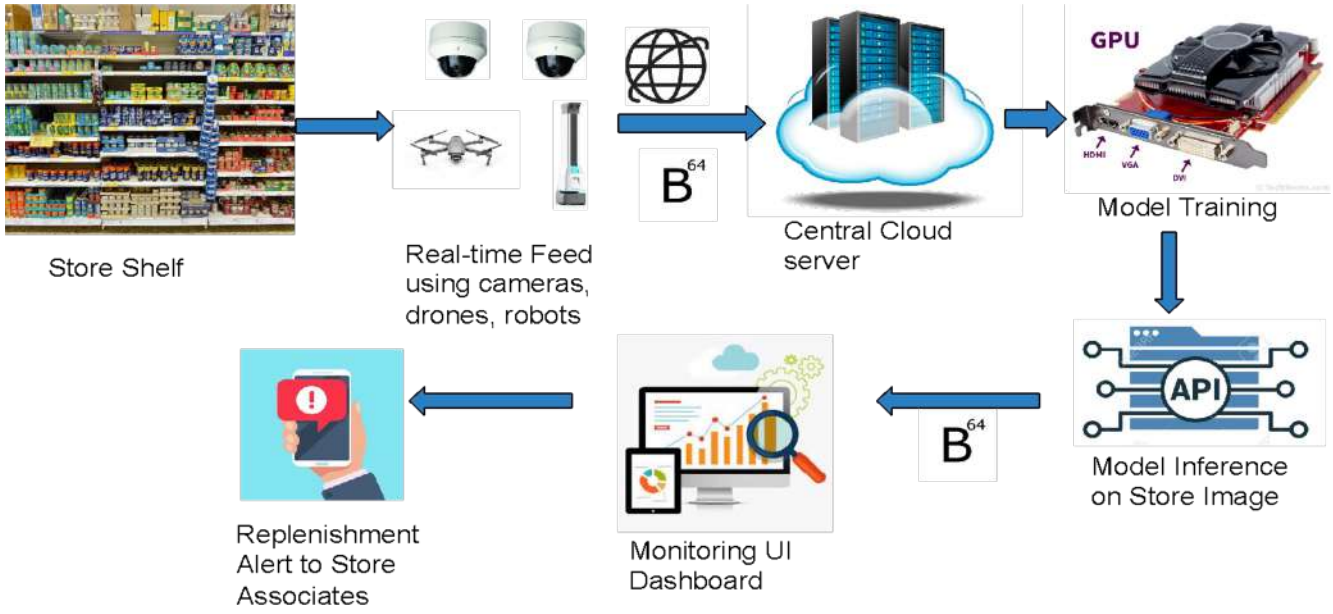
❖ **Overview of the system:**

Our system is primarily developed as a real-time Artificial Intelligence system, which keeps capturing feeds of images from either drones, shelf-scanning robots or static cameras, stitches together the shelf layout, runs a series of models for out-of-stock detection and finally generates a report through an automated alert to the responsible store manager or associates to replenish the items. The input to our system mainly consists of Shelf images from Stores, Planogram details like product names, and horizontal and vertical facing quantities. The target is to identify and detect the voids present in the Shelf image as Partial Voids or Out of Stock. Here Partial void is referred to the configuration when some of the products of a given type are sold out. Out of Stock indicates that the product is completely sold out. The images are collected from the cameras or drones are then converted to an encoded string format and then passed through an API to the central computing device hosted in the cloud.

Column of ALUMNUS

The ensemble model is deployed as a service through the API, which takes as an input the actual store image in an encoded format, then processes the image using pre-trained weights, and outputs the presence or absence of complete or partial out-of-stocks, along with the location coordinates of the void. The output image, with the demarcated voids, is encoded and sent back to the store systems as alerts. The model training happens in a batch process at regular intervals to keep updating the model parameters with the new data received.

Following is a diagrammatic overview of our system.



The images from the drones or cameras, once stitched are passed on to the system mentioned above. For each image frame, the steps are run iteratively and then a comprehensive report on the shelf availability is generated for regular monitoring. If there are instances of partial or completed out-of-stock present in any particular aisle, then an alert would be sent across to the particular store associates or the store manager, on their mobile devices or through email, with details on which products need to be replenished. The system also cross-references with back-room inventory to decide if the supply chain manager needs to be alerted to expedite the next batch of shipment.

❖ **Ensembled Out-of-Stock Detection Framework:**

For the system to work very fast and be scalable, we need to optimize the performance. Also, we cannot rely upon only one variant of the model, since it needs to be generalized across all types of products in the store. We can use a variety of supervised techniques to create an Ensembled Out-of-Stock framework. We have the following main algorithms:

❑ **Region-Based Convolutional Neural Networks:**

The goal of R-CNN is to correctly identify the regions of the main object in the image by proposing bounding boxes having objects and classifying them accurately. The approach proposed to apply high-capacity convolutional networks (CNNs) to bottom-up region proposals to localize and segment objects and when there is minimal supervision, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly.

❑ **Masked Region-Based Convolutional Neural Networks:**

This solves Instance segmentation problems in a 2-stage framework. In the first stage, it detects the bounding boxes and in the second stage predicts the object class and generates a mask at the pixel level for the object. Masked R-CNN presents a simple, flexible, and robust framework for object instance segmentation which efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance.

❑ **Single Shot Detection:**

SSD, need to take one single shot to detect multiple objects within the image and is much faster compared with two-shot RPN-based approaches. The approach presents an object detection framework using a single deep neural network by discretizing the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. During inference, the model generates scores for the presence of each object category in each default box and suggests robust and efficient adjustments to the box and enhances the mapping.

These steps are followed in sequence to properly identify the most probable region where the out-of-stock can be present. This ensemble framework ensures high accuracy of the model with very fast response time of the model within milliseconds.

❖ **Modelling & Inferencing Framework:**

In this section we would discuss in more details the actual model training and inferencing framework to determine out-of-stock scenarios. Broadly we follow the below-mentioned series of steps:

❑ **Dataset Preparation and Augmentation:**

For our case, the primary challenge is availability of good quality labeled images with complete and partial out-of-stock present. We overcome this challenge through a novel data augmentation technique. The primary task comprises i) Dataset creation and formulation in an Object Detection framework, ii) Dataset Augmentation for Object detection. The augmentation strategies for object detection tasks are much more complex than in simple classification tasks as we must keep a track of the position of the object while rotating and translating the image. We leverage concepts from [6] which helps us in learning and augmenting high-quality data with limited features.

❑ **Feature Extraction:**

When it comes to working with deep learning models there is no explicit need for extracting the features from the data to train the models as compared to traditional approaches. Convolutional Neural Networks act as feature extraction layers and these features are used downstream like in this case detecting void regions in planogram. Convolutional features are used for classification as well as localization for the task in hand and sometimes features from multiple layers are also used to make the network predict accurate outcomes irrespective of the object size.

❑ **Model calibration:**

The model calibration is done by minimizing on deviation: i) Where the object actually is (location loss), ii) What is the object (class loss)

These techniques help in increasing confidence of detection of out-of-stock. This is further validated with the information present in the planogram about the number of products planned to be present in the shelf to accurately determine the partial out-of-stock and predict beforehand the estimated time when the product will go out-of-stock based on the rate of purchase.

❖ **Advantages of the System:**

We have faced multiple challenges while building our system, and we have made our system resilient to these. The primary advantages of the system are following:

❑ **Semi-supervised Out of Stock Detection Methodology:**

- Very few examples of void images are fed as input to the model.
- Created an intelligent data augmentation framework to enhance the training set intelligently.
- Ensembled model approach to enhance the semi-supervised performance by capturing the complimentary information.

❑ **Detection of Partial Void and Out of Stock Detection in very less time:**

- High accuracy of the model helps in detecting both the Partial void and Out of Stock.
- It is very crucial for businesses to understand when there is an out of stock or partial voids so that they can take immediate actions and our model is efficient and can inference in very less time.
- It is scalable and can be implemented for any Shelf images and it will be able to identify the partial voids and out of stock.

Column of ALUMNUS

Our system also works under different lighting conditions in the store and is robust to partial image presence, and approximate matching.

❖ **Results and Model Accuracy:**

We have tried out our models on many different categories of products. Overall accuracy of the model is around 90 % which is quite high, given that it is a semi-supervised model.

Here are the model results for a few sample shelf-images:

This validates our model accuracy on a variety of product images.

Semi-Supervised Void Detection Results		
	No of Voids Detected by the model	Actual Void Count
HRes_DJI_0369.jpg_0000_orig	6	6
HRes_DJI_0369.jpg_0001_orig	1	1
HRes_DJI_0369.jpg_0002_orig	1	1
HRes_DJI_0369.jpg_0003_orig	5	6
HRes_DJI_0369.jpg_0004_orig	3	3
HRes_DJI_0369.jpg_0005_orig	2	3

❖ Conclusion:

We have implemented the solution for the US with Store Shelf images provided by Drone cameras in Stores, Dec 2019.

Business Impact:

- The model helps in providing an automated alert to the store personnel whenever there is an out-of-stock scenario, so that the necessary steps can be taken else an out of stock in a shelf is responsible for customer dissatisfaction.
- It helps the business in merchandising, replenishment & assortment of decisions effectively.
- Solution has been shared with the International Store Operations business and is in the process of getting implemented in the pipeline under Image Analysis.

This model will ensure better compliance of pre-emptive out-of-stock detection which will have significant uplift in incremental sales and improve customer experience. Estimates on optimal planogram showcase potential of around 10 % lift in incremental sales, from predicted demand models.

❖ Acknowledgements:

We are grateful to the Business partners in the US and International Store Operations team for their support and guidance, and for providing us with necessary product images, Shelf images and planogram information.

❖ **References:**

- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 1. Curran Associates Inc., USA, 1097-1105.
- Murinto, Murinto & Prahara, Adhi & Winiari, Sri & Pramudi Ismi, Dewi. (2018). Pre- Trained Convolutional Neural Network for Classification of Tanning Leather Image. International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.090129.
- R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 142-158, 1 Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.
- K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.
- Liu, Wei & Anguelov, Dragomir & Erhan, Dumitru & Szegedy, Christian & Reed, Scott & Fu, Cheng-Yang & Berg, Alexander. (2016). SSD: Single Shot MultiBox Detector. 9905. 21-37. 10.1007/978-3-319-46448-0_2.

AN OVERVIEW ON MARKET BASKET ANALYSIS

Sukanya Mukherjee, Kankana Ghosh

1st Year, M.Sc. in Data Science

❖ Introduction:

Many of us have visited online retail stores such as Amazon and Flipkart to meet our daily needs. What we typically do is, search for the item, select the product and head towards the billing counter to purchase it. But in today's world the goal of any organization is to increase revenue. Can this be done by pitching one product at a time to the customer? The simple answer to it is no. Hence, organizations begin mining data related to frequently bought items. So, market basket analysis is one of the key techniques used by large retailers to uncover association between items.

It is typically a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analysing large datasets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together. An example would be

that a customer who would purchase a laptop would more likely purchase a laptop bag along with it.

There are two types of market basket analysis:

Predictive market basket analysis: This type considers items purchased in sequence to determine cross-sell.

Differential market basket analysis: This type considers data across different stores, as well as purchases from different customer groups during different times of the day, month or year. If a rule holds in one dimension (like store, time period or customer group), but does not hold in the others, analysts can determine the factors responsible for the exception. These insights can lead to new product offers that drive higher sales.

Now, if a customer buys an item A, then there is a slight possibility or chance that they might buy B. This type of relationship is called Single Cardinality. But there can be cases where the customer who buys A and B also buys C or the customer who buys A, B and C also buys D. In these cases, the cardinality increases thus increasing the number of combinations around the data/item sets. If we have 10,000 or more than 10,000 data items then there will be so many rules that we have to create for each product. Thus, Association Rule Mining uses certain measures and that is where apriori algorithm comes in.

Association Rule Mining is an efficient algorithm which helps the business make profit. It is all about building rules. It can be thought of as an if-then relationship. Just to elaborate in that, we have come up with a rule that suppose if item A is bought by the customer then the chances of item B picked by the customer too under the same transaction id is found out. It's not a casualty rather a co-occurrence pattern that comes to the force. There are two elements in this rule i.e. 'if' and 'then'. Now 'if' is also known as antecedent. This is an item or a group of items that can typically be found in an itemset. And the later one is called the consequent. This comes along as an item with an antecedent group or the group of antecedent approaches. Now A=>B indicates that if a person buys an item A, he will also buy an item B or he will most probably buy an item B.

There are 3 types of metrics which help to measure the association such as:

Support: Support is the frequency of item A or the combination of item A or B. It is the frequency of the item which we have bought, by what the combination of the frequency of the item which we have bought are. With this we can filter out the item which was bought less frequently.

$$\textbf{Support} = \frac{freq\ (A,B)}{N}$$

Confidence: Confidence gives us how often A and B occur together given the number of times A occurs. This also helps us solve a lot of problems, such as if somebody is buying A and B together and not buying C, we can simply rule out C at that point of time. According to this we can define our minimum support and confidence. After setting these values we can put them in the algorithm, filter out the data and create different rules.

$$\textbf{Confidence} = \frac{freq\ (A,B)}{freq(\ A)}$$

Lift: Lift is the strength of any rule. In the denominator we have the independent support values of A and B. This gives the independent occurrence probability of A and B. Now, if the denominator of lift is more, it means that the occurrence of randomness is more rather than the occurrence because of any association.

$$\textbf{Support} = \frac{Support}{Supp\ (A) \times Supp\ (B)}$$

If Lift (A => B) = 1, means that there is no correlation within the item set.

If Lift (A => B) > 1, means that there is a positive correlation within the item set, i.e., products in the item set, A and B, are more likely to be bought together.

If Lift (A => B) < 1, means that there is a negative correlation within the item set, i.e., products in item set, A and B, are unlikely to be bought together.

❖ **Apriori Algorithm:**

In Business sectors, according to the sales, the marketing teams have a minimum threshold value for confidence as well as the support.

It uses frequent item sets to generate association rules. It is based on the concept that a subset of a frequent itemset must also be a frequent itemset.

What is a frequent itemset?

Frequent Itemset is an itemset whose support value is greater than a threshold value.

Let us consider these Transactions where TID is the transaction ID and Items depict the i^{th} item picked where i=1(1)5.

TID	Items
T1	1 3 4
T2	2 3 5
T3	1 2 3 5
T4	2 5
T5	1 3 5

Now, we build a list of itemset of size 1 by using this transactional data and calculate the respective support values. Let us assume that the minimum support count is 2 for the organization.

Iteration 1:

C1

Item set	Support
{1}	3
{2}	3
{3}	4
{4}	1
{5}	4

Here, we see that the itemset 4 has a support 1 because in the transactional data the frequency of the item et 4 to occur is 1. As this support is less than

the minimum support count 2, so we eliminate item set 4. Hence, the final table F1 becomes,

Item set	Support
{1}	3
{2}	3
{3}	4
{5}	4

We build a list of items set of size 2 and calculate the support.

Iteration 2:

C2

Item set	Support
{1, 2}	1
{1, 3}	3
{1, 5}	2
{2, 3}	2
{2, 5}	3
{3, 5}	3

Here, we reject the item set {1, 2} as it does not satisfy the minimum support count. Hence, the final table F2 is,

Item set	Support
{1, 3}	3
{1, 5}	2

{2, 3}	2
{2, 5}	3
{3, 5}	3

Moving forward we calculate the support for the item sets of size 3 where the combinations are used from the item set F2 for this iteration. The item sets in C3 are {1, 2, 3}, {1, 3, 5}, {2, 3, 5}, {1, 2, 5}. Before calculating support, we perform pruning on the dataset. That is after the combinations are made, we divide C3 item sets to check if there is another subset whose support is less than the minimum support value that is what frequent itemset means.

Iteration 3:

C3

Item set	In F2?
{1, 2, 3}, {1, 2}, {1, 3}, {2, 3}	No
{1, 3, 5}, {1, 3}, {1, 5}, {3, 5}	Yes
{2, 3, 5}, {2, 3}, {2, 5}, {3, 5}	Yes
{1, 2, 5}, {1, 2}, {1, 5}, {2, 5}	No

Here, the item sets {1, 2, 3} and {1, 2, 5} are eliminated as the subsets are not present in the transactional data.

F3

Item set	Support
{1, 3, 5}	2
{2, 3, 5}	2

In the last step we calculate the support for the itemset of size 4.

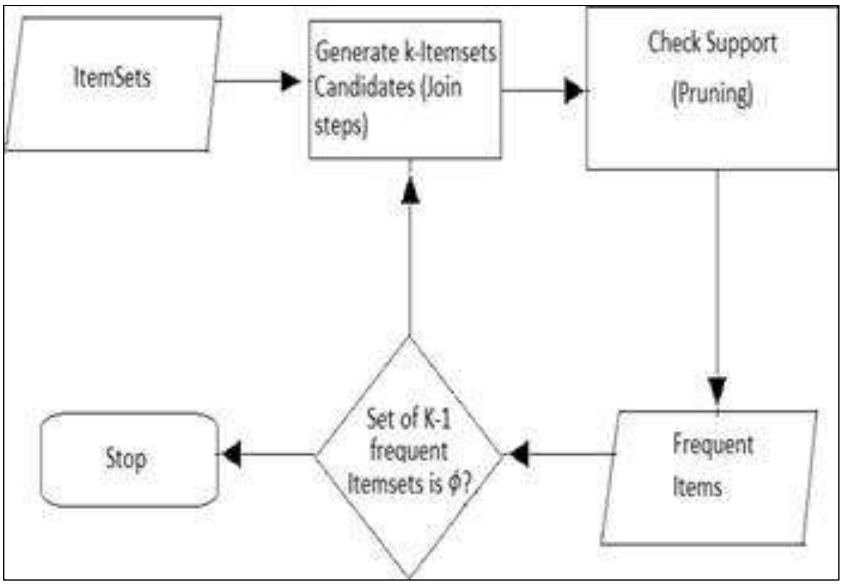
Iteration 4:

C4

Item set	Support
{1, 2, 3, 5}	1

Here, as the support is less than the minimum support count, we stop the iteration here and move to the previous item set C3.

Flowchart for discovering frequent item sets for mining Boolean Association rules:



Let's assume our minimum confidence value is 60% for the organization. To compute the confidence, we generate all the non-empty subsets for each frequent item sets.

For I= {1, 3, 5} the subsets are {1, 3}, {1, 5}, {3, 5}, {1}, {3}, {5}

For I= {2, 3, 5} the subsets are {2, 3}, {2, 5}, {3, 5}, {2}, {3}, {5}

For each subset S of I, we output the rule: S → (I-S) (S recommends I-S)

If, $\frac{Support(I)}{Support(S)}$ >= Minimum Confidence Value,

Applying this rule to itemset F3,

1: {1, 3, 5}

Rule 1: {1,3}→({1,3,5} - {1,3}) means 1 & 3 → 5

Confidence = $\frac{Support(1,3,5)}{Support(1,3)} = \frac{2}{3}$ = 66.67% > 60%

Rule 2: {1, 5}→({1, 3, 5} - {1, 5}) means 1 & 5 → 3

Confidence = $\frac{Support(1,3,5)}{Support(1,5)} = \frac{2}{2}$ = 100% > 60%

Rule 3: {3, 5} → ({1, 3, 5}-{3, 5}) means 3 & 5 → 1

Confidence = $\frac{Support(1,3,5)}{Support(3,5)} = \frac{2}{3}$ = 66.67% > 60%

Rule 4: {1} → ({1, 3, 5} - {1}) means 1 → 3 & 5

Confidence = $\frac{Support(1,3,5)}{Support(1)} = \frac{2}{3}$ = 66.67% > 60%

Rule 5: {3} → ({1, 3, 5} - {3}) means 3 → 1 & 5

Confidence = $\frac{Support(1,3,5)}{Support(3)} = \frac{2}{4}$ = 50% < 60%

Rule 6: {5} → ({1, 3, 5} - {5}) means 5 → 1 & 3

Confidence = $\frac{Support(1,3,5)}{Support(5)} = \frac{2}{4}$ =50% < 60%

Here, Rules 5 and 6 do not satisfy the minimum confidence value so we reject these two rules. Hence the Rules 1, 2, 3 and 4 can be used by the

organizations to increase their revenue. In a similar manner, we can find out the confidence for the itemset {2,3,5}.

❖ **Advantages:**

1. Helps in setting prices: Market basket analysis can point out that whenever a customer buys milk, they end up purchasing coffee as well. So, whenever the sale of milk and coffee is expected to rise, retailers can mark down the price of cookies to increase the sales volume.

2. Arranging SKU (Stock Keeping Unit) Display: Market basket analysis helps identify items that have a close affinity to each other even if they fall into different categories. With the help of this knowledge, retailers can place the items with higher affinity close to each other to increase their sales. For instance, if chips are placed relatively close to a beer bottle, customers may end up buying both. In contrast, if they were placed in two extremes, then the customer would just walk in the store, buy beer and leave the store causing lost sales of chips.

3. Identifying Sales Influencers: All items in a retail store have some relationship with each other – be it strong or weak. In most cases, the sale of one item is driven by the increase or decrease in the sale of other items. Market basket analysis can be used to study the purchasing trend of a certain SKU.

❖ **Disadvantages:**

1. Shortcomings: Although the market basket analysis is a data mining technique with considerable usability, it is by no means an infallible study of consumer behaviour.

Firstly, even if an association between products shows promising evaluation metrics, it cannot directly prove the causality between the products. After all, correlation is not equal to causation.

Secondly, like any data mining technique, the market basket analysis is prone to errors. It can falsely omit significant associations or falsely include insignificant associations. We perform our analysis keeping these shortcomings in mind, lest we draw wrong conclusions from our findings.

Thirdly, when working with large datasets the Apriori algorithm is slow, inefficient and uses a lot of resources as it has to scan the database many times, generate a large number of candidate sets and check each of them. Hence, the cost to calculate the support increases.

2. Iffy correlations: Way back before 'big data' (the '90's), a retail company ran SQL queries against its store data and discovered that beer was often purchased with diapers. This discovery quickly caught wind, and stores started to put diapers and beer nearby on store shelves. Naturally, sales of the two went up together.

The issue is, when the sales went up for the combination of items, stores were merchandising them next to one another in a highly trafficked area. The root of the problem is that we cannot seek out and validate correlations in data that we had a hand in creating. Thus, although market basket analysis may help us spot a trend but once we act on it, it becomes difficult to assess the validity of the correlation.

3. Test and learn lag times: Because there are no clear calls to action, retailers who use these solutions must dedicate time to A/B test any actions they take as a result of the data. But how do we decide which correlation to A/B test? Testing even just the strongest correlations takes enormous time and effort.

Let's say we manage to pick a product pairing that we think will generate the most revenue. There's a lot of work to make the test happen. For instance, if we're going to run an in-store cross-promotion, we'll need to re-organize our shelves, rework our planograms, and send directives down to all stores. From there, we'll need to train staff on the new locations and make them aware of the promotion. We'll then need an adequate amount of time (weeks? months?) to test whether or not we were right.

Costs aside, from learning the correlations to making changes and then testing their efficacy and moving forward by using what we learned, this entire process is slow.

❖ **References:**

1. https://www.researchgate.net/publication/351168385_Research_and_Case_Analysis_of_Apriori_Algorithm_Bas ed_on_Mining_Frequent_Item-Sets

2. https://cb4.com/blog/market-basket-

analysis/#:~:text=What%20are%20the
%20Limitations%20of,still%20leave%2
0room%20for%20improvement.&text

=Averages%20tend%20to%20lie.,ll%2
0hit%20some%20speed%20bumps.

3. https://towardsdatascience.com/understanding-consumer-behavior-with-the-market-basket-analysis- 3d0c017e5613

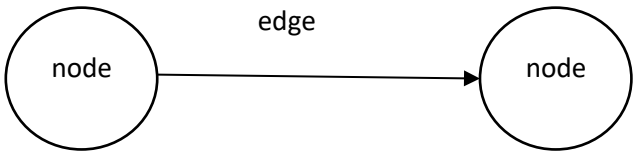
CAUSAL MODELS

Soham Choudhury, 2nd Year
Ranit Sarkar, 2nd Year

❖ Introduction:

Mathematical models which represent causal relationships within an individual system or population are known as Causal models. Causal models are used to make inferences about causal relationships from statistical data. A causal model makes predictions about the behaviour of a system. Several types of Causal models can be developed such as Structural Equation Models, Probabilistic Causal Models, etc.

Generally, Causal models have two components - node and edge.



In Casual models, we need at least two nodes and one edge.

The node represents variables and the edge represents the causal relationship between the variables.

❖ Variables:

Variables are the basic building blocks of Causal models. The variables can be nominal or numeric. The values of a variable can represent the occurrence or non-occurrence of an event, a range of incompatible events, a property of an individual or of a population of individuals, or a quantitative value. Suppose we want to model a situation where a killer shoots a bullet, and a person dies. We define our variables A and B in such a manner that,

A=1 represents the killer shooting a bullet, A=0 represents the killer not shooting

B=1 represents the person's death, B=0 represents that the person remains alive

❖ Types of Causal Models:

Causal models can be of various types. Some major types of Causal Models are –

❑ **Graphical Models:**

Graphical Models display causal relationships between variables through a directed graph. Here each variable is connected by an arrow to one or more other nodes upon which it has a causal influence. Here, an arrowhead describes the direction of causality, e.g. an arrow connecting variables X and Y with the arrowhead at Y indicates that a change in X causes a change in Y.

❑ **Rubin Causal Model:**

Rubin's Causal Model is an approach to the statistical analysis of cause and effect based on the framework of potential outcomes. This model is also known as the Neyman-Rubin **Causal Model**.

This model is based on the idea of potential outcomes. For example, a person would have a particular income at age 30 if they had attended college, whereas they would have a different income if they had not attended college. To measure the causal effect of going to college for this person, we need to compare the potential outcomes for the same person in both alternative futures.

❑ **Structural Equation Models:**

An SEM or Structural Equation Model characterises a causal system with a set of variables and a set of equations about how each variable depends upon the other. It is mostly used in social and behavioural sciences. Suppose a person wants to understand the factors that influence a student's performance in an exam. Then he might take the factors and form some equations between the factors which are related to each other.

Using SEM one may estimate the strength and direction of the relationships between variables and test the fit of the model on the data.

❖ **Important Statistical Tools in Causal Modelling:**

Various statistical analysis methods are used in Causal Models to understand the relationship between the variables more precisely.

❑ **Regression analysis in Causal Modelling:**

Regression analysis finds its use in causal modelling to identify and estimate the relationships between dependent and independent variables. Several types of regression analysis can be used in causal modelling such as linear regression, logistic regression, multivariate regression, etc. The effects of the variables that affect both the independent and dependent variables can be easily controlled using regression analysis.

For example, if anyone is interested in the relationship between exercise and weight loss, one may select some individuals, measure their weights, and study their exercise habits. Here exercise habit is the independent variable and weight is the dependent variable. Through regression analysis, the relationship between these two variables can be estimated by controlling the other factors for weight loss like diet, age, etc. One may find that individuals who are engaged in more exercise tend to lose more weight. This would imply exercise has a causal effect on weight loss.

- **Instrumental Variable Analysis in Causal Modelling:**

If there is a confounder (A variable that affects both the dependent and independent variables) present then Instrumental Variable Analysis can be used to estimate the causal effect between the dependent and independent variables unbiasedly.

If we are interested in the relationship between education and income, then we might collect data on education level (independent variable) and income (dependent variable). There may be some other variables that could affect both education and income like family background, access to job opportunities, etc. These factors could make it difficult to identify the true causal effect of education and income. Distance to the nearest school is a variable related to education but does not directly affect income. This is known as the instrumental variable that can be used to control the effect of confounders. Using the following equation causal effect of education on income can be estimated:

The causal effect of education on income = (Coefficient of education on income) x (Correlation between instrumental variable and education)

Instrumental Variable Analysis is a useful tool in causal modelling to estimate the causal effect in the presence of confounders.

- **Practical Applications of Causal Models:**

Causal Models are used in a variety of fields to make predictions, identify the cause of a particular phenomenon and evaluate the impact of interventions. Some practical applications of Causal Models are:

1. Medicine: To identify the relationship between risk factors and the likelihood of developing a particular disease, Causal Models are used. They can also be used to evaluate the efficacy of different treatment options and to predict the outcome of a particular treatment.

2. Economics: In Economics, to understand the relationship between variables such as prices, demand and supply, causal models are frequently used. They are helpful to make predictions about how changes in one variable will affect others, and to evaluate the impact of policies or interventions on the economy.

3. Social Science: Causal models are commonly used in social science to understand the factors that influence human behaviour and social systems. To evaluate the effectiveness of interventions aimed at improving outcomes such as education, health and crime, causal models are effective.

4. Marketing: In marketing, Causal models can be used to understand the relationship between variables like advertising, sales and customer loyalty. They can be used to make predictions about how changes in marketing strategies will affect sales, and to evaluate the effectiveness of different marketing campaigns.

5. Environmental Science: To study the relationship between variables like climate, land, biodiversity, Causal models are effective. They can be used to predict the impact of environmental changes on ecosystems and to evaluate the effectiveness of interventions aimed at conserving or restoring natural habitats.

❖ **Conclusion:**

In conclusion, Causal models are a fundamental tool in statistical analysis used to understand and explain the relationships between variables in a system. These models rely on statistical techniques, such as regression analysis, structural equation modelling, and instrumental variable analysis, to identify and estimate the strength and direction of causal relationships between variables. Causal models are widely used in a variety of fields and are a crucial tool for understanding the underlying mechanisms that produce observed relationships between variables.

Articles from Students

❖ **References:**

1. https://youtu.be/mbt6W5E1m9Y

2. https://plato.stanford.edu/entries/causal-models/

3. https://en.m.wikipedia.org/wiki/Causal_model

4. https://towardsdatascience.com/causal-models-for-regression-96270bf464e0

5.https://towardsdatascience.com/apply-instrumental-variables-method-in-causal-analysis- 9fd55e39da7a

6. https://en.m.wikipedia.org/wiki/Rubin_causal_model

7. https://en.m.wikipedia.org/wiki/Structural_equation_modeling

Data-Driven Decision-Making in Criminal Justice

Arka Roy, 2nd Year
Bideepta Saha, 2nd Year

INTRODUCTION:

Effective decision-making is critical to the success of any organisation or business.

According to Nobel laureate **Professor Daniel Kahneman,** a person follows and makes decisions in two ways.

- Intuitive Method
- Logical, Evidence-based Method.

The first method is more convenient and often faster but has a higher risk of error. The second method, on the other hand, is more time-consuming but generally is more reliable because it relies on reasoning and is evidence-based.

Statistical approaches allow organisations to collect data and analyse patterns and other trends related to research. Predictive Analytics uses current and historical data and eventually applies statistical techniques coupled with AI to predict future trends and draw conclusions about the anticipated outcomes.

Why is Data-driven Decision Making necessary in Criminal Justice?

While digging deep into criminal justice, we typically come across three main questions:

1. Whom are we arresting?
2. Whom are we charging?
3. Who are we putting in the nation's jail?

The big problem with answering these questions is that the decisions are often made across the "police → prosecutor → court → prison" system. Based on intuition and experience, this procedure has its advantages and disadvantages. Subjective assessments often lead to wrong decisions. Here comes the need to introduce data and rigorous statistical analysis to make decisions in the field of criminal justice.

Data-driven Criminal Justice is revolutionising many areas. These include:

- **Response Plan**: By analysing law enforcement decisions and outcomes in responding to thousands of crimes, the experts can determine which response methods are appropriate in different types of situations, so that the judges can make more rational decisions when responding to acts about the crime being committed.
- **Crime Prevention**: By combining data on crime with data reflecting skiving rates, rate of unemployment, instances of vandalism, and more; law enforcement can see both important and finer-grained correlations that affect crime. Once these segments of information are put together, analysts can use them to predict when and where different types of crime are most likely to occur. For instance, in a pilot program in Manchester, the police used advanced data analytics to apply preventative measures, resulting in a reduction of 12% in robberies, 21% in burglaries and 32% in thefts in motor vehicles.

- **Risk Assessment**: In Texas, Leslie Chew was arrested for stealing four blankets on a cold winter night and kept in jail for $3500 bail which he could not afford to pay and was retained in jail for 8 months at a cost of taxpayers of $9000. This person possessed a low risk to public safety but was still retained and on the other hand, criminals possessing a higher risk to public safety are often released because of the way decisions are taken. Facilitating data-driven decision-making not only keeps high-risk criminals in jail and citizens safe but also ensures that the taxes citizens pay are used correctly.

- **Public Safety Assessment(PSA):**

The pre-trial phase of the criminal justice system is intended to protect public safety and enable defendants to appear in court while ensuring that the

constitutional rights of citizens are upheld. However, research shows that low-risk and non-violent offenders who cannot post bail are often imprisoned, while high-risk offenders are released from prison in many circumstances. This system causes serious harm to many innocent defendants and is a threat to our society. Hence, several jurisdictions are reforming their pre-trial systems to change the way they used to make decisions relating to pretrial release and detentions.

They are shifting from decision-making based primarily on a defendant's charge to decision-making that considers the level of risk that the offender may possess. The **Public Safety Assessment, or PSA**, provides reliable, evidence-based information that helps judges to determine if the defendant should be released before the next trial. The PSA tool uses information related to the defendant's age, criminal record, and current convictions to determine whether a defendant is likely to commit a new crime or fail to appear for their court hearing if released before his next trial. This risk-based approach helps in allowing a relatively small number of guilty suspects to remain in prison and low-risk offenders to be released and hence returned to society to await trial.

How does the PSA work?

PSA uses those factors that are the strongest predictors of whether a defendant will commit a new crime, commit a violent crime, or fail to return to court if released before trial. The factors are:

- whether the current offence is an act of violence
- whether the person had a pending charge during the current offence
- whether the person has a prior conviction for a misdemeanour
- whether the person has a prior conviction of a felony
- whether the person has prior convictions for committing violent crimes
- age of the person at the time of the arrest.
- number of times the person failed to appear at a pre-trial hearing in the last two years
- whether the person failed to appear at a pre-trial hearing more than two years ago
- whether the person has previously been sentenced to imprisonment.

With the help of this information, the PSA produces **two risk scores**: one used for predicting the chances that an individual will commit a new crime if released before trial, and another, for predicting the chances that he will fail to return to the court for a future hearing. The tool will also calculate an elevated risk of the defendant committing a violent crime. The PSA risk score is expressed on a scale of 1 to 6, with higher scores indicating higher levels of risk. This neutral and authoritative method helps judges understand the risk that a defendant possesses if he is released.

Pre-Trial Risk Assessment (PTRA):

In the Federal System, when a person is arrested and charged with a crime, court officials must decide whether to release the accused or imprison him until the case is resolved. The **Pretrial Risk Assessment (PTRA) Instrument** was introduced in the United States to assess a defendant's likelihood of pre-trial misconduct. This includes failure to appear in court, pre-trial withdrawal, or re-arrest for new criminal offences.

Development: The current study began with all defendants (n=565,178) who entered the Federal System between FY2001 and FY2007. After certain refinements and missing data, the final study was conducted on 185,827 and 215,338 defendants.

This study included **two dependent measures** (outcomes). The first measure, FTA/NCA, is an indicator of failure if the defendant either failed to appear in court or was charged with a new criminal arrest while on pretrial release. The second dependent measure, FTA/NCA/TV, is also an indicator of failure if the defendant either failed to appear; was arrested for a new criminal charge while on pretrial release, or had his/her bond revoked due to technical violations.

A split sample process for construction and validation was applied. First, potential risk factors were identified based on results from previous studies and supplementary logistic regression analysis using split-sample processes and bootstrapping. Once a set of risk factors was identified, points were assigned to those risk factors and a risk score was calculated. The relationship between this score and outcomes of interest was assessed. Then apply the risk calculation to the remaining 50% of the sample to determine whether the risk instrument holds across the two halves of the larger sample.

Table 1.
Risk factors and failure rates by sample

Variable	FTA/NCA			FTA/NCA/TV		
	A	C	V	A	C	V
Number of felony convictions						
0-None	6	6	6	10	10	10
1-One to four	11	12	11	19	19	19
2-Five or more	16	15	16	26	26	26
Prior FTAs						
0-None	6	6	6	11	11	11
1-One to four	12	12	11	22	22	21
2-Five or more	15	15	14	26	26	26
Pending cases						
0-No	6	6	6	11	11	11
1-Yes	12	12	12	22	22	22
Current offense type						
0-Theft/fraud, violent, other	4	5	4	8	8	8
1-Drug, firearms, immigration	10	10	10	18	18	18
Offense class						
0-Misdemeanor	4	4	5	6	6	6
1-Felony	8	8	7	14	14	14
Age at interview						
0-47 and older	4	3	4	6	6	6
1-27 to 46	7	7	7	13	13	13
2-26 or younger	9	9	9	17	17	16
Highest education						
0-College degree	3	3	3	5	5	5
1-High school degree, vocational, some college	6	6	6	11	11	11
2-Less than high school or GED	10	10	10	19	19	19
Employment status						
0-Employed	6	6	6	10	10	10
1-Unemployed	9	9	9	17	17	17
Residence						
0-Own/purchasing	4	4	4	7	7	7
1-Rent, other, no place to live	8	8	8	15	15	15
Current drug problems						
0-No	5	5	5	7	7	7
1-Yes	10	10	10	19	19	19

After conducting a series of bivariate analyses and multivariate logistic regression models, several factors relevant to predicting pre-trial outcomes and scoring schemes for each of these factors were identified. As indicated in **Table 1**, most factors are related to criminal history and the specifics of the current offence.

Most items are scored in a 0 and 1 format. Items with multiple point values also use a simple weighting process(0, 1, or 2 points).

Table 1 reports the failure rates based on the two outcomes (measures) for all defendants (column labelled A), the construction sample (column labelled C), and the validation sample (column labelled V). The relationship between the design pattern and the validation pattern remains largely unchanged. All relationships are statistically significant at the $p < .001$ level.

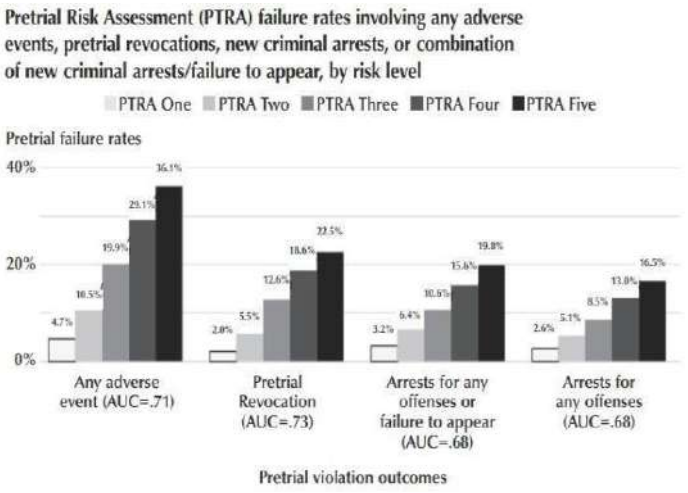
As shown in **Table 3,** a full 30% of the accused fall into the lowest risk category (Category I).

Table 3.
Failure Rates, Odds of Failure, and PSO Release Recommendations.

Risk Category	N	%	FTA/ NCA*	Odds of Success	FTA/NCA/TV*	Odds of Success	PSO Release Recommendation
Category I (0-4)	55,243	30	2%	49:1	3%	32:1	86%
Category II (5-6)	53,193	29	6%	16:1	10%	9:1	60%
Category III (7-8)	47,915	26	10%	9:1	19%	4:1	41%
Category IV (9-10)	20,833	11	15%	6:1	29%	2:1	28%
Category V (11+)	4,555	3	20%	4:1	35%	2:1	13%
* P < .001							

Almost similar percentages fall into categories II and III (29 per cent and 26 per cent, respectively). A very small percentage of defendants belong to categories IV and V. It is to be noted that both measures of failure rates increase from one category to the next. The failure rates for category V are 10 times the failure rates for category I defendants when considering FTA/NCA. The FTA/NCA/TV measure also notes a similar trend.

Results: The predictive efficiency of PTRA was tested on a sample of 85,369 released defendants. The figure beside represents the percentage of people committing pre-trial violations, re-arrest, adverse events, etc. Based on crime type, samples have been distributed in 4 PTRA classes. The **AUC** ranges from 0.0 to 1.0, with 0.5 representing the value associated with the chance prediction.



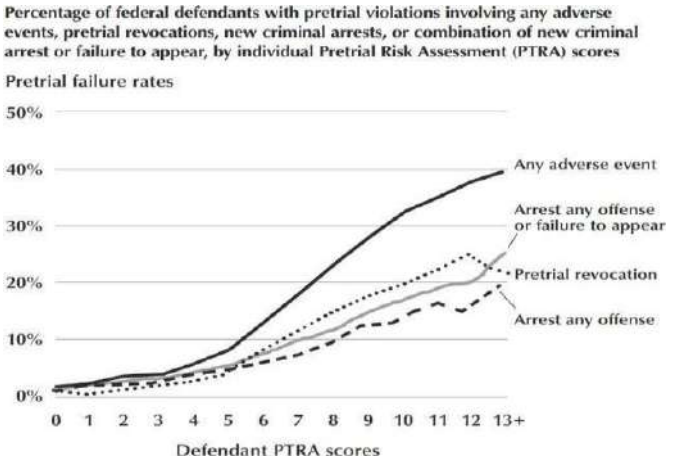
Minimum AUC-ROC scores of 0.56, 0.64, and 0.71 correspond to "small," "medium," and "large" effects, respectively.

Higher AUC implies a better performance of the model in distinguishing between the positive and negative classes.

This shows that PTRA effectively predicts pre-trial violations irrespective of whether the outcome of interest involves revocation from pre-trial release, re-arrest for any felony or misdemeanour offences, or a combination of these outcomes. This histogram (any adverse event as) shows that while on pretrial release, a tendency to commit any adverse event increased in the following

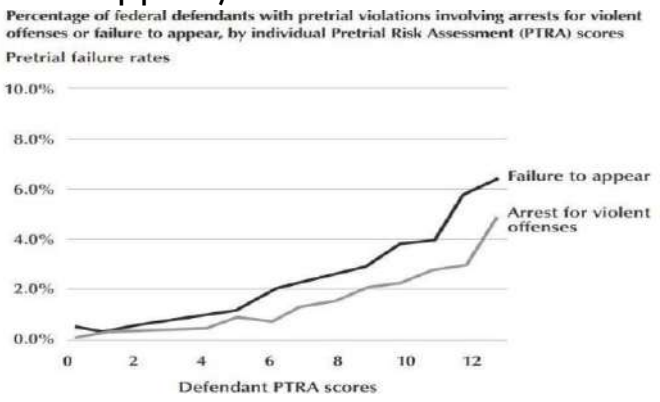
incremental fashion by PTRA risk category: 4.7 per cent (PTRA ones), 10.5 per cent (PTRA twos),

19.9 percent (PTRA threes), 29.1 per cent (PTRA fours), and 36.1 per cent (PTRA fives). This was expected as, the higher the classifications, the higher the FTA rate.



This figure represents that as the defendant's PTRA scores have been increasing, violations in pre-trials are also increasing which is to be expected. We can see that pretrial revocation has been increasing up to 22 per cent, corresponding to a PTRA score of 12, decreasing to 20 per cent. Defendants with PTRA scores of 13 or above were recorded into PTRA 13s, as there were relatively few defendants with these very high PTRA scores (n= 19) to produce statistically reliable estimates.

This figure illustrates that FTA (failure-to-appear) and arrests for violent crime offences surge as PTRA scores rise. A one-point increase in PTRA points results in incrementation in violations. Slight exceptions can be seen - when the PTRA score increases from 0 to 1 there is a decrease in FTA before increasing again. The violent re-arrest rates are essentially the same for defendants with PTRA scores of 1/2 and 5/6.



Public Safety Assessment Dashboard:

Developed by **Arnold Ventures,** the PSA Dashboard is a tool that uses nine risk factors to generate a score and predict three outcomes: failure to appear pre-trial, new criminal arrests, and new arrests for violent crimes. Decision

makers use the PSA score along with the release conditions matrix to inform pretrial release conditions.

APSA uses a variety of factors to predict 3 outcomes: FTA (failure to appear), NCA (new criminal arrest), and NVCA (new violent criminal arrest).

PSA uses 4 factors to calculate the FTA score:

- Pending charge at the time of the arrest
- Prior convictions
- Prior failure to appear in the past 2 years
- Prior failure to appear older than 2 years.

The table below shows how PSA assigns points to FTA factors:

Failure to Appear: Points		
PSA FACTOR	**RESPONSE**	**POINTS**
Pending charge at the time of the arrest	No	0
	Yes	1
Prior conviction (misdemeanor or felony)	No	0
	Yes	1
Prior failure to appear in the past 2 years	No	0
	Yes, just 1	2
	Yes, 2 or more	4
Prior failure to appear older than 2 years	No	0
	Yes	1

Each factor is weighted according to its strength of relationship with the specific outcome. Similarly, the following tables show how PSA assigns factor weights to NCA and NVCA.

New Criminal Arrest: Points

PSA FACTOR	RESPONSE	POINTS
Age at current arrest	23 or older	0
	22 or younger	2
Pending charge at the time of the arrest	No	0
	Yes	3
Prior misdemeanor conviction	No	0
	Yes	1
Prior felony conviction	No	0
	Yes	1
Prior violent conviction	No	0
	Yes, 1 or 2	1
	Yes, 3 or more	2
Prior failure to appear in the past 2 years	0	0
	Yes, just 1	1
	Yes, 2 or more	2
Prior sentence to incarceration	No	0
	Yes	2

New Violent Criminal Arrest: Points

PSA FACTOR	RESPONSE	POINTS
Current violent offense	No	0
	Yes	2
Current violent offense and 20 years old or younger	No	0
	Yes	1
Pending charge at the time of arrest	No	0
	Yes	1
Prior conviction (misdemeanor or felony)	No	0
	Yes	1
Prior violent conviction	No	0
	Yes, 1 or 2	1
	Yes, 3 or more	2

CONCLUSION:

PSA is highly accurate in predicting pre-study outcomes. However, PSA can be dramatically improved if better-quality data can be integrated. Data-driven Decision Making is the future of the Criminal Justice system.

REFERENCES:

1. https://www.uscourts.gov/sites/default/files/82_2_3_0.pdf
2. https://www.uscourts.gov/sites/default/files/73_2_3_0.pdf
3. https://www.uscourts.gov/sites/default/files/73_2_1_0.pdf
4. https://www.tableau.com/learn/articles/data-driven-decision-making
5. https://bja.ojp.gov/doc/policymakers-use-data-inform-cj-decisions.pdf
6. https://blog.ipleaders.in/data-driven-policing-changed-law-enforcement/
7. https://www.snowflake.com/blog/three-ways-data-driven-insights-improve-public-safety/
8. https://statetechmagazine.com/article/2021/09/public-safety-agencies- seek-tools-improve-data-driven-decision-making

EFFECT OF WORDING BIAS IN SURVEY

Bishwayan Ghosh, 2nd Year
Dyuti Manik, 2nd Year

❖ Wording Bias:

Wording bias is a type of bias in which the respondent is influenced by the wording of the question. For example, a question asking people's views on the statement: "Welfare helps people to get back on their feet" will get an extremely different response than the one asking their views on the statement: "Welfare pays people who don't work". The former will get a positive response, whereas, the latter is expected to get a negative one. This bias can also occur as a result of the questions that are being framed in a way, in which it is difficult to understand for the respondents.

In this article, our goal is to demonstrate the effect of wording bias in a survey. In order to do so, we have conducted a survey, where we chose questions that are susceptible to wording bias. The details of the survey and data collection are given in the next section.

❖ Data Collection:

We chose four topics and decided to see what are the opinions that people have on them. For each topic, we chose a question and have framed it in two different ways. This also resulted in two different questions, which had the same meaning but different wordings. Each question had possible answers as "YES" or "NO". Clearly, if wording bias does not have an effect, both the questions should have the same proportion of "YES" and "NO" answers. The topics chosen were:

❑ Topic 1:

Whether people think that meditation helps in increasing our concentration.

Question 1: It is believed that meditation increases concentration on work and academics. Do you think so?

Question 2: Researchers from a well-known university claimed that meditation does NOT affect concentration on work and academics. Keeping

this in mind, do you think that meditation increases concentration on work and academics?

- **Topic 2:**

  Whether people think that the violence shown in movies and video games have a detrimental effect on the mindset of the youth.

  Question 1: Many studies, including one from Dr. Gentile at Iowa State University, have proven that playing and watching violent video games cum movies can give rise to aggression in children. Keeping this in mind, do you think children playing violent video games tend to have aggressive behaviour when they grow up?

  Question 2: Do you think children playing and watching violent video games and movies tend to have aggressive behaviour when they grow up?

- **Topic 3:**

  Whether people think that classical and pop music have the same effect on logic.

  Question 1: It is believed that listening to music increases our IQ and logic. But classical and pop music are quite different. Do you think classical music and pop music have the same effect on our logical abilities?

  Question 2: A social experiment carried out by National Geographic Channel revealed that both classical and pop music have the same effect on our logical abilities. Do you think so?

- **Topic 4**:

  Whether diet is more important than exercise for reducing weight.

  Question 1: Do you think cutting down your calorie intake WITHOUT much exercising will help you to lose weight?

  Question 2: Several studies have shown that reducing your calorie intake is more efficient in reducing weight as compared to that caused by exercising. Do you agree with this opinion?

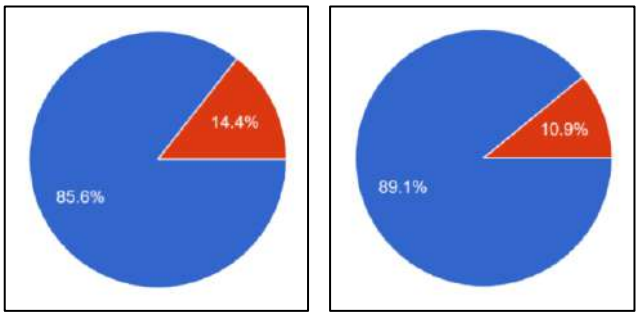
Clearly, for each topic, both the questions have effectively the same meaning. We constructed two surveys, each containing one question from each topic. Survey A consisted of Question 1 while Survey B consisted of Question 2 from all the topics.

We conducted the survey using google forms. Each respondent was given one survey. In order to randomise the two surveys among the respondents, we created two different "sections" in our google form. At the start of the form, we kept a question with two similar looking images as options. Each option led to one of the two sections corresponding to the two surveys. We further used the "shuffle option order" in order to minimise any other kind of bias.

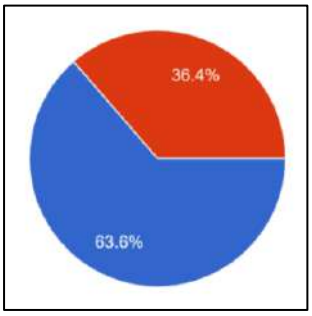
❖ **Observations:**

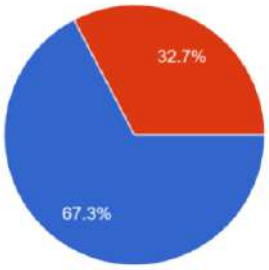
A total of 214 responses were received. Out of them, Survey A was filled by 110 respondents and the remaining 104 filled Survey B.

For topic 1, the percentage of 'Yes' (blue) and 'No' (red) answers for questions 1 and 2 are respectively:

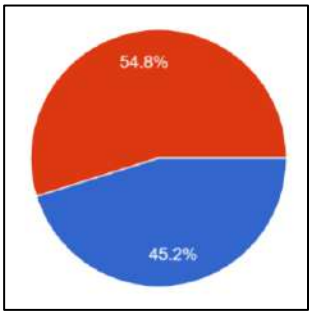


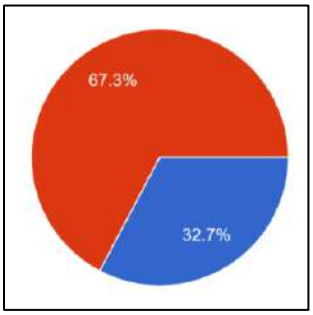
For topic 2, the percentage of 'Yes' (blue) and 'No' (red) answers for questions 1 and 2 are respectively:



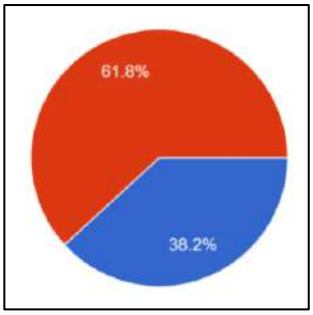


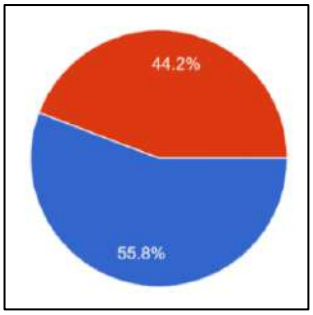
For topic 3, the percentage of 'Yes' (blue) and 'No' (red) answers for questions 1 and 2 are respectively:





For topic 4, the percentage of 'Yes' (blue) and 'No' (red) answers for questions 1 and 2 are respectively:





❖ **Statistical Analysis:**

While having a glance at the pie charts, it seems as if the proportion of "YES" and "NO" as answers for the two questions in topics 1 and 2 are almost the same, while the proportion is different for topics 3 and 4. Now, we shall prove our claims or 'eye-estimates' with the tool of **hypothesis testing**. Below is the procedure of **testing of two independent binomial proportions** that we have used as the aforementioned proportions follow Binomial distribution.

Let X and Y be two random variables denoting respectively the number of times "YES" has been selected as an answer in question 1 of a topic from a randomly selected sample of 110 responses and in question 2 of a topic from a randomly selected sample of 104 responses.

Clearly, $X \sim \text{Binomial}(110, p_1)$ independently of $Y \sim \text{Binomial}(104, p_2)$, where, p_1 and p_2 are the proportion of "YES" answers for question 1 and question 2 respectively.

Let us define the **test statistic** as, $\mathbf{Z = X+Y}$
Let x_0, y_0 and z_0 be the observed values of X, Y and Z =(X+Y) respectively.
Null Hypothesis H_0 : $p_1 = p_2 = p$ (say), where p is known.
Null Distribution: Under H_0, $Z \sim \text{Binomial}(214,p)$
For $X = x_0$, $Y = y_0$ and $Z = z_0$, the distribution of X | X+Y (or equivalently, X | Z) is

$$P(X = x_0 \mid X + Y = z_0) = \frac{P(X = x_0, X + Y = z_0)}{P(X+Y=z_0)} = \frac{\binom{n_1}{x_0}\binom{n_2}{z_0-x_0}}{\binom{n_1+n_2}{z_0}}$$

$$Thus, X \mid X + Y \sim Hypergeometric(n_1 + n_2, x_0, \frac{n_1}{n_1 + n_2})$$

Here, sample sizes are $n_1 = \mathbf{110}$ (Question 1), $n_2 = \mathbf{104}$ (Question 2), for each topic.

We are going to use the **p-value approach** to test for equality of proportion of "YES" in the two questions, at 5% level of significance **($\alpha = 0.05$)**. We reject the null hypothesis if the p-value (p*) is less than the level of significance and we fail to reject the null hypothesis if the p-value is greater than level of significance. The calculations and conclusion had been tabulated below in the following table: In the table, x_0 and y_0 denote the number of "YES" given as answer for Question 1 (out of 110) and Question 2 (out of 104) respectively.

In each case, the **Null Hypothesis** is the same, as stated above, i.e., **H_0: $p_1 = p_2$.**

Topic No.	x_0	y_0	Alternate Hypothesis	p-value	Decision	Conclusion
1	98	89	$p_1 < p_2$	0.830	Fail to reject H_0	the proportion of "YES" can be considered to be the same for questions 1 and 2 of topic 1.
			$p_1 > p_2$	0.285	Fail to reject H_0	
2	70	70	$p_1 < p_2$	0.337	Fail to reject H_0	the proportion of "YES" can be considered to be the same for questions 1 and 2 for topic 2.
			$p_1 > p_2$	0.760	Fail to reject H_0	
3	36	47	$p_1 < p_2$	0.042	Reject H_0	the proportion of "YES" for question 1 is significantly less than that of question 2 for topic 3.
4	42	58	$p_1 < p_2$	0.007	Reject H_0	the proportion of "YES" for question 1 is significantly less than that of question 2 for topic 4.

From the above calculations, we see that wording bias has a significant effect on the answer of respondents for topics 3 and 4, whereas it did not have such impact in topics 1 and 2. Thus, it is surely evident that a change in the wording of a question can potentially change the views of the respondents.

❖ **Conclusion:**

Wording bias, which makes the collected data less valuable, is not acceptable in a survey. The questions itself are phrased in a manner so as to direct respondents to a desired answer. It can be reduced by framing questions using simple and lucid language. They should be neutral and straightforward. The respondent should have a basic knowledge and

understanding of the topics that are being surveyed. Wording bias can also be reduced if the respondents have a firm opinion about the topic and do not get influenced by some extra information given in the question. For instance, most people firmly believe that meditation increases our concentration. Hence, even after trying to influence their opinion by giving some extra meditation, their answers remained the same. Also, in general, people in our society firmly believe that violent video games tend to result in aggressive behaviour among children. Hence, no wording bias was observed in topics 1 and 2.

Usually, a question is framed in a single way in a survey unlike ours, where each was framed in two different ways. We should prefer the one with least amount of additional information, as it may influence the respondents' answer resulting in wording bias.

If we are to choose a single question from each topic, we should choose, Question 1 for Topic 1, Question 2 for Topic 2, Question 1 for Topic 3, and Question 1 for Topic 4 as they do not try to influence people's opinion by quoting the results of other studies.

❖ **References:**

1) https://www.nextiva.com/blog/response-bias.html

2) https://en.wikipedia.org/wiki/Response_bias

THE IDEA OF ENTROPY IN STATISTICS AND ITS APPLICATION IN THE DECISION TREE CLASSIFIER MODEL

Spandan Ghosh, 1st Year, M.Sc. in Data Science
Sayan Das, 1st Year, M.Sc. in Data Science

❖ Introduction:

In our day-to-day activities, we often come across some decision-making problems that we have to solve on our own, based on past experience. For example, we may have to make a decision about whether we should carry an umbrella or not before we go out or whether an email is spam or not. In the glossary of Data science, these problems are called **classification problems.** Where we need to classify some objects according to some attribute. In our first example - The object can be a day, and we have to classify whether it is a sunny day or it is a rainy day. And in the second example, we need to classify whether the emails are spam or not.

Now, the decision-making in these problems will be better in accuracy, that is, the decisions will grasp reality better if we allow it to be data-driven. With respect to the first example, suppose we have a data set on several meteorological measurements on a daily basis and the observations of whether the day was a rainy day or a sunny day, then based on this dataset we may be able to predict if today is going to be a sunny day or a rainy day, provided we have the similar metrological measurements at hand for today.

There are several algorithms that are used in an automated system to implement this decision-making feature using previous data. One of those algorithms is a **Decision Tree Classifier** model, a non-linear classifier algorithm used for decision-making purposes. It is based on the ID3 (Iterative Dichotomiser 3) algorithm developed by **Ross Quinlan.** The ID3 algorithm is based on the foundation of **Statistical Entropy.**

In this article, we are going to discuss the ID3 algorithm for building up a Decision Tree with a simple example, for which we are going through the idea of Entropy.

❖ What is Entropy?

The concept of Entropy in Information Theory was first proposed by **Claude Shannon** in his paper **"A Mathematical Theory of Communication"**. It is basically a measure of "uncertainty" or "surprise" in the values of a random variable. Let's revise how the concept of Entropy was developed.

❑ Development of the Idea of Entropy:

To understand the concept of entropy we need to understand "surprise" or "uncertainty". Let us do it with an example.

Consider a random experiment of throwing a die. Suppose we are interested only in the occurrence of a "six". Now, let $p(0 \leq p \leq 1)$ be the probability of getting a "six" in a single throw of the die.

Now, assume p is very low (say $p = 0.01$). That means if we go on throwing the die, we hardly ever come across a "six". Now in this scenario, if by fortune a "six" occurs, then we will be surprised as it is a nearly impossible event under this setup. Also if in this scenario some of the faces that are other than "six" (like "one", "five" etc.) turned up, then we wouldn't have been surprised at all, because when the probability for "six" to turn up is such low then the compliment cases will automatically have a higher probability of occurrence.

From the discussion of the above example, we are now in a state to appreciate that, the amount of surprise that we get from the occurrences of an event is inversely related to the probability of the occurrence of that event. So, we can write,

$$amount\ of\ surprise\ of\ an\ event \propto \frac{1}{probability\ of\ occurrence\ of\ the\ event}$$

But there is one problem with using the inverse of probability to measure the amount of surprise of an event. Suppose in the previous example - say the probability of occurrence of "six" in a single throw of a die, i.e. p, is very high, (say p=1). Now if in a throw of a die, we get a "six", we will not be surprised at all. Meaning that here our amount of surprises is 0. But this fact is not reflected if we calculate surprise using $\frac{1}{P("six")}$. Here as p = 1, we will get surprised $= \frac{1}{1} = 1$. To let go of this confusion we use to take $\log(\frac{1}{p})$ while calculating the amount of surprise of an event.

Now, we are in the stage to get an idea of what entropy is. The entropy of a random experiment is the expected amount of "surprise" or "uncertainty" or "randomness" that is inherent in the possible outcomes of that random experiment.

Consider a random experiment W that has n number of possible outcomes e_1 ,e_2 ,e_3 ,...e_n and the probability of occurrences of the event e_i , is P_i. Then, the entropy of the W is given by,

$$E = \sum_{i=1}^{n} \left(\frac{1}{Pi}\right) P_i$$

$$= -\sum_{i=1}^{n} Pi\ )$$

We can define it for a random variable too. If X be a random variable assuming the values x_1, x_2, x_3,...x_n and $P(X= x_i) = p_i$ then the entropy associated with the random variable X is given by

$$E = -\sum_{i=1}^{n} pipi\ )$$

❑ **The interpretation of Entropy:**

From the previous discussion, we have seen that entropy measures the surprise or randomness in the system. In our previous example, if the probability of occurrence of "six" is very high or very low, then in those cases only one type of outcome will occur mostly - either it will be "six" (when the probability of "six" is high) in most of the cases or the outcome will be "not six" (when the probability of "six" is low).

Now, consider 100 throws of the die, Let's say p = 0.97. Then those hundred outcomes will be more or less homogenous (most of them are "six"). Ideally if p = 1 then all the 100 outcomes will be "six". In that case, we will say the outcomes are pure. Also when ideally p = 0 then all the outcomes will be "not six". Then also we will say that the outcomes are pure. As p moves between 0 and 1, the heterogeneity in the outcomes varies and so does the impurity in the outcomes.

So, the entropy of a random experiment is measuring the impurity in the outcomes of a random experiment. The higher the value of the entropy is, the lesser the impurity in the outcome of the random experiment.

The main application of entropy in the implementation of the decision tree is associated with the ability of entropy to measure the impurity in the outcomes of the random experiment. We will use it to calculate the impurity of the leaf node of a decision tree. But what is a leaf node? Let's first study the decision tree.

❖ **Introduction to Decision Tree:**

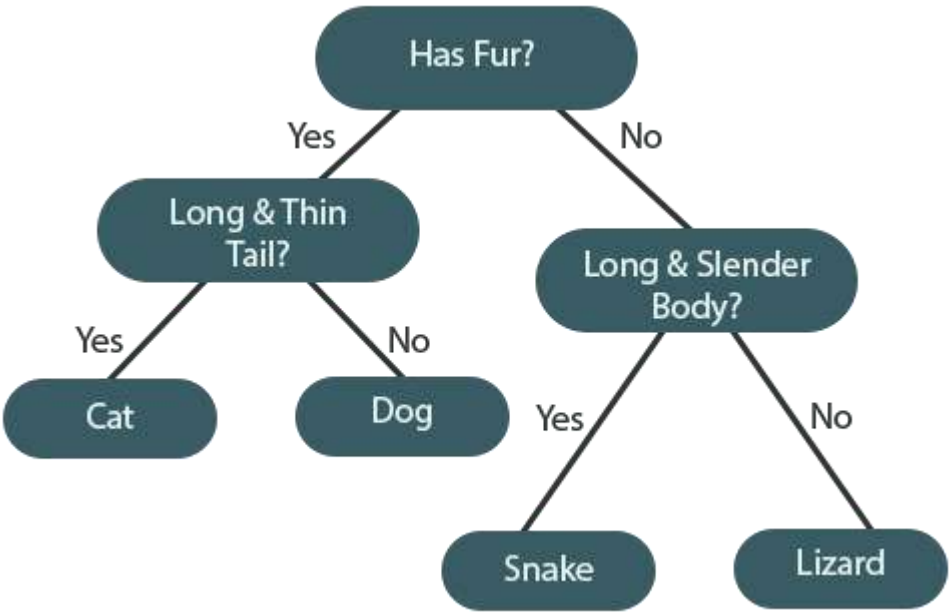
Decision trees are a form of machine learning technique that may be applied to both classification and regression applications. A decision tree learns a hierarchy of judgments based on a set of features in order to build a model that can make predictions based on those features.

A decision tree works by creating a tree-like model of decisions based on the features of the data. At the top of the tree is a root node that represents the whole dataset. The root node is then split into two or more branches, each representing a different decision or feature. The branches are further split into smaller branches until they reach a leaf node, which represents a prediction or classification.

Decision trees are a simple and effective way to make predictions based on a set of features, and they are widely used in many different fields, including finance, healthcare, and marketing.

❖ **Illustration with an example:**

Here's an example of a decision tree for predicting whether an animal is a cat or a dog:



In this example, the root node is the decision to determine whether the animal has fur. If the answer is "yes," then the algorithm will consider the length and shape of the animal's tail to make a prediction. If the animal has a long and

thin tail, the algorithm will predict that it is a cat. If the animal does not have a long and thin tail, the algorithm will predict that it is a dog.

If the answer to the root node is "no," then the algorithm will consider the length and shape of the animal's body to make a prediction. If the animal has a long and slender body, the algorithm will predict that it is a snake. If the animal does not have a long and slender body, the algorithm will predict that it is a lizard.

This is just one example of how a decision tree could be used to make a prediction. Decision trees can be used for a variety of purposes, including classification, regression, and decision-making.

❖ **Use of Entropy in Decision Tree:**

In decision tree learning, entropy is used as a measure of the impurity of an attribute. The entropy of an attribute is calculated based on the frequency of each unique value (or class) in the attribute. If the attribute is pure, meaning that all of the values in the attribute belong to the same class, then the entropy is zero. On the other hand, if the attribute is completely mixed, meaning that there are equal numbers of values belonging to different classes, then the entropy is at its maximum.

The goal of decision tree learning is to create a tree with nodes that are as pure as possible, meaning that the values within each node all belong to the same class. To do this, the decision tree algorithm will try to split the data at each node in a way that maximizes the purity of the resulting nodes. This is where entropy comes in: the decision tree algorithm will use entropy to determine the best split at each node.

Let's discuss this with the Help of an example, consider the following data of n = 10 cases of Human Survival with respect to the temperature and presence of Water and Flora & Fauna.

Temperature	Water	Flora & Fauna	Human Survival
Hot	Present	Present	No
Hot	Not Present	Present	Yes
Hot	Present	Present	Yes
Cool	Not Present	Not Present	No
Cool	Present	Not Present	No
Cool	Not Present	Not Present	No
Cool	Present	Present	No
Hot	Not Present	Present	Yes
Hot	Present	Not Present	Yes
Cool	Not Present	Present	No

❖ **Procedure of Building a Decision Tree:**

In the above example, the variable of interest is Human Survival. We have to classify a newly given place and whether it is possible for humans to survive based on the temperature of the place and the presence of the water and Flora and Fauna in the place. Here there are 3 predictors that are respectively temperature, presence of water, and presence of flora and fauna. To build a decision tree, on this data, we first have to decide which of these three predictors is Most informative. We will find this out by following the procedures shown below.

In the first step, we are going to calculate the Entropy of the column named 'Human Survival' i.e. $E_{HumanSurival}$ given by the formula -

$$E_{Human\ Survival} = -P("Yes").log_2(P("Yes")) - P("No").log_2(P("No"))$$

Note that this quantity $E_{Human\ Survival}$ gives the idea of randomness within the values of the Human Survival Column.

In this example $E_{Human\ Survival} = 0.971$

Now in the next step, for each predictor variable, we are going to calculate the Entropy of the Human Survival for each of the labels of the predictor. For example, consider the predictor variable temperature. The temperature variable has two labels, 'Hot' and 'Cool'. So we split the dataset into two parts, one where the temperature is 'Hot' and the other where the temperature is 'Cool'. Now for each of the two data sets, we calculate the Entropy of Human Survival.

Specifically for the Label 'Hot' under the Temperature variable, we consider only those records for whom Temperature is labelled as 'Hot', then we calculate Entropy $E_{Human\ Survival\ |Hot}$ as -

$$E_{Human\ Survival\ /\ Hot} = -P("Yes"|'Hot').log_2(P("Yes"|'Hot')) - P("No"|'Hot').log_2(P("No"|'Hot'))$$

This quantity gives the idea of the randomness of the Human Survival columns of the places where the temperature is labelled as 'Hot'.

Similarly, we can calculate $E_{Human\ Survival\ /\ Cool}$ as -

$$E_{HumanSurvival/Cool} = -P("Yes"|'Cool').log_2(P("Yes"|'Cool')) - P("No"|'Cool').log_2(P("No"|'Cool'))$$

This quantity gives the idea of the randomness of the Human Survival columns of the places where the temperature is labelled as 'Cool'.

Now, to determine how much information, the predictor variable Temperature holds to determine the response variable Human Survival, we can determine the Gain(temperature) as

$$\text{Gain(Temperature)} = E_{\text{Human Survival}} - P(\text{'Hot'}).E_{\text{Human Survival/Hot}} - P(\text{'Cool'}).E_{\text{Human Survival/Cool}}$$

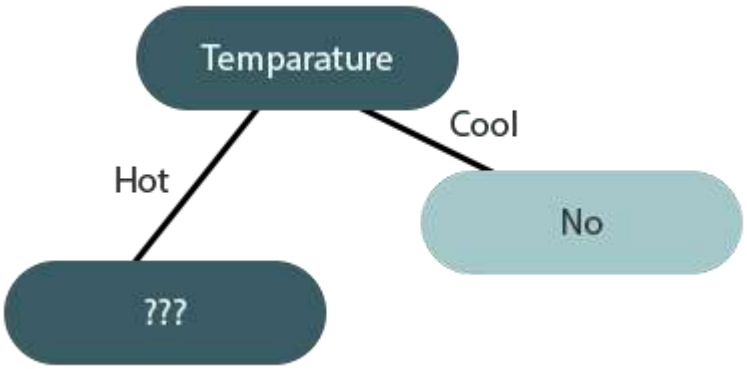
Since in this quantity, the average entropy of Human Survival due to temperature is subtracted from the unconditional entropy of the same, the quantity gives us the amount of information gained for treating the temperature variable as a predictor variable. The more the value of Gain(temperature) the more informative the Temperature variable is in guessing Human Survival.

Now our job is to get this Gain() value for all the predictor variables and set that variable as the root node of the decision tree which has the highest information gain.

Here in the example, we got Gain (Temperature) = 0.610, Gain (presence of Water) = 0, and Gain (presence of Flora & Fauna) = 0.046. Since the Gain(temperature) has the highest information gain we will select the temperature as the root node of the decision tree.

Once we select the root node as Temperature, it will have two branches (One for each), since it can assume only two values 'Hot' and 'Cool'. Now observe that in the data table, for the records where the Temperature value is 'Cool' the corresponding value of the Human Survival is always a 'No' So, the branch corresponding to 'Cool' in the root node will result in a leaf node - 'No'. We will call this type of branch a 'Pure Branch'. It always results in a leaf node. On the contrary, the records that hold the

Temperature value as 'Hot' contain both 'Yes' and 'No' in the Human Survival Column. This type of branch is called an impure branch. This leads us to some undecided situations. The Decision tree at this current stage will look like the following -

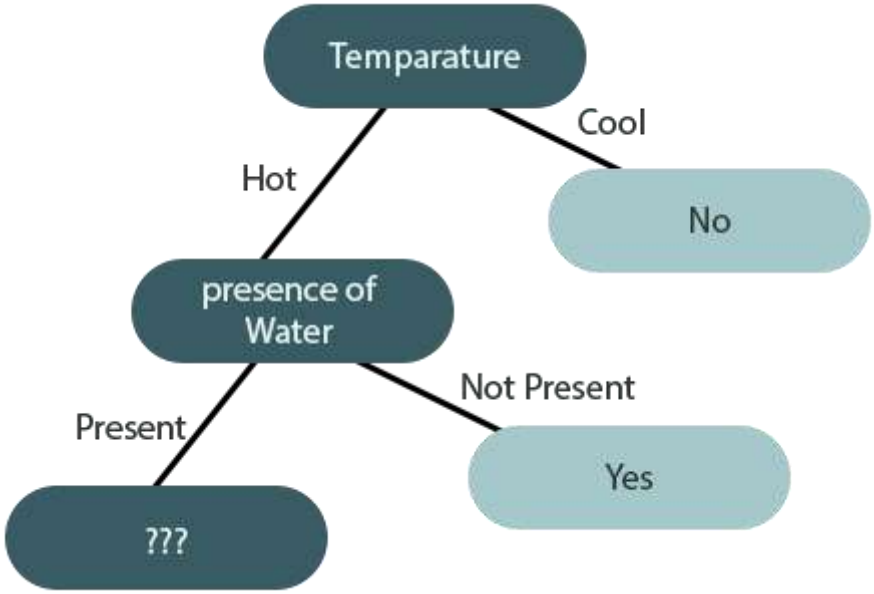


Now, here we can see that the right part of the tree is complete. Now we are stuck in the left part where all the Temperature observation is 'Hot'. We now have to decide on the basis of the other features like the presence of Water and the presence of Flora & Fauna, which were previously ignored as they contain less information compared to temperature.

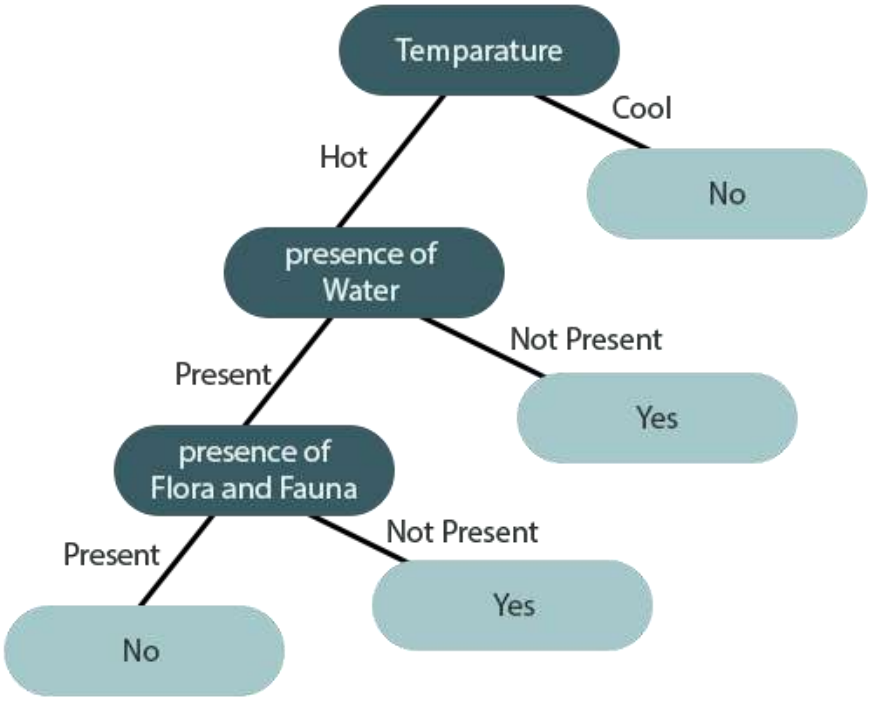
What we can do is we can consider those records that have Temperature = 'Hot' and calculate the Gain of the presence of Water and presence of Flora & Fauna predictors by the method of entropy (done previously), to see which one among them holds the most of information about deciding Human Survival.

So here, proceeding in a similar fashion as before, for the data where the Temperature = 'Hot' we get Gain (presence of Water | Temperature = 'Hot') = 0.171 and Gain (presence of Flora and Fauna | Temperature = 'Hot') = 0.171. Since under this situation, the presence of Water has more information gain, we can add the presence of Water in the left branch of the previous tree. Since the presence of water has only two types of values 'Present' or 'Not Present', this node will also have two branches. The branch where the presence of Water = 'Present' (also Temperature = 'Hot' previously) has 2 out of 3 'Yes' as Human Survival value and 1 out of 3 'No' resulting in an impure branch but all the records that have the presence of Water = 'Not

Present' (also Temperature = 'Hot' previously) result in 'Yes' as the value of Human Survival resulting a leaf node. Then in this stage of modification, the decision tree becomes -



Now, we are left with only one predictor to decide with this undecided situation and that is the presence of Flora and Fauna. We will choose this variable as a node. Similar to previously, here also there will be two different branches. The branch with 'Not Present' (presence of Water = 'Present' and Temperature = 'Hot' as previously) has only one record with 'Yes' as the Human Survival value. This concludes with a leaf node. On the other hand, the branch with 'Present' (presence of Water = 'Present' and Temperature = 'Hot' as previously) has 1 out of 2 saying Human Survival = 'Yes' and 1 out of 2 saying Human Survival = 'No'. This again leaves us in an undecided state, but here the case is different. Here we are out of predictor variables. That means we have no other predictor variable for the further split. So, here is the presence of Flora and Fauna = 'Not Present' then we will predict Human Survival = 'No' since it is the most probable in the whole dataset. So, our final decision tree is like this -



Now the decision tree is complete. It can now be used for prediction purposes.

❖ **Prediction using the Decision Tree Classifier Model:**

A decision tree classifier can be used to make predictions by first dividing the data into a training set and a test set. The training set is used for building the model, while the test set is used for evaluating the model's performance. After choosing the features to be used as inputs for the model, the decision tree is constructed by starting with the root node and branching out based on the values of the selected features. Predictions on new data can then be made by following the branches of the tree corresponding to the values of the input features until a leaf node is reached. The label of this leaf node is the prediction made by the model.

Using a decision tree classifier model, we can predict the possibility of human existence on a given planet by considering the following factors:

- **Temperature**: If the temperature is cool, human existence is not possible on that planet. If the temperature is hot, we move on to the next factor.

- **Water**: If the planet has water, we move on to the next factor. If it does not have water, the existence of humans is not possible.

- **Flora and fauna**: If flora and fauna are present, human life is possible on the planet. If they are not present, the existence of humans is not possible.

Overall, the decision tree classifier model uses these three factors to predict the possibility of human existence on a given planet.

❖ **Conclusion:**

A decision tree is a very powerful tool for the purpose of prediction. It is a Non-linear classification model, whose main foundation is Shanon's Entropy. It can give an excellent prediction result when it comes to the classification problem. Sometimes, one uses more than one decision tree, to classify problems which gives rise to a new model called Random Forest.

Python Package named 'sklearn' (Scikit learn) provides us API called DecisionTreeClassifier() (usage: sklearn.tree.DecisionTreeClassifier()) that helps us to easily implement a Decision Tree model on some dataset.

Here, we have only considered the predictor variables to be qualitative variables. If one or more of the predictor variables are quantitative, then those variable values should be treated in a slightly complex and different way, but the idea is the same as in qualitative variables. Also, there is a model called Decision Tree Regressor, here the response variable is a quantitative variable.

❖ **References:**

1. https://www.vtupulse.com/category/machine-learning/
2. https://www.researchgate.net/profile/Mahesh-Huddar

3. https://statquest.org/
4. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
5. https://en.wikipedia.org/wiki/Entropy

LADY TASTING TEA EXPERIMENT

Anushka Bose, 2nd Year

❖ Introduction:

One of the most important topics in Statistics lies in the study of Statistical Inference. Inferential Statistics is the process of estimating our parameter of interest (for example, height of people aged 18 in India) based on a sample of data, under the assumption that our observed data set is sampled from a large population. This can be done mainly, in three ways - Point Estimation, Interval Estimation and Testing of Hypothesis. In Testing of Hypothesis, we accept or reject certain conjectures regarding our parameter of interest on the basis of our sample dataset; all while allowing a certain level of error. This assumption / conjecture is referred to as the "Null Hypothesis". This term was coined by Sir Ronald Aylmer Fisher, one of the founding fathers of modern statistics; in his work, 'The Lady Tasting Tea' experiment reported in his book ***The Design of Experiments***.

The lady in question was Dr. B. Muriel Bristol, an algologist. One afternoon at the research station Rothamsted, when Fisher drew a fresh cup of tea and offered it to the lady, she declined it, stating that she preferred to add the milk to the tea instead of adding tea to milk as Fisher had prepared it. Fisher laughed it off protesting that it made no difference. However, she was adamant in her claim and was determined to defend her tastes. It was then that Fisher decided to test her by performing an experiment which is now known as 'The Lady Tasting Tea' experiment.

❑ Statement of the Experiment:

The first question that comes to mind while preparing for the experiment, is how many cups of tea should be used in the test; whether they should be paired or not and in what order should the cups be presented. Here, the numbering and ordering of cups should be determined in such a way so as to prevent the correct discrimination of the order of pouring in the cups of tea, simply by pure chance.

Fisher first prepared eight cups of tea, half of which, i.e., 4 cups were prepared with milk added first and then tea while the remaining 4 cups were prepared with tea added first which was eventually followed by milk. The cups were then arranged in a 'random order' and hence, presented to

the lady. She had previously been told what the test would consist of, i.e., to taste the eight cups and distinguish, if possible, the 4 cups in which milk was added first and consequently the other 4 will have tea poured in first. The random order in which the cups were placed were not determined arbitrarily by human choice; but by using a game of chance. This could have been done in many possible ways. The lady was first asked to leave the room and then eight identical cards could have been prepared; marked as T1, T2, T3, T4 to identify those cups in which tea was added first and M5, M6, M7, M8 to identify those cups in which milk was added first. The cards were then shuffled thoroughly and placed one by one faced up on the table. We then had a random order of eight cards. A cup was then placed behind each card. If the card had a "T" on it, the cup was prepared with tea added first and if it had an "M", milk was added first. The lady was then invited back to taste and identify accordingly.

- **Interpretation of the Experiment:**

While performing any statistical experiment, one of the first and foremost steps is to discuss all possible results of the experiment and to decide what interpretation can be made out of them. In our case, if the subject can correctly identify the 4 cups in which milk is added first then the remaining 4 would automatically have tea added first. A person can choose four cups out of 8 in 70 ways by using the following formula through Combinatorics:

$$(8\ 4\ )\frac{8!}{4!\,(8!-4!)}=70$$

This result of 70 is useful in our interpretation. A subject can divide all 8 cups of tea correctly into two categories in 1 trial out of 70, or with a frequency which approaches 1 in 70 more and more nearly; the more often the test is performed. This odds can be much higher by enlarging the experiment, however, if the experiment were smaller than our present case, even the greatest possible success will give odds so low that it may be attributed to chance.

- **The Null Hypothesis:**

Fisher coined the term null hypothesis while reporting this experiment and stated that 'the null hypothesis is never proved or established but is possibly disproved in the course of experimentation.' The experiment is only

performed to 'give the facts a chance of disproving the null hypothesis.' Fisher did not talk about alternative hypotheses in his approach. The null hypothesis must be exact and free from vagueness and ambiguity.

In the context of the present experiment, the null hypothesis can be considered to be the fact that 'the judgements given are in no way influenced by the order in which the ingredients have been added' or in simpler terms, the lady has no ability to distinguish the cups of tea against the alternative one being that she has the ability to distinguish the cups. If the null hypothesis is rejected, the alternate hypothesis can be supposed to be true on the basis of the sample data; however, may not be so in reality.

- **Test Statistic:**

A test statistic is a random variable(or, any function of the random variable) whose value is calculated based on our sample data. It is used to determine whether to accept or reject the null hypothesis. In this case, the test statistic can be a ***simple count of the number of successful attempts to select the 4 cups prepared by a specific method***. This can be easily calculated using the method of 'permutations and combinations'.

Let us denote any successful attempt of identifying a cup with correct guess of ingredient added first by '×' and any unsuccessful attempt by an 'o'. Suppose the lady has 0 success, so, the combination of selections will look like: 'oooo', as is indicated by the first row of the table below. 0 success can be chosen out of 4 cups in ${}^4C_0$ ways which equals to 1. Similarly, if the lady successfully chooses 1 cup out of 4, she can do that in ${}^4C_1 = 4$ different ways as has been illustrated in row 2 of the table. Similarly, she can correctly choose 2 cups out of 4 in ${}^4C_2 = 6$ different ways as shown in row 3. Therefore, we can select any two correct cups and the remaining two incorrect cups are $6 \times 6 = 36$ ways. The remaining number of combinations can be calculated in a similar pattern. The frequencies of the possible number of successes are given in the final column of the table.

Tea-Tasting Distribution Assuming the Null Hypothesis		
Success Count	**Combinations of Selection**	**Number of Combinations**
0	oooo	1 x 1 = 1
1	ooo×, oo× o, o× oo, × ooo	4 x 4 = 16
2	oo× ×, o× o×, o× × o, × o× o, × × oo, × oo×	6 x 6 = 36
3	o× × ×, × o× ×, × × o×, × × × o	4 x 4 = 16
4	× × × ×	1 x 1 = 1
Total		70

We can clearly see that the number of successes follows a Hypergeometric Distribution.

Let X be a random variable denoting the number of successes.

Then,

$$X \sim \text{Hypergeometric}(N=8, K=4, n=4);$$

where, N denotes the population size which is the total number of cups i.e., 8 in our case; K denotes the number of success states i.e., 4 cups of each type and n is the number of draws, i.e., 4 cups.

- **Test of Significance:**

It depends on the experimenter to decide upon the smallness of the probability which he would require before he can admit that his observations have yielded a positive result. Usually and conveniently, we take 5 percent as the level of significance such that we reject all those results from our experiment which fail to reach this standard. By using this, we can divide our possible results of the experiment into two classes, one when we accept the null hypothesis and the other when we reject it. However, no such selection can eliminate the whole possibility of coincidence. In accepting 5 percent as the level of significance, we accept the fact that the lady selects all the cups correctly by 'pure chance' in 5 trials out of 100.

Under the null hypothesis, the probability of selecting all four correct cups is 1 out of 70 which is equal to 0.014 which is less than the Size of the test (considering 5 percent as the level of significance). However, if the lady correctly chooses only three cups out of 4 correctly, the probability of such an event will be 16 out of 70 which equals 0.22 which is way over our critical region. Thus, we can reject the null hypothesis only when the lady correctly categorizes all four cups, effectively acknowledging the lady's ability to distinguish the cups at 1.4 % level of significance. Even the rare case of '3 right and 1 wrong' could not be judged significant simply because it is rare, and it can occur simply because of mere chance by our level of significance.

- **Methods of Increasing Sensitiveness:**

One was to increase the sensitivity of the experiment, i.e., to reduce the possibility of success by chance, we can enlarge the experiment. On the other hand, instead of enlarging the experiment we can increase its sensitiveness by reorganising the structure. One way to do so is not by fixing in advance that there should be 4 cups of each kind. However, we can determine by some random process so as to how the subdivision should occur. We might allow a random process such as the flipping of a coin to decide which ingredient is to be added first to the tea for each of the 8 cups. The chance of correctly classifying all 8 cups prepared by this process would be 1 in 28 or 1 in 256 chances and there are only 8 chances of classifying 7 cups right and 1 wrong. We can thus increase the sensitiveness of the experiment while still using the same number of cups.

- **Conclusion:**

The statistical test associated with this experiment is known as Fisher's exact test. It is now employed in the analysis of 2×2 contingency tables. Although there are some points of controversy regarding this test which raise debate up until this day; Fisher's test showed very conclusively that Dr. Bristol could indeed tell the difference between tea with milk added first and tea with milk added after.

- **References:**

Articles from Students

1. **https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2012.00620.x**
2. **https://en.wikipedia.org/wiki/Lady_tasting_tea**
3. **https://home.iitk.ac.in/~shalab/anova/DOE-RAF.pdf**

A STUDY ON PREDICTORS OF GDP

Sreetama Maitra, 2nd Year

❖ Introduction:

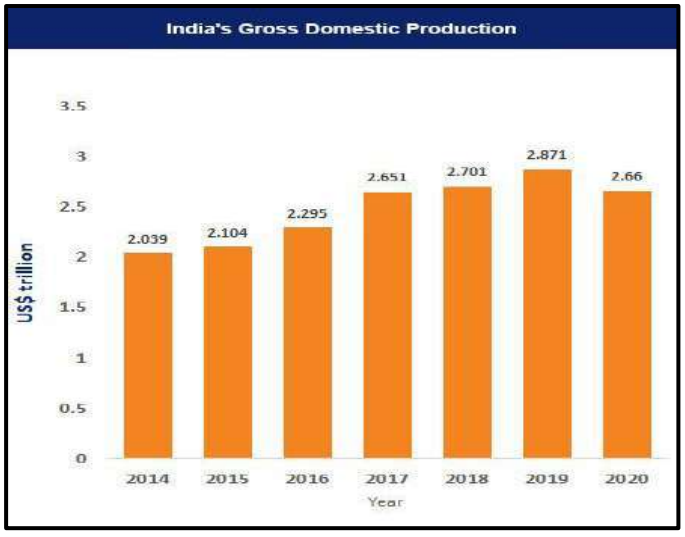
GDP is a very common word in today's economic world. The Bureau of Economic Analysis gives a clear definition of the term. GDP or Gross Domestic Product measures the monetary value of final goods and services i.e., those goods that are bought by the final user, produced in a country in a given time period (say a quarter or a year). It counts all the output generated within the borders of a country.

Here, our purpose is to analyse GDP using the Multivariate Regression Model.

❖ Purpose:

❑ GDP which depends on Agriculture, Service, and industry performance.

The study to derive the actual relationship between dependent variable GDP and independent variables agriculture, industry, and service sector by using multivariable regression model is an application of the Multivariable Regression Model.



India's GDP across the years

❖ **Methodology:**

Multivariable Regression Model technique is a widely recognised technique to find out the nature of relationship between the predictor and response variables accurately and estimate the impact of independent variables on dependent variables. The process of finding the mathematical function which describes the relationship between a dependent variable and one or more independent variables is regression analysis.

❖ **Rationale of Research:**

Most of the article in descriptive cum survey, researcher will go through quantitatively and analyse the highly significant data (GDP and agriculture, agriculture and Industry, Industry and Service sector, Service Sector and GDP) and establish the linear relationship between all the factors and find out the situation of GDP after using initial conditions. No research using the Multivariable Regression Model has been done as it is a newly developed model. So, the researcher intended to conduct the research as empirical validity of their own developed model.

❖ **Research Methodology:**

The linear relationship between four factors will be given as,

$$\frac{A_{11}x_1}{\sigma_1} + \frac{A_{12}x_2}{\sigma_2} + \frac{A_{13}x_3}{\sigma_3} + \frac{A_{14}x_4}{\sigma_4} = 0$$

Where $A_{11}, A_{12}, A_{13}, A_{14}$ are cofactors of correlation coefficient matrix.

$\sigma_1, \sigma_2, \sigma_3, \sigma_4$ are standard deviations of the individual data.

$x_i = X_i - \mu_i.$

Here researchers find cofactors of correlation coefficient matrix, mean and standard deviation of the individual data and represent the linear relationship between dependent and independent variables.

❖ **Results and Discussion:**

❑ **Linear relationship between Agriculture, Industry, Services and GDP**

The association that we have in the method and methodology used to interpret and analyse the data taken from the source of the bureau of statistics is given in table 1.

Table 1: Sectoral Contribution to GDP

FY	Real GDP (X_1) Rs. In Billion	Agriculture (X_2) Rs. In Billion	Industry (X_3) Rs. In Billion	Services (X_4) Rs. In Billion
2009/10	565.76	205.52	91.29	293.27
2010/11	587.53	214.79	95.25	303.32
2011/12	614.64	224.73	98.11	318.52
2012/13	637.77	227.19	100.73	336.76
2013/14	674.23	237.52	107.84	357.69
2014/15	694.27	240.14	109.40	374.26
2015/16	695.69	240.68	102.44	383.06
2016/17	749.55	253.20	115.14	414.04
2017/18	797.15	260.33	126.16	444.06
2018/19	850.93	273.51	135.90	476.27
2019/20	870.25	280.59	140.29	485.76
Total	$\sum X_1 = 7737.77$	$\sum X_2 = 2658.20$	$\sum X_3 = 1222.55$	$\sum X_4 = 4187.01$

❖ **Computation:**

Mean $= \mu_1 = \frac{\sum X_1}{N} = \frac{7737.77}{11} = 703.4336$

The population standard deviation σ is applicable where the entire population can be known along with the square root of variance.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^2 - \mu^2)}$$

where x_i is an individual value.

μ is the mean or expected value.

N is the total number of values.

Standard Deviation $(\sigma_1) = \sqrt{\frac{1}{11}\sum_{i=1}^{11}(x_i^2 - \mu^2)} = 98.1987$

Similarly, one can also compute mean and standard deviation for agriculture, industry and services respectively.

Mean $(\mu_2) = 241.6545$	Standard deviation $(\sigma_2) = 22.4746$
Mean $(\mu_3) = 111.1409$	Standard deviation $(\sigma_3) = 15.7210$
Mean $(\mu_4) = 380.6373$	Standard deviation $(\sigma_4) = 64.1748$

A Pearson's correlation coefficient (r) whose value lies between -1 to 1 inclusively. The formula is given by,

$$r = \frac{n[\sum xy - (\sum x)(\sum y)]}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Using the formula for computation one can find all the association values,

$r_{11} = 1.0000$	$r_{12} = 0.9952$	$r_{13} = 0.9835$	$r_{14} = 0.9986$
$r_{21} = 0.9952$	$r_{22} = 1.0000$	$r_{23} = 0.9753$	$r_{24} = 0.9833$
$r_{31} = 0.9835$	$r_{32} = 0.9753$	$r_{33} = 1.0000$	$r_{34} = 0.9743$

$r_{41} = 0.9986$	$r_{42} = 0.9833$	$r_{43} = 0.9743$	$r_{44} = 1.0000$

Now we will put these values in matrix form and the matrix is called correlation coefficient matrix:

$$A = \begin{bmatrix} 1 & 0.9952 & 0.9835 & 0.9986 \\ 0.9952 & 1 & 0.9753 & 0.9833 \\ 0.9835 & 0.9753 & 1 & 0.9743 \\ 0.9986 & 0.9833 & 0.9743 & 1 \end{bmatrix}$$

Then the cofactors of the matrix will be as follows:

- $A_{11} = (-1)^{1+1}|1\ 0.9753\ 0.9833\ 0.9753\ 1\ 0.9743\ 0.9833\ 0.9743\ 1| = 0.001382$
- $A_{12} = (-1)^{1+2}|0.9952\ 0.9753\ 0.9833\ 0.9835\ 1\ 0.9743\ 0.9986\ 0.9743\ 1| = -0.0004912$
- $A_{13} = (-1)^{1+3}|0.9952\ 1\ 0.9833\ 0.9835\ 0.9753\ 0.9743\ 0.9986\ 0.9833\ 1| = -0.0001206$
- $A_{14} = (-1)^{1+4}|0.9952\ 1\ 0.9753\ 0.9835\ 0.9753\ 1\ 0.9986\ 0.9833\ 0.9743| = -0.00077984$

By substituting cofactor of correlation coefficient matrix, standard deviation and $x_i = X_i - \mu_i$ in the first equation, we get,

$$0.001382 \times \frac{(X_1 - 703.4336)}{98.1987} + (-0.0004912) \times \frac{(X_2 - 241.6545)}{22.4746} + (-0.0001206) \times \frac{(X_3 - 111.1409)}{15.721009} + (-0.00077984) \times \frac{(X_4 - 380.6373)}{64.1748} = 0$$

This is the required linear relationship between dependent variable GDP and independent variables agriculture, industry and service.

❖ **Conclusion:**

GDP can only be hypothetically negative. Whereas, in real life scenarios, it cannot be negative or even 0 as a country needs forex (foreign exchange market) to import which will not be available for a zero production country. But hypothetically, if a country has negative net exports and it is not covered by consumption or investment, then GDP can be negative. The negative GDP in the Multivariable Regression Model at initial conditions reflects that the stated amount will be consumed from savings of the previous year for operation of the economy at said hypothetical zero production. The Multivariate Regression Model should be applied to explain all national accounting indicators such as Gross National Income, National Income, Net National Production, Personal Income and Disposable Personal Income.

RATIONALISING MUSIC USING MEASURE THEORY

Arshiya Paul, 1st Year
Aditya Saha, 1st Year

Measure theory is the formal underpinning for how mathematicians define integration and probability. Statistics is founded on probability, and the modern formulation of probability theory is founded on measure theory. It is the tool mathematicians use to formalize and study the idea of mass, especially in the continuum. Music, on the other hand, is less straightforward – rather, quite mysterious for a number of reasons. Primarily, because it's nearly impossible to define. What might be music to some might not be so for others. But even if we were to ignore its ambiguous and subjective nature, topics seemingly as simple as harmony and cacophony have had people stumped for ages. Although some might consider it sacrilege to tie music and maths – given that both are infinitely complex in their respective domains – it is possible, in some ways, to link them.

Pythagoras made the first concrete argument for a fundamental link between music and maths. Reportedly, he experimented with the notes produced when plucking strings of different lengths. He found that some specific ratios of string lengths created pleasing combinations (harmonies) and others did not.
To add more clarity: let us play a musical note of a given frequency, say 220 Hz. Let us then choose some number 'r' that lies between 1 and 2, and play a second musical note whose frequency is 'r' times the frequency of the first note – 'r' * 220 Hz.
It has been observed that for some values of this ratio 'r', like 1.5, the two notes sound harmonious together, but for others – such as $\sqrt{2}$, they sound cacophonous.

But if we discard the obvious task of listening, **how can we determine, just by analysing this number 'r', whether two separate frequencies (or notes), when played together, will sound harmonious or cacophonous?**

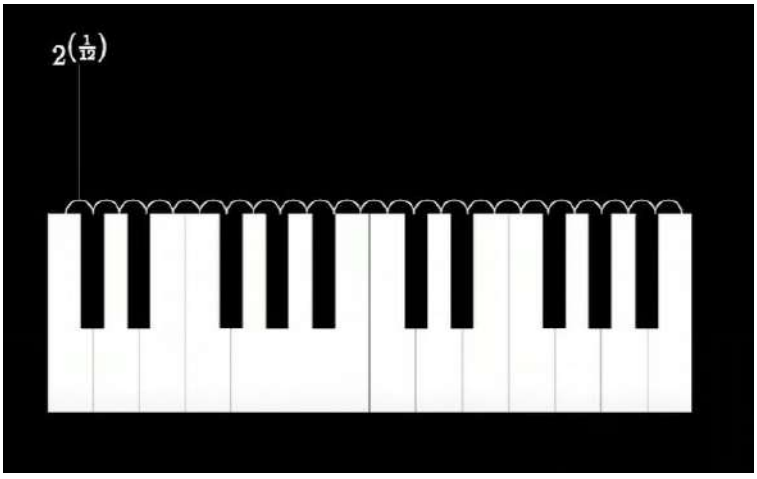
Just from the above example, we might be led to conclude that the two notes sound good when 'r' is a **rational** number and bad when it is **irrational**.
Pythagoras observed several ratios of sound wave frequencies and the

corresponding intervals between them, including 4:3 (known to musicians as the interval of a perfect fourth), 3:2 (a perfect fifth), and 8:5 (a perfect sixth). The ratio of 2:1 is known as the octave (8 tones apart within a musical scale). When the frequency of one tone is twice the rate of another, the first tone is said to be an octave higher than the second tone, yet interestingly the tones are often perceived as being **almost identical** (this is why we have chosen 'r' to range from 1 to 2).

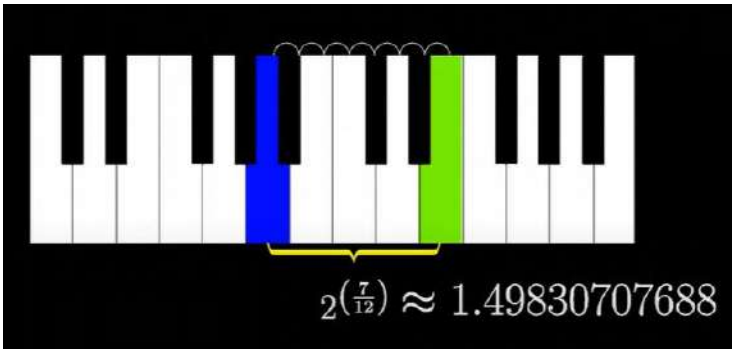
However, most rational numbers actually sound bad!

For instance, the ratios 211:198, 1093:826, or 2138:1873 all lead to unpleasant sounds. Now, if we compare the two types of ratios, we realise that the issue may be that a ratio like 211:198 is more "**complicated**" than one like 4:3 (one simple way to measure the complexity of a rational number is to consider the size of its denominator when it is written in reduced form).

Even still, we find that it is not possible to limit harmoniousness to rational numbers only. Though it may seem counter-intuitive, **plenty of notes sound pleasant together even if the ratios between them are irrational**, hence debunking our initial assumption.
In fact, some instruments are tuned in terms of **irrational** intervals. A piano, in particular, is tuned such that each half-step increase corresponds to multiplying the original frequency by the twelfth root of 2.



This means that if we were to consider a harmonious interval – say, a fifth – the ratio of frequencies when played on a piano will NOT be 3/2 (as we had defined earlier), but some power of the twelfth root of 2, in this case $2^{7/12}$, which is very close to 3/2.



This method of tuning is also known as **musical temperament.** The reason why this method works (albeit reliant on irrational intervals) is because the powers of the twelfth root of 2 tend to remain within a 1% margin of error of simple rational numbers.

$2^{(\frac{1}{12})} \approx 1.05946309$ is close to $16/15$ with 0.67% error

$2^{(\frac{2}{12})} \approx 1.12246204$ is close to $9/8$ with 0.22% error

$2^{(\frac{3}{12})} \approx 1.18920711$ is close to $6/5$ with 0.89% error

$2^{(\frac{4}{12})} \approx 1.25992104$ is close to $5/4$ with 0.79% error

$2^{(\frac{5}{12})} \approx 1.33483985$ is close to $4/3$ with 0.11% error

$2^{(\frac{7}{12})} \approx 1.49830707$ is close to $3/2$ with 0.11% error

$2^{(\frac{8}{12})} \approx 1.58740105$ is close to $8/5$ with 0.78% error

$2^{(\frac{9}{12})} \approx 1.68179283$ is close to $5/3$ with 0.90% error

$2^{(\frac{10}{12})} \approx 1.78179743$ is close to $16/9$ with 0.22% error

To finally edit our definition of harmoniousness, we now know that two notes sound pleasant when the ratio 'r' of their frequencies **is sufficiently close to a rational number with a low denominator.**

But it is also possible that someone with a particularly acute musical sense would be able to hear and find pleasure in the patterns resulting from more complicated ratios (like 211:198 or 1093:826), as well as numbers closely

approximating these ratios. This raises an interesting question - **suppose there is a musical savant who finds pleasure in all pairs of notes whose frequencies have a rational ratio. In that case, would she find all ratios r between 1 and 2 – even the irrational ones – harmonious?** Because after all, for any given real number, we can always find a rational number arbitrarily close to it, just as $3/2$ is close to $2^{7/12}$. This is because, by construction $\forall\, r \in R$ and $\forall \varepsilon > 0$, $\exists\, q \in Q$ such that $|r - q| < \varepsilon$, i.e., the **set of rational numbers is dense in the set of real numbers** (R represents the set of real numbers and Q represents the set of rational numbers).

In this context, we have a challenging question - **can we cover all the rational numbers between 0 and 1 with open intervals such that the sum of their lengths is strictly less than 1?**

The answer is yes, although the task feels impossible as rational numbers are **dense** in real numbers. So how could we possibly cover all the rational numbers without just covering the entire interval from 0 to 1 itself?

Let us first enumerate all the rationals in the interval (0,1) so that they are countable in the following fashion - $1/2$, $1/3$, $2/3$, $1/4$, $3/4$, $1/5$, $2/5$, $3/5$, $4/5$, $1/6$, $5/6$, ... and so on.

Next, to ensure that all the rational numbers are covered, we are going to assign one specific interval to each rational. Now, it seems much clearer that the sum of their lengths can be less than 1, since each particular interval can be as small as we want and still cover its designated rational. **In fact, the sum of their lengths can be any positive number!**

In order to prove this, we can just choose an infinite sum that converges to 1 (of the form $\sum_{n=1}^{\infty} a_n = 1$). For example,

$1/2 + 1/4 + 1/8 + 1/16 + \ldots\ldots\ldots\ldots \rightarrow 1$, which is basically the long-hand form of $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$

We pick ε such that $0 < \varepsilon < 1$, then we have

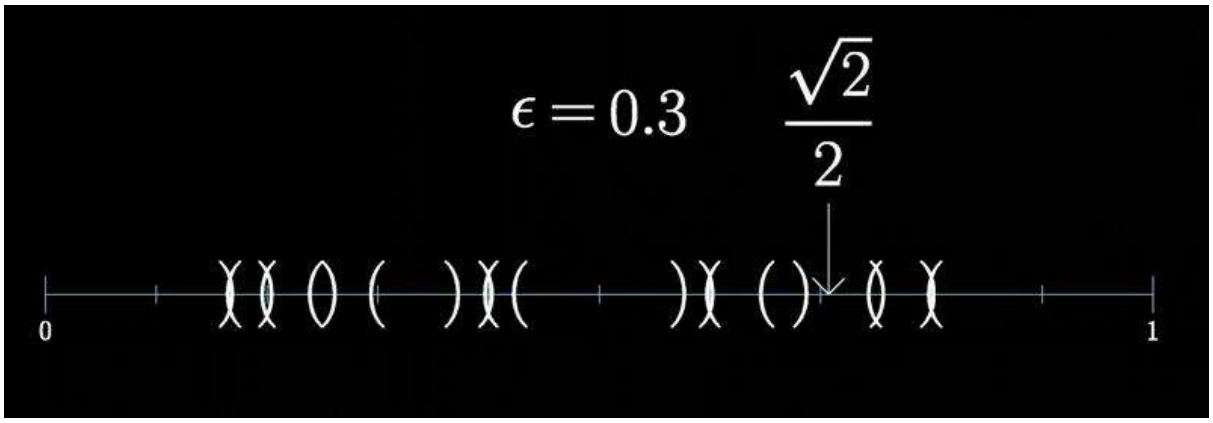
$\varepsilon/2 + \varepsilon/4 + \varepsilon/8 + \varepsilon/16 + \ldots\ldots\ldots\ldots \rightarrow \varepsilon$

Now, let us scale the n^{th} interval covering the n^{th} rational number in the given list to $\varepsilon/2^n$. Hence the sum of their lengths will be less than 1 as, $\sum_{n=1}^{\infty} \varepsilon/2^n \rightarrow \varepsilon$ and $0 < \varepsilon < 1$. We note that our sum can be arbitrarily small as we have the freedom to choose the value of ε.

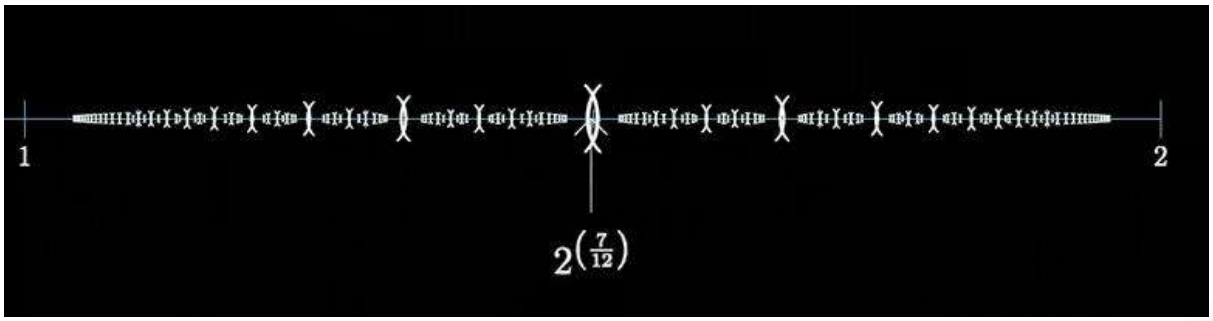
Hence, it is possible to cover all the rationals in (0,1) using infinitely many open intervals such that the sum of their lengths is less than 1. It is to be noted that Lebesgue's measure is used to determine the mass of a subset of real numbers and is defined as **the greatest lower bound for the sum of lengths of open intervals covering the set.** From the proof shown above, it can be said that **the Lebesgue measure of the rational numbers is 0** (as ϵ can be as small as we want).

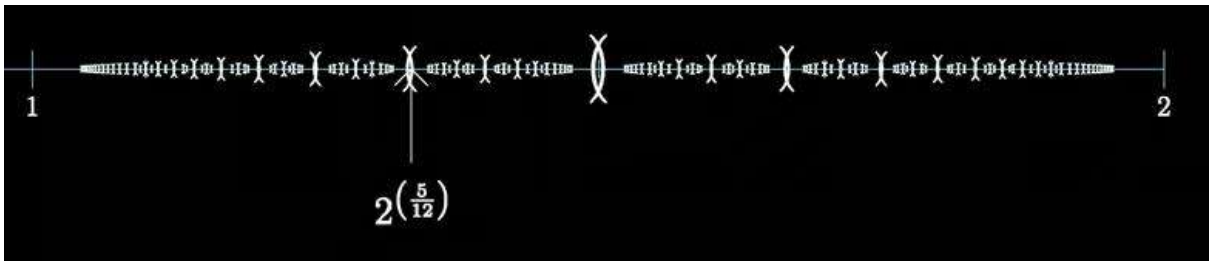
Now, if we pick $\epsilon = 0.3$ and choose one number between 0 and 1 at random, there is a 70% chance that it is outside those infinitely many intervals. For example, $\sqrt{2}/2$ is among those 70%. Hence, from our previous discussion,

$\sqrt{2}/2$ is not covered.
=> rationals which are close to $\sqrt{2}/2$ must have large denominators.
=> if r=$\sqrt{2}/2$ then the two notes will be unpleasant to hear.



Now let us pick an even smaller value of, say, $\epsilon = 0.1$ and shift our setup from the interval (0,1) to (1,2). On observing which numbers fall among the special 1% covered by our tiny open intervals, we see that almost all of them are harmonious. For instance, the harmonious irrational number $2^{7/12}$ is very close to 3/2 and also the interval around 4/3 is smaller but large enough to cover $2^{5/12}$.





Now let us go back to the musical savant who finds pleasure in all pairs of notes whose frequencies have a rational ratio. For her, the harmonious numbers are precisely those 1% covered by the intervals, provided her tolerance for error goes down exponentially for more complicated rationals. In other words, the seemingly paradoxical fact that a collection of intervals can densely populate a range while only covering 1% of its values, implies that harmonious numbers are rare, even for the savant. It is indeed surprising that the savant we defined could find 99% of all ratios cacophonous.

It is strange and beautiful to see how these two seemingly unrelated topics are connected. In fact, discoveries such as these aren't simply discoveries, but reservoirs of history. What we know about the mathematics of musical harmony has its roots in the findings of Pythagoras, who while describing his theory of music and mathematics, wrote: **"There is geometry in the humming of the strings, and there is music in the spacing of the spheres."** We hope we have been able to explain the first part of this statement to some degree in this article.

The second part refers to the prevalent cosmological theory in proto-Greece at the time, which was that Earth resided at the centre of a series of concentric "celestial spheres." These spheres rotated around the Earth, and the various objects visible in the sky were each attached to one of these spheres, which Pythagoras believed made music that could cure ailments. Along with probably being one of the first to propose the idea of music therapy, he predicted that the universe, indeed, produces a variety of sounds.
We now know that he was right about what he said – and we wouldn't without the mathematics and technology we possess today. If there is anything we can infer from all of this, it is that mathematics isn't just a subject to study in class, but a tool that paves the way for answers from the

beyond. As scientists of our time, it is our duty to use what we have to unravel the mystery and music of our universe.

❖ **References:**

1. The Theory of Statistics and Its Applications by Dennis D. Cox, Rice University

https://www.stat.rice.edu/~dcox/Stat581/chap1-2.pdf

2. The Music of the Spheres – Sensory Studies

https://www.sensorystudies.org/picture-gallery/spheres_image/#:~:text=Auditory%20culture%20is%20thereby%20extended,the%20visual%20and%20the%20aural

3. Music and Mathematics – A Pythagorean theory

https://www.unyp.cz/news/music-and-mathematics-pythagorean-perspective#:~:text=Pythagoras%20observed%20several%20ratios%20of,fifth%2C%20seven%20semitones%20apart

RELATIONSHIP BETWEEN MATHEMATICS AND STATISTICS

Manopriya Pal, 3rd Year

❖ **Introduction:**

Mathematics is a science that deals with the logic of scope, quantity & agreement. It is used in every step of our life as a building block. On the other hand, statistics is the science that collects numerical information and analyses it in large quantities.

Studying mathematics and statistics facilitates us to develop the ability to think creatively, critically, strategically, and logically. We learn to structure and organise, carry out procedures flexibly and accurately, process and communicate information, and enjoy intellectual challenges.

Now there's a common question about whether statistics is a part of mathematics or not. The answer is - Statistics is a branch of applied mathematics. The mathematical theories behind statistics are virtually supported by differential and integral calculus, linear algebra, and probability theory.

Mathematics deals with the knowledge of space, measures, and structures in their rudimentary and statistics is a mathematical science, not a subfield of mathematics. We even have aphorisms to precise some ways in which our science differs from mathematics. George Cobb: "In mathematics, context obscures structure. In data analysis, context provides meaning." David Moore: "Mathematical theorems are true; statistical methods are sometimes effective when used with skill." That version of those aphorisms apply whenever mathematics models phenomena in another field solely emphasising that statistics is another field.

Let us now focus on the difference between mathematics and statistics in the aspect of modelling different phenomena.

Math always follows a uniform definition-theorem-proof structure. No matter what branch of mathematics we're studying, whether it is algebraic number theory or real analysis, the structure of a mathematical argument is more or less constant.

Statistics is a field of study that deals with the collection, organisation, analysis, interpretation and display of data. Broadly, there are two main goals of statistics. The first is statistical inference: analysing the data to understand the processes that gave rise to it; the second is a prediction: using patterns from past data to predict the future.

Statistical models depict the random behaviour of relationships among variables which are generally not considered in mathematical models. It uses equations which involve some observed variables and some disturbances that consist of all the variables, which are considered irrelevant for this model as well as unforeseen events.

❖ **Modelling the Real World:**

A model is a way of expressing the relationship between one set of variables to another set of variables through some functional form. This functional form may involve some unknown parameters. When the model involves a random error term it's called a statistical model.

Both mathematics and statistics are tools we tend to use to model and understand the world, but they do so in very different ways. Maths creates an idealised model of reality where everything is obvious and deterministic; statistics accepts that all knowledge is uncertain and tries to form a sense of the data despite all the randomness. As for which approach is better — both approaches have their benefits and downsides.

Maths is good for modelling domains where the rules are logical and can be expressed with equations. One example of this is physical processes: simply a small set of rules is remarkably good for predicting what happens in the real world. Moreover, once we have figured out the mathematical laws that govern a system, they're infinitely generalizable — Newton's laws can accurately predict the motion of celestial bodies although we've only observed apples falling from trees. On the other hand, maths is awkward at handling error and uncertainty. Mathematicians produce a perfect version of reality and hope that it's close enough to the real thing. It uses the model of the type:

$$Y = \alpha + \beta x$$

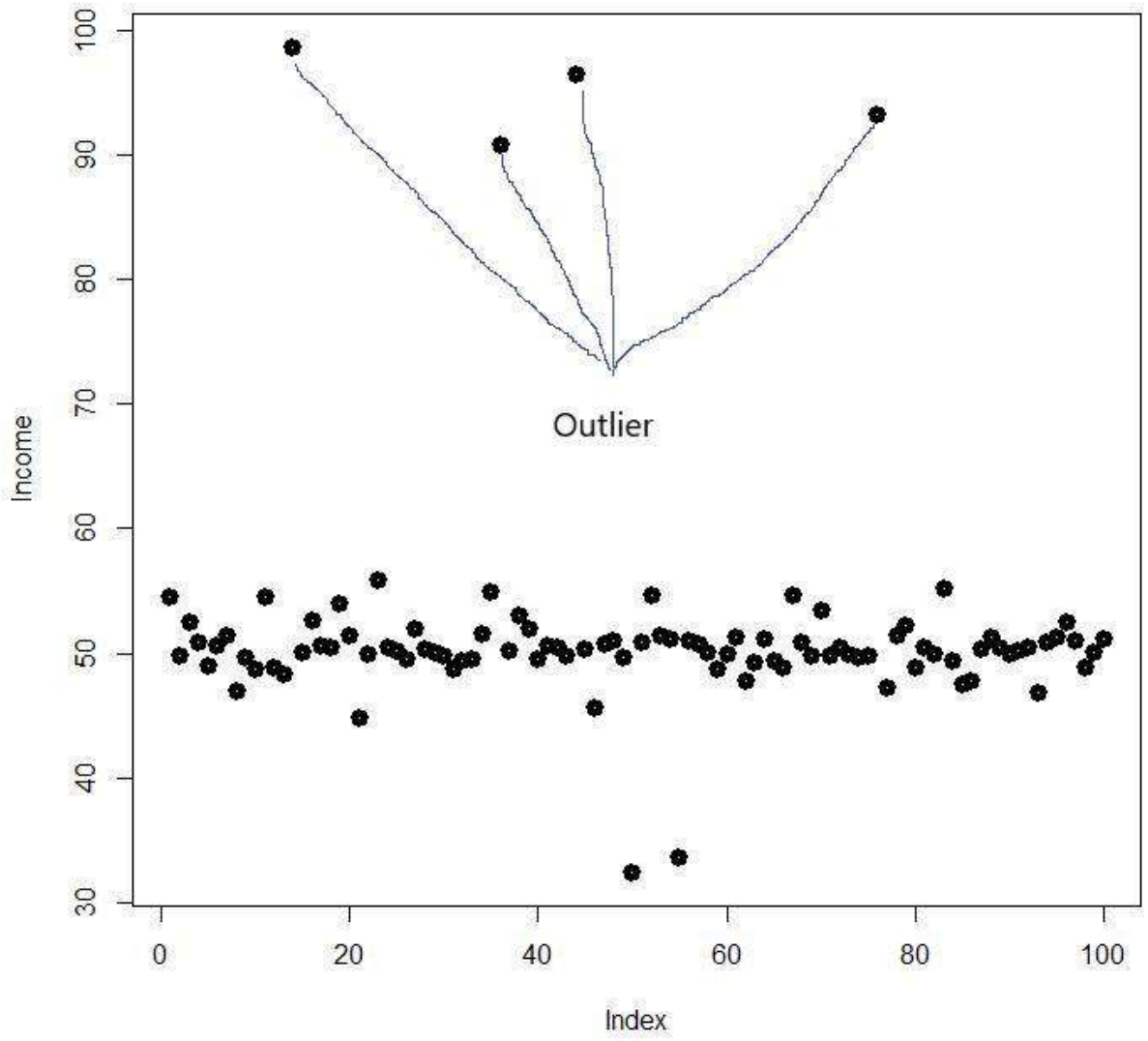
where x is the explanatory variable, Y is the dependent variable and α, and β are the model parameters. We see that it doesn't involve any error term that takes into account the effect of extraneous variables or includes the error of choosing the functional form.

Statistics come into the picture when the rules of the game are uncertain. Instead of ignoring the error, statistics embraces uncertainty. Every value has a confidence interval where we can expect it to be right about 95% of the time, but we can never be 100% sure about anything. But given enough data, the correct model will separate the signal from the noise. This makes statistics a strong tool when there are many unknown confounding factors, like modelling sociological phenomena or anything involving human decisions.

The drawback is that statistics only works on the sample space where you have data; most models are bad at extrapolating past the range of data that it's trained on. In other words, if we use a regression model with information about apples falling from trees, it'll eventually be pretty sensible at predicting other apples falling from trees, but it fails to predict the path of the moon. Thus, maths helps us to understand the system at a deeper, more elementary level than statistics.

Statistics work with real data, which tends to be messy and doesn't lend itself easily to clean rigorous definitions. For example, we consider the concept of an "outlier". Many statistical methods behave badly when the data contains outliers, so it's a standard practice to identify outliers and remove them. But what exactly constitutes an outlier depends on several criteria, like how many data points we are having, how far it is from the

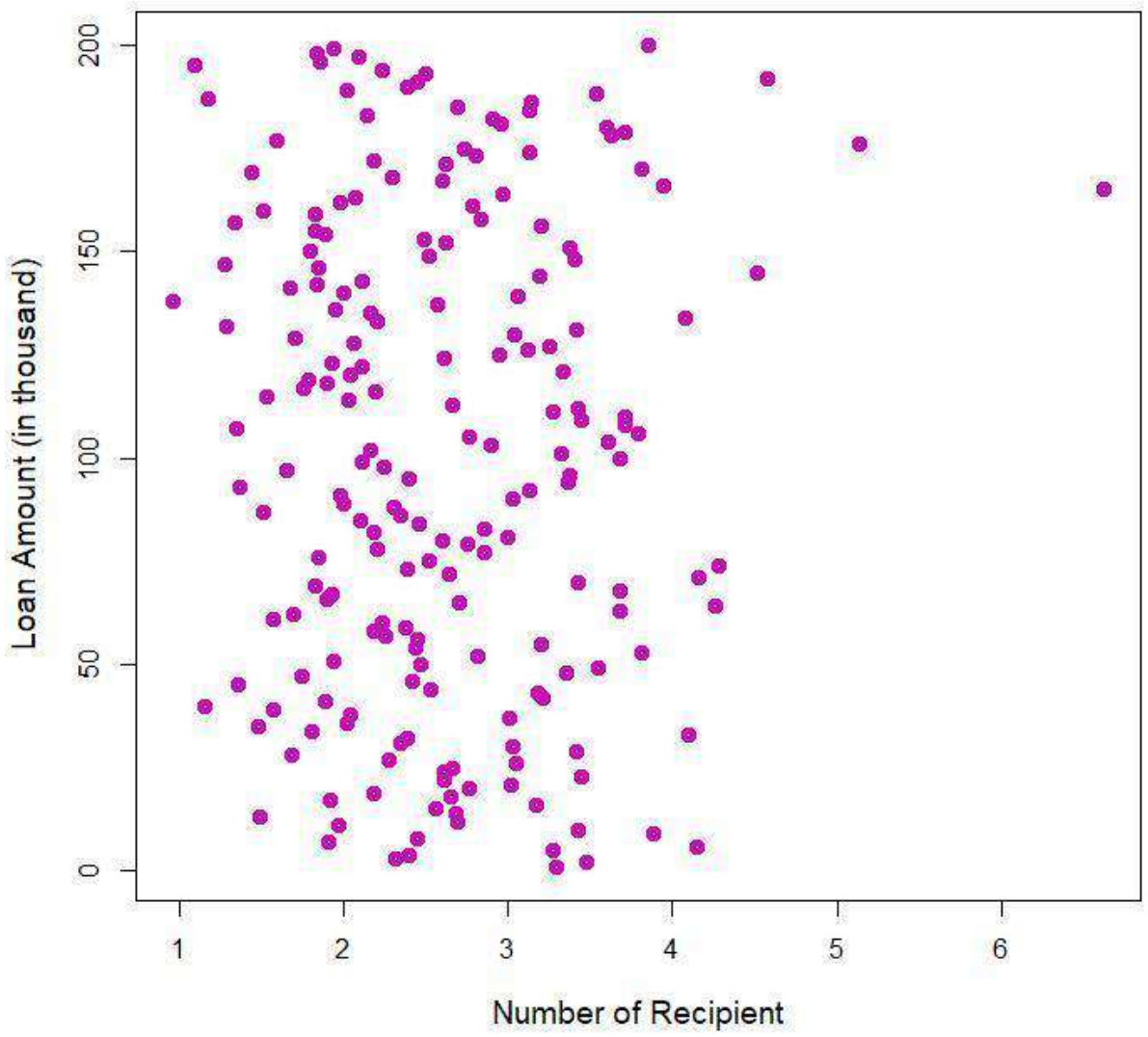
rest of the points, and what kind of model we want to fit.



In the above plot, 4 points are potential outliers. We remove them, or keep them, or maybe remove one or two of them that depend upon us.

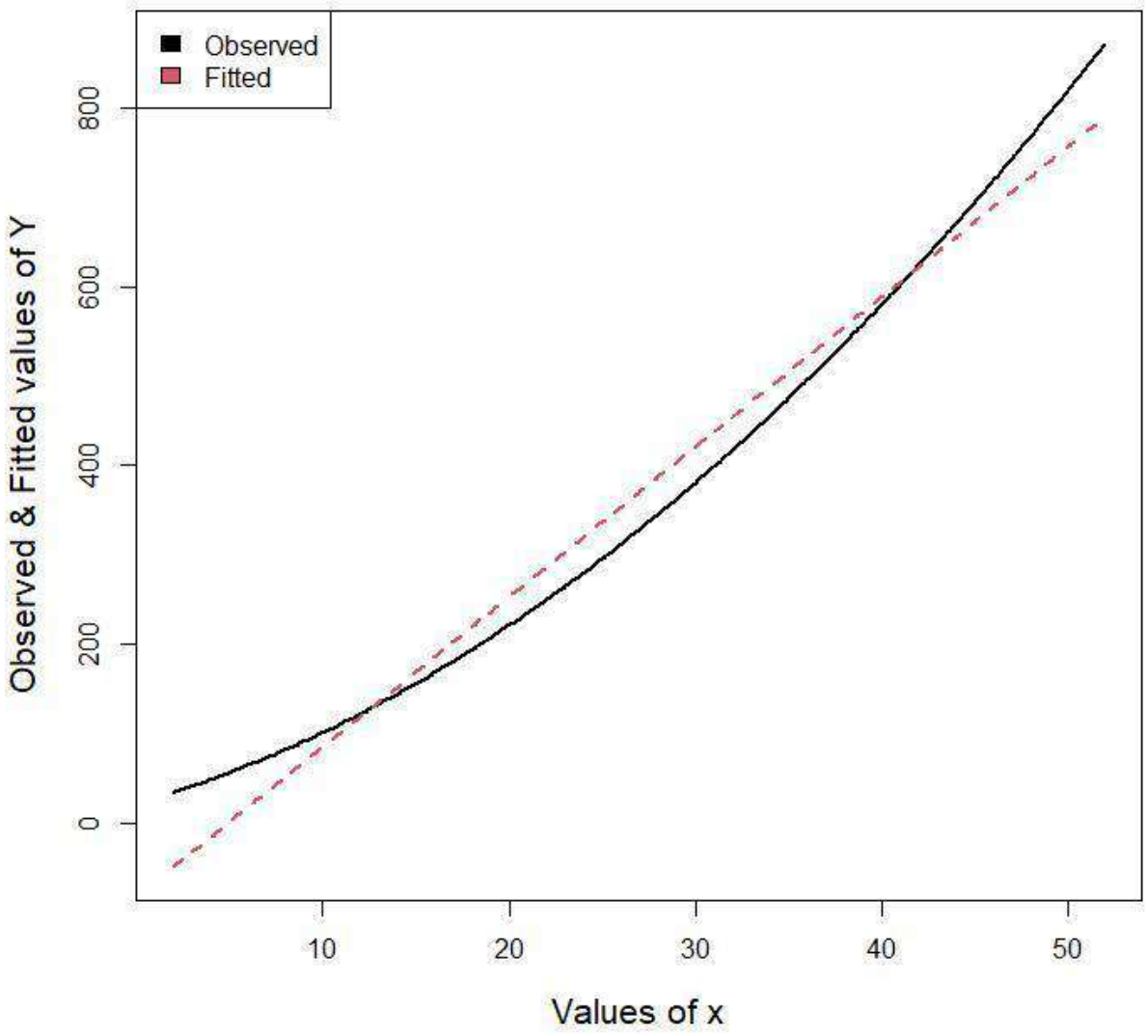
For another example, consider p-values. Usually, if a p-value is under 0.05, it may be thought of as statistically significant. But this value is just a guideline, not a law — it's not like 0.049 is properly significant and 0.051 is not. So to consider whether it is significant or not is up to us.
Take another example: heteroscedasticity. This means the variance is not equal for various components of the dataset. Heteroscedasticity is not good because a lot of models, like the classical linear model, assume that the variance is constant, and when this assumption is profaned then we'll get

wrong results, so we need to use a different model.



Is this data heteroscedastic, or does it seem like the variance is uneven as there are so few points to the right of 4? Is the downside serious enough that fitting a linear model is invalid? There's no correct answer, we've to use our judgement.

Another example: consider a linear regression model with two variables X and Y. When we plot the points on a graph, we should expect the points to roughly lie on a straight line. Not exactly a line, of course, just roughly linear. But what if we get this:



There is some evidence of non-linearity, but how much "bendiness" can we accept before the data is not "roughly linear" and we have to use a different model? Once more, there's no correct answer, and we have to use our judgement.

We can see in statistics unlike maths, there is no universal procedure which will tell whether or not the information satisfies these assumptions.

Here are some common things that statistical models assume:

1) A stochastic variable is drawn from a normal (Gaussian) distribution.
2) Two random variables are independent.
3) Two random variables satisfy a linear relationship.
4) Variance is constant.

The data isn't going to exactly fit a normal distribution, so all of these are approximations. In practice, we may have samples from exponential or log-normal distributions. Various configurations in the games of cards were of special interest to gamblers, these are count data and Poisson distribution is a classic choice here. Interest could be in proportion- the presence or absence of disease. Here Bernoulli distribution is a classic choice. If our purpose is to know the number of diseased persons, here Binomial distribution is appropriate.

A proverb in statistics goes: "all models are wrong, but some are useful". But if data deviates considerably from the model assumptions, then the model breaks down and we get garbage results. For example, when we use the following model:

$$Y = \alpha + \beta x + e$$

We usually assume that the conditional mean of error is zero that is $E(e|x) = 0$. Also, we assume that Y values are distributed symmetrically around their respective (conditional) mean values and the regression line passes through these conditional mean values. Otherwise, there is no meaning in using such a model. So, there is no universal black-and-white procedure to decide if data is normally distributed, at some point, we have to step in and apply our judgement.

❖ Conclusion

Mathematics follows a rigid theorem and proof structure throughout the complete discipline. There are well-defined facts which are laid down as a part of proven human knowledge which has minimal scope for modification.

However, Statistics is a discipline where people handle real-life data. This factor makes this field of study more abstract, where individuals have to develop newer solutions to issues which will be novel.

Mathematics is a very broad domain of study, encompassing just about all quantitative disciplines whereas Statistics is a specific discipline within it, deeply related to Applied Mathematics. Mathematical theory is rarely informative about functional forms. We have to use statistical methods to choose the functional form, as well.

Statistics is strictly associated with physical data and its interpretation; hence it has a restricted scope. Mathematics, however, also deals with abstract ideas which may be metaphysical. Hence, Mathematics has a much wider scope than Statistics.

❖ **References:**

1) https://luckytoilet.wordpress.com/2017/09/06/whats-the-difference-between-mathematics-and-statistics/
2) https://leverageedu.com/blog/what-is-the-difference-between-mathematics-and-statistics/

ROBUSTNESS IN ESTIMATION THEORY

Tamasha Dutta, 3rd Year

❖ Introduction:

In Statistical theory, Robustness of an estimator is defined as the problem of finding estimators that are likely to have small departures from the assumed statistical model, for a wide range of probability distributions, especially for non-normal distributions.

The robust statistics was first introduced by John Tukey (in 1960), then, over time it has been developed by Peter Huber (in 1964), and Frank Hampel (in 1971).

Robust statistical methods have been introduced for many problems, like, the estimation of parameters, for different regression models, probabilistic model, etc. The basic objective of this procedure is to remove the effect of outliers in a dataset. Also, this procedure helps us to provide a good observation of the parameters, when there are some small departures from a specified distribution.

In Statistical theory, there are many given definitions of robustness. A robust statistic is defined as the resistant to the errors, produced due to the deviations from assumption, i.e., if the assumptions only occurred, approximately, then the robust estimator will also be having an efficiency, and small bias, also the bias is tending towards 0 as the sample size tends towards infinity.

❖ Measurements of Robustness:

To measure robustness, there are generally, three methods –
1) The Breakdown Point Method
2) The Sensitivity Curve Method
3) The Influence Function Method

❑ The Breakdown Point Method:

The breakdown point method of an estimator is defined as the smallest fraction of the arbitrarily large observations or the outliers, in the given dataset that overestimates the estimator.

According to the theory, the higher the breakdown point value of the estimator, the more robust it is.

❑ **The Sensitivity Curve Method:**

The sensitivity curve method is defined as the measure of the effect of a single outlier on the estimator.

This method is used to compute the difference between the estimate value for a given sample $(x_1, x_2, \dots, x_n)$ and the estimate, when a particular observation (x) is added to the sample. The difference is obtained by the fraction of contamination $1/(n+1)$, where n is the sample size.

Hence, for a given estimator θ, the sensitivity curve function is defined as,

$$\mathbf{SC}\ (x_1, x_2, \dots, x_n, \theta) = \frac{[\theta(x_1,x_2,\dots,x_n,x) - \theta(x_1,x_2,\dots,x_n)]}{\frac{1}{(n+1)}}$$

An estimator is considered as a robust estimator if its sensitivity curve function values are bounded.

❑ **The Influence Function Method:**

The influence function method is defined as the asymptotic part of the sensitivity curve. It does not depend on any finite set of observations, but depends on the specific distribution of the estimator.

The influence function measures the changes in estimation when a certain contamination is added to the distribution. The contaminated function of the distribution (f) can be given as,

$$\mathbf{I} = (1 - \xi) \times f + \xi \times \delta_x$$

Here, δ_x is the Dirac Measure, which is 1 at the point x, and 0 otherwise.

Now, the influence function, is given as,

$$\mathbf{IF}\ (x, f, \theta) = \frac{\theta(I) - \theta(f)}{\xi})$$

An estimator is considered as a robust estimator if the influence function values are bounded.

❖ **Robust Estimate for Central Tendency:**

1) Median:

We know that, median is the middle most value of a certain dataset. For a given set of observations, say, $(x_1, \dots, x_n)$, median is $x_{(n+1)/2}$, for n to be odd, and, $(x_{n/2} + x_{(n+2)/2})/2$, for n to be even.

Now, according to theory, it is stated that the breakdown point of the median is 0.5, i.e., the median can resist 50% of the outliers, without creating any discrepancy in the estimation. Thus, this measure can also be considered as a robust measure of central tendency.

2) Trimmed Mean:

We know that, trimmed mean is the simple average of the sampled data, computed by ignoring the smallest and the largest observations. Thus, in presence of outliers, the trimmed mean ignores the outlier values and then computes the simple mean. Thus, this measure can also be considered as one of the robust measures of central tendency.

❖ **Robust Estimate for Dispersion:**

1) InterQuartile Range:

For a set of observations, the InterQuartile Range (IQR) is defined as, IQR = $Q_3 - Q_1$, where, Q_i is the i^{th} quantile of the set of observations; i = 1, 3, here.

Similarly, as median, $Q_3 - Q_1$ has respectively, 25% outlier resistant power, i.e., the breakdown point of IQR is 0.25.

Thus, in presence of outliers, the IQR is able to ignore the outlier values 25%. Thus, this measure can also be considered as one of the robust measures of dispersion.

❖ **Properties of Robust Estimator:**

1) The Robust estimators provide us with the largest possible breakdown point.
2) The Robust estimators provide us with the highest possible efficiency.

3) The Robust estimators provide us a smooth influence function.

❖ **Example:**

For example, let us consider the dataset {5, 2, 9, 5, 11, 18, 3, 10, 19, 13}, obtained from Binomial (20, 0.5) distribution, without considering any replacement of value.

Now, according to theory, mean is not a robust statistic, while median is a robust statistic of central tendency.

Now, the mean and median of the data is 9.5.

Now, if we add a particular outlier, say 1000, then the mean is 99.5, while the median is coming out as 10, nearly to 9.5.

Thus, according to the Breakdown Point Method, the breakdown point of the sample mean is nearly 1/11. Also, as $n \to \infty$, the breakdown point of the sample means is tending towards 0, which is the worst possible case, here.

Also, in case of median, it is the middle most value of the sample, thus, taking any random sample, Sensitivity Curve value will always be bounded, but, for mean, the difference between the sample mean and the new sample mean may or may not be bounded. Here, for mean, that difference is nearly 90, which is a high value.

Again, using the Influence Curve function, we must conclude the same comment as for Sensitivity curve.

Other than this, the robust estimator helps us to estimate the parameters (Scale and Location Parameters) of a distribution.

❖ **Real life examples of robust estimation:**

In 2004, Simon Newcomb considered a data set, which is related to the speed of light measurements, in the Bayesian Data Analysis.

According to the theory, the plot of the data looked more or less to be normally distributed, with two obvious outliers, which had a very large effect on the mean, taking mean towards them, and took the plot away from the centre of the data.

In that case, if the mean is taken as the measure of the location parameter of the data, then in the presence of outliers, the mean is absurd. Also, due to the

central limit theorem, the distribution of the mean is known to be asymptotically normal, but in presence of outliers the distribution of the mean may be non-normal, even for a fairly large amount of data. Thus, mean is not efficient, in the presence of outliers and the measurement of the location parameter is also absurd.

Now, we came to know that, the trimmed mean is a simple robust estimator of location parameter, which deletes around 10% percentage of observations, from each end of the data, and then computes the mean in the general way.

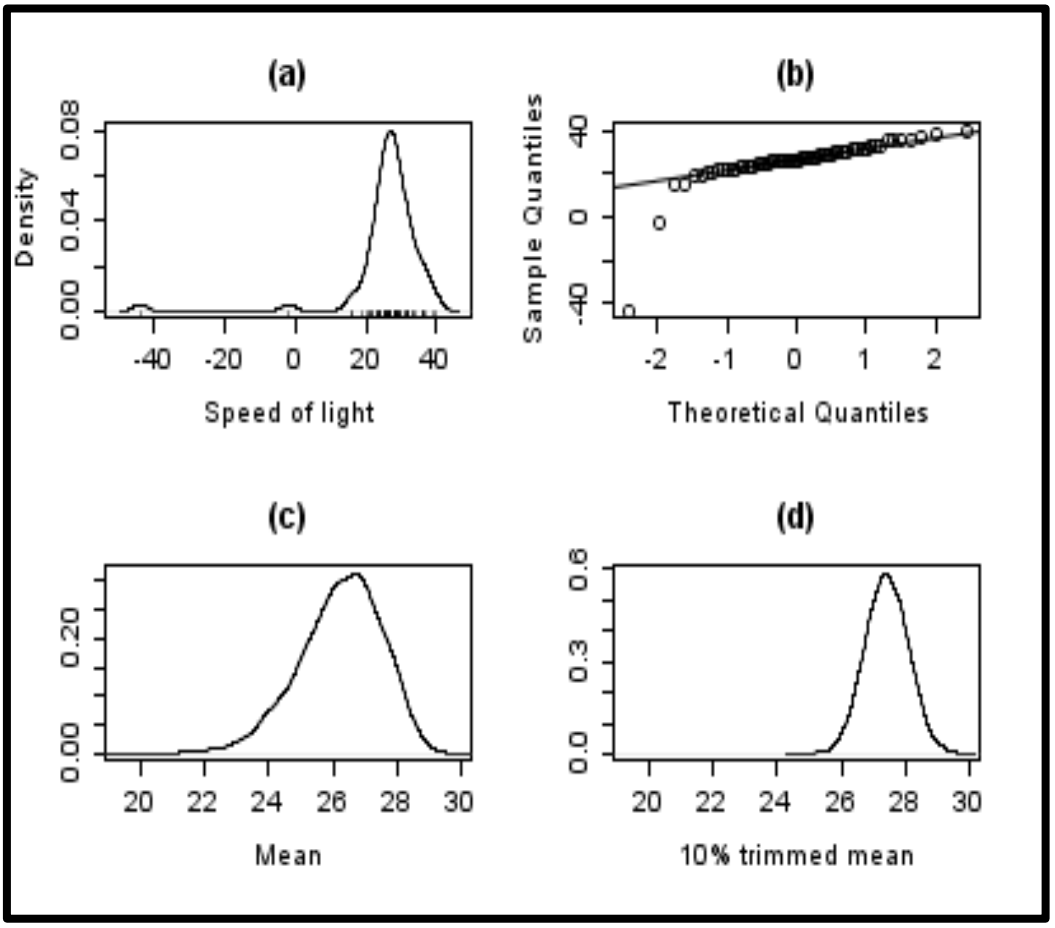


Figure 1: Plots of the Speed of Light data, obtained by Simon Newcomb, for 10000 sample data - (a) Density Plot of the Speed of Light data; (b) Q-Q plot of the Speed of Light data; (c) Plot of the Distribution Mean of of the Speed of Light data; (d) Plot of the 10% Trimmed Mean of of the Speed of Light data.

Thus, the distribution of mean is clearly wider than that of the 10% trimmed mean. Also, according to theory, the distribution of the trimmed mean comes to be approximately normal distribution.

Here, it is known that the trimmed mean performs well in comparison to the arithmetic mean.

Thus, these measurements of robustness of an estimator are very helpful in real-life data analysis.

❖ **References:**

1) https://www.baeldung.com/cs/robust-estimators-in-robust-statistics
2) https://en.m.wikipedia.org/wiki/Robust_statistics

DRIVING INTO THE UNKNOWN: HOW STATISTICS AND PROBABILITY HELP AUTONOMOUS VEHICLES TO HANDLE UNCERTAINTY

Ranit Sarkar, 2nd Year
Soham Choudhury, 2nd Year

❖ Introduction:

An autonomous vehicle, or a driverless vehicle, is one that can operate itself and perform necessary functions without any human intervention, through its ability to sense its surroundings. In this article, we'll explore how statistics and probability are helping to drive the development of autonomous vehicles, and how they are essential for ensuring the safety and reliability of these innovative technologies.

Autonomous vehicles use a variety of sensors, such as cameras, lidar, and radar, to gather data about the environment. This data is then processed using computer vision algorithms, which help the vehicle to identify and classify objects in the environment, such as pedestrians, vehicles, and traffic signs. By interpreting this data, the autonomous vehicle can make informed decisions about how to safely navigate through its environment.

❖ Algorithms Used in Autonomous Vehicles:

Deep learning algorithms, such as **convolutional neural networks** (CNNs) and **recurrent neural networks** (RNNs), are widely used in autonomous vehicles for a variety of tasks such as object detection, semantic segmentation, image recognition, and control. Among these algorithms, YOLO (You Only Look Once) is considered to be one of the best object detection algorithms for autonomous vehicles, due to its speed, accuracy, and real-time performance. However, it is important to note that there are other object detection algorithms such as **Faster R-CNN**, **Single Shot MultiBox Detector (SSD)**, **Mask R-CNN**, **Gradient Boosting**, **k-Nearest Neighbours (k-NN)** and **RetinaNet** which are also widely used and are considered to be competitive alternatives that can be utilised in autonomous vehicles. The choice of algorithm will depend on the specific requirements of the application and the sensor data available.

❑ YOLO:

YOLO (You Only Look Once) is a real-time object detection algorithm. It is a convolutional neural network (CNN) based algorithm that is able to detect and classify objects in an image or video stream. In autonomous vehicles, YOLO is used to detect and track objects in the vehicle's environment, such as other vehicles, pedestrians, and traffic signs. YOLO is also used to process images and video from cameras mounted on the vehicle, such as from cameras used for lane detection, traffic sign recognition, or obstacle detection. The algorithm detects and classifies objects in real-time, making it suitable for use in autonomous vehicles, where fast and accurate object detection is crucial for safe navigation. Additionally, YOLO is used to perform semantic segmentation, which is the process of classifying each pixel in an image to a specific class. This is useful for tasks such as free space detection and driveable area detection.

In this article, we will implement the YOLO algorithm using the OpenCV library.

- **OpenCV:**

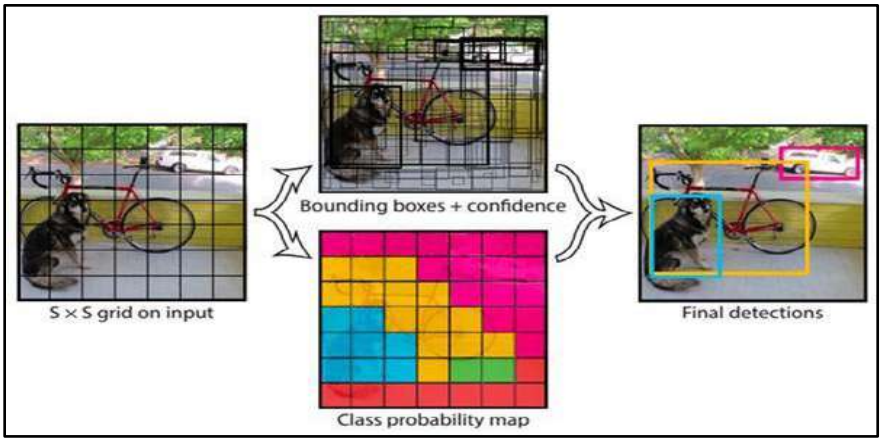
OpenCV is an open-source computer vision library that includes several hundreds of computer vision algorithms. It is used for image processing, object detection, and more. In autonomous vehicles, OpenCV is commonly applied for tasks such as object detection, image recognition, and depth estimation. It provides a variety of other functions, such as edge detection, feature extraction, and image filtering, that are used to extract features from images and videos, crucial for autonomous navigation.

- **Difference between YOLO and OpenCV:**

OpenCV and YOLO are related to each other as they are both used for computer vision tasks, specifically object detection. However, they are not the same. OpenCV is a computer vision library that can be used to implement YOLO and other object detection algorithms. YOLO is a specific object detection algorithm that can be implemented using OpenCV or other libraries that are used for image processing and object recognition.

- **How YOLO works – Graphical Example:**

YOLO helps to detect and classify objects in an image or video frame in a single pass, allowing for fast and efficient processing. It works by dividing the image or video frame into a grid of cells and using machine learning techniques to predict the likelihood that each cell contains an object. This allows YOLO to process the data very efficiently, making it an attractive choice for use in autonomous vehicle applications where fast response times and low computational overhead are important.



❖ **Implementation of YOLO Algorithm using OpenCV in Python:**

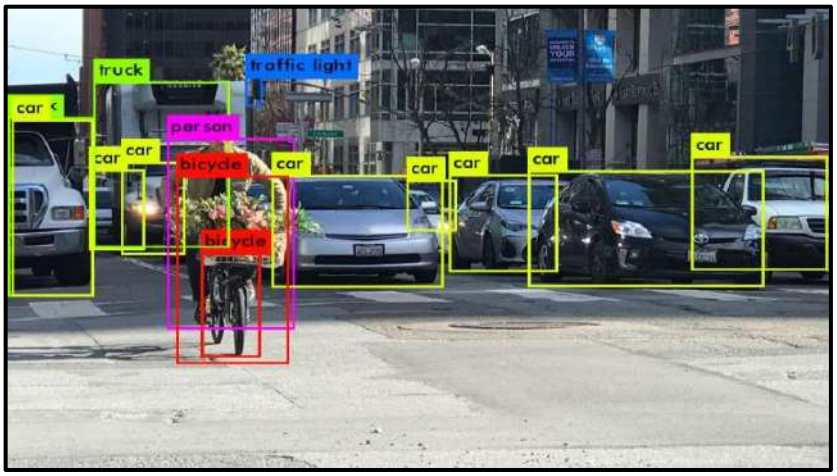
To implement the YOLO algorithm using OpenCV, we need three files viz - 'yolov3.weights', 'yolov3.cfg' and "coco.names" . First, we are going to load the model using the function "cv2.dnn.ReadNet()". This function loads the network into memory and automatically detects configuration and framework based on file name specified.

```
import cv2
import numpy as np
# Load YOLO model
net = cv2.dnn.readNet("yolov3.weights", "yolov3.cfg")
classes = []
with open("coco.names", "r") as f:
    classes = [line.strip() for line in f.readlines()]
layer_names = net.getLayerNames()
output_layers = [layer_names[i[0] - 1] for i in
net.getUnconnectedOutLayers()]
colors = np.random.uniform(0, 255, size=(len(classes), 3))
```

```
# Load image
img = cv2.imread("image.jpg")
img = cv2.resize(img, None, fx=0.4, fy=0.4)
height, width, channels = img.shape
# Detect objects
blob = cv2.dnn.blobFromImage(
    img, 0.00392, (416, 16), (0, 0, 0), True, crop=False)
net.setInput(blob)
outs = net.forward(output_layers)
# Show detections
class_ids = []
confidences = []
boxes = []
for out in outs:
    for detection in out:
        scores = detection[5:]
        class_id = np.argmax(scores)
        confidence = scores[class_id]
        if confidence > 0.5:
            # Object detected
            center_x = int(detection[0] * width)
            center_y = int(detection[1] * height)
            w = int(detection[2] * width)
            h = int(detection[3] * height)
            x = center_x - w / 2
            y = center_y - h / 2
            boxes.append([x, y, w, h])
            confidences.append(float(confidence))
            class_ids.append(class_id)
# Perform non-maximum suppression
indexes = cv2.dnn.NMSBoxes(boxes, confidences, 0.5, 0.4)
# Draw bounding boxes
for i in range(len(boxes)):
    if i in indexes:
        x, y, w, h = boxes[i]
        label = str(classes[class_ids[i]])
        color = colors[class_ids[i]]
        cv2.rectangle(img, (x, y), (x + w, y + h), color, 2)
        cv2.putText(img, label, (x, y + 30),
                    cv2.FONT_HERSHEY_SIMPLEX, 1, color, 3)
```

```
# Show image
cv2.imshow("Image", img)
```

Output of the program will look like this:



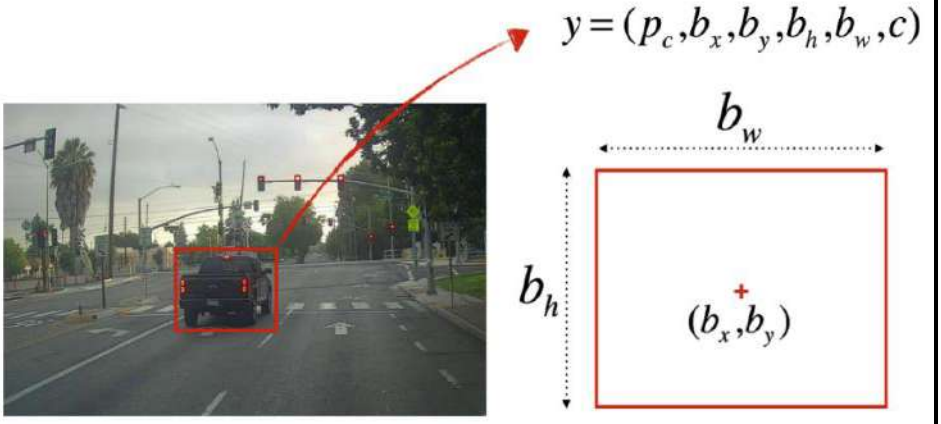
❖ **Application of Statistics and Probability Distribution in Modelling Autonomous Vehicles:**

Statistics plays a central role in deep learning, especially in the fields of computer vision and image processing. Deep learning algorithms, such as those used in YOLO, rely on statistical techniques to analyse and process images and videos, and extract useful information from them.

Probability distributions are a useful tool for modelling and predicting the likelihood of various outcomes in autonomous vehicles. These distributions allow developers to analyse and interpret the data collected by the vehicle's sensors, and to make informed decisions based on that data. For example, a probability distribution is used to model the likelihood of a pedestrian crossing a particular road at a certain time of the day. By analysing data on pedestrian traffic patterns and other relevant factors, the probability distribution is used to predict the likelihood that a pedestrian will be present at a particular location at a particular time. If the probability of a pedestrian being present is high, the autonomous vehicle may choose to slow down or come to a stop to avoid a potential collision.

❖ **Use Of Statistics in YOLO Algorithm:**

YOLO uses statistical techniques, majorly probability distributions. The algorithm divides an image into a grid of cells and assigns each cell a probability of containing an object. Then it uses probability to determine which cells are most likely to contain an object. YOLO also uses the technique of bounding boxes. To calculate the location and size of bounding boxes around each detected object statistical techniques are useful. These bounding boxes are used to enclose and identify the objects in the image.



❖ **Significance of Mahalanobis distance in Autonomous Cars:**

The Mahalanobis distance is a statistical measure of distance between a point and a distribution. It considers the covariance of the data, allowing for a more accurate assessment of the distance between the two.

For autonomous vehicles, the Mahalanobis distance is used to measure the similarity between a detected object and a known object, like a pedestrian or a traffic sign. The likelihood that the detected object is the same as the known object is determined by the autonomous vehicle by calculating Mahalanobis distance between the two objects. This information is very helpful for autonomous cars about how to respond to the detected object. Suppose the Mahalanobis distance between the detected object and a known pedestrian is low. Then the car might slow down and stop to avoid a potential collision.

❖ Conclusion:

Statistics and probability are used in autonomous vehicles to make predictions and decisions based on data. Advantages of using these techniques include improved safety and efficiency, as well as the ability to handle complex and dynamic environments. However, there are also potential disadvantages, such as the risk of errors or biases in the data and the possibility of the autonomous vehicle behaving unexpectedly in certain scenarios. Additionally, there is also a risk of over-reliance on data and automation, which can lead to decreased awareness and decision-making skills among human operators.

In India, several companies and organizations are working on autonomous cars and are testing the vehicles on public roads. The Indian government has approved testing of autonomous cars in several cities, and companies such as Tata Motors and Ola are working on autonomous car projects. It is likely that autonomous cars will continue to advance and become more widespread in the future. However, there are still many challenges to overcome in terms of technology, regulation and public acceptance. It is difficult to predict exactly when and how autonomous cars will be fully deployed and adopted.

❖ References:

1. https://neilnie.com/2018/11/18/implementing-yolo-v3-object-detection-on-the-autonomous-vehicle/
2. https://medium.com/@feiqi9047/the-data-science-behind-self-driving-cars-eb7d0579c80b
3. https://towardsdatascience.com/yolo-object-detection-with-opencv-and-python-21e50ac599e9
4. https://medium.com/@MrBam44/yolo-object-detection-using-opencv-with-python-b6386c3d6fc1
5. https://www.mygreatlearning.com/blog/yolo-object-detection-using-opencv/#:~:text=To%20use%20YOLO%20via%20OpenCV,model%20has%20been%20trained%20on).
6. https://medium.com/analytics-vidhya/object-detection-using-yolov3-d48100de2ebb

THE MONTY HALL PROBLEM

Dishari Datta, 2nd Year

Who would have thought that an old TV game show could inspire a statistical problem that has tripped up mathematicians and statisticians with PhDs? The Monty Hall problem has confused people for decades. In the game show, let us Make a Deal, Monty Hall asks you to guess which closed door a prize is behind. The answer is so puzzling that people often refuse to accept it!

❖ **Understanding the Monty Hall Problem:**

The Monty Hall Problem is a counter-intuitive statistics puzzle:

- There are 3 doors, behind which are two goats and a prize.
- You pick a door (call it door A). You are hoping for the prize of course.
- Monty Hall, the game show host, examines the other doors (B & C) and opens one with a goat. (If both doors have goats, he picks randomly.)

The role of the host is as follows, under standard assumptions:

- The host must always open a door that was not picked by the contestant.
- The host must always open a door to reveal a goat and never the prize.
- The host must always offer the chance to switch between the originally chosen door and the remaining closed door.

Here is the game: Do you stick with door A (original guess) or switch to the unopened door?

❖ **Simple Solution!**

❑ **Understanding The Game Filter:**

Instead of the regular game, imagine this variant:

- There are 100 doors to pick from in the beginning

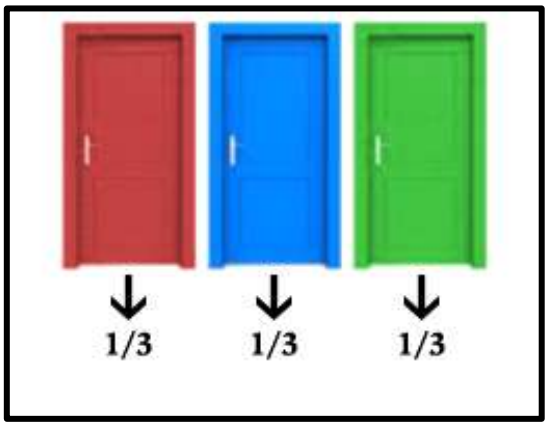
- You pick one door
- Monty looks at the 99 others, finds the goats, and opens all but 1

Do you stick with your original door ($\frac{1}{100}$), or the other door, which was filtered from 99?

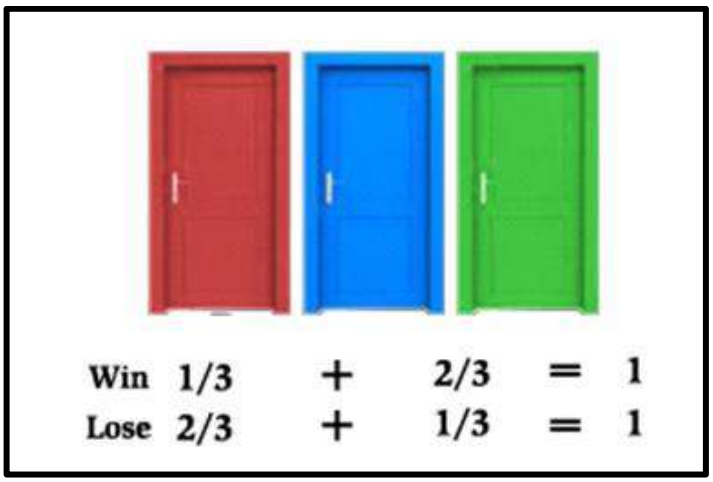
It is a bit clearer: Monty is taking a set of 99 choices and improving them by removing 98 goats. When he has done, he has the 'top' door out of 99 for you to pick.

❑ **Now back to our original problem:**

Let us begin with a simple diagram:

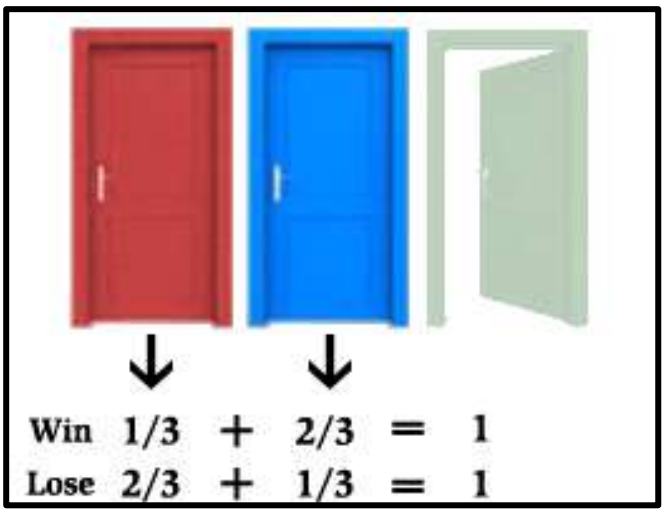


These are the probabilities we face when we are confronted by these three doors: the probability of one door being the door which hides a prize is $\frac{1}{3}$ and the probability that it is not the prize-hiding door is $\frac{2}{3}$. Look at the following representation is a modified diagram.



The probability that any two doors do hide a prize is $\frac{2}{3}$, and the probability that any one door does not hide a prize is $\frac{2}{3}$. They are the same because the sum of the probabilities for individual doors containing a prize must be one.

After you have selected a door, Monty Hall then opens one of the two remaining doors, and **reveals to you that it does not contain a prize**. Once he has done this, we can modify our diagram a bit:



Here you can see that he has eliminated one of the three doors from consideration. Note that he has not just eliminated any random door but has eliminated one of the doors which does not hide a prize.

Case 1: You did select the correct door initially: Since you will be right one-third of the time, that means that if you stay with your first choice, you will get the prize one-third of the time.

Case 2: Your initial choice is incorrect: You are going to choose the wrong door $\frac{2}{3}$ of the time. Your initial choice has only one chance in three of being right – the remaining door has two chances in three.

❖ **Solution using Bayes' Theorem:**

Initially, the prize is equally likely to be behind any of the three doors: the odds-on door 1, door 2, and door 3 are 1:1:1. This remains the case after the player has chosen door 1, by independence. According to Bayes' rule, the posterior odds on the location of the prize, given that the host opens door 3, are equal to the prior odds multiplied by the Bayes factor or likelihood, which is, by definition, the probability of the new piece of information (host opens door 3) under each of the hypotheses considered (location of the prize). Now, since the player initially chose door 1, the chance that the host opens door 3 is 50% if the prize is behind door 1, 100% if the prize is behind door 2, 0% if the prize is behind door 3. Thus, the Bayes factor consists of the ratios $\frac{1}{2}$:1:0 or equivalently 1:2:0, while the prior odds were 1:1:1. Thus, the posterior odds become equal to the Bayes factor 1:2:0. Given that the host opened door 3, the probability that the prize is behind door 3 is zero, and it is twice as likely to be behind door 2 than door 1. Given that the prize is not behind door 1, it is equally likely that it is behind door 2 or 3. Therefore, the chance that the host opens door 3 is 50%. Given that the prize is behind door 1, the chance that the host opens door 3 is also 50%, because, when the host has a choice, either choice is equally likely. Therefore, whether the prize is behind door 1 or not, the chance that the host opens door 3 is 50%. The information "host opens door 3" contributes a Bayes factor or likelihood ratio of 1:1, depending on whether the prize is behind door 1 or not. Initially, the odds against door 1 hiding the car were 2:1. Therefore, the posterior odds against door 1 hiding the car remain the same as the prior odds, 2:1.

❖ **Direct calculation:**

Consider the event C_i, indicating that the prize is behind door number i, takes value X_i, for the choosing of the player, and value H_i, for the host opening the door. The player initially chooses door $i = 1, C = X_1$ and the host opens door $i = 3, C = H_3$.

In this case, we have:

$$P(C_1, X_1) = \frac{1}{2}$$

$$P(C_2, X_1) = 1$$

$$P(C_3, X_1) = 0$$

$$P(C_i) = \frac{1}{3}$$

$$P(C_i, X_i) = P(C_i)P(X_i)$$

$$P(X_1) = \frac{1}{2}$$

$P(X_1) = \frac{1}{2}$ because this expression only depends on X_1, not on any C_i. So, in this particular expression, the choosing of the host does not depend on where the car is, and there are only two remaining doors once X_1 is chosen (for instance, $P(H_1|X_1) = 0)$ and $P(C_i, X_i) = P(C_i)P(X_i)$ because C_i and X_i are independent events (the player does not know where the car is in order to make a choice).

Then, if the player initially selects door 1, and the host opens door 3, we prove that the conditional probability of winning by switching is:

$$P(H_3, X_1) = \frac{2}{3}$$

From the Bayes' rule, we know that $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$. Extending this logic to multiple events, for example A, B and C, we get that we can play with the different subsets of {A, B, C} to calculate the probability of the intersection, as a tool to simplify the calculation of our conditional probability:

$$P(A, B, C) = P(B, C)P(B, C)$$

$$= P(A, C)P(A, C)$$
$$= P(C|A, B)P(A, B)$$
$$= P(A, B|C)P(C)$$
$$= P(A, C|B)P(B)$$
$$= P(B, C|A)P(A)$$

In our case, since we know that $P(H_3|C_2, X_1) = 1$, we are in luck:

$$P(H_3, X_1) = \frac{P(C_2,H_3,X_1)}{P(H_3,X_1)} = \frac{P(H_3|C_2,X_1)P(C_2,X_1)}{P(H_3,X_1)} = \frac{P(C_2)P(X_1)}{P(H_3|X_1)P(X_1)} = \frac{1/3}{1/2} = \frac{2}{3}$$

❖ CONCLUSION:

So, to answer our initial question: **Should we switch? Yes.** Switching doors substantially increases our probability of winning the prize.

❖ REFERENCES:

- https://betterexplained.com/articles/understanding-the-monty-hall-problem/
- https://en.wikipedia.org/wiki/Monty_Hall_problem

IMPORTANCE OF DATA IN FOOTBALL: A PEEK INTO FIFA WORLD CUP 2022

Soham Choudhury, 2nd Year
Souhardya Biswas, 2nd Year

❖ Introduction:

Numbers are everything, aren't they? Any player in any sport can be claimed to be the best among his competitors if he has better statistics than his other competitors. When it comes to football, players are judged based on the number of goals and assists (the final pass by a player that helps to score a goal) they provide during matches. However, as time passed, new methods were introduced for a better understanding of a player's performance with the help of numbers. **Expected goals (xG), Expected assists (xA), Big chances created by the player** are some of the widely used factors to judge the performance of a particular player or a team.

❑ Expected Goals(xG):

Expected goals (xG) measures the quality of a shot using statistical models based on several variables such as *shot angle, assist type, distance from goal, whether the shot was a header, whether it was defined as a big chance,* etc. In simple words, **xG** tells us the probability that a shot is likely to be converted into a goal. The closer the shot is to the goal, the higher the **xG**. The wider the angle from the goal, the lower the **xG**. Some **xG** models are built using a logistic regression model. An **xG** model uses historical information from thousands of shots with similar characteristics to estimate the probability that the shot will result in a goal. For example, a shot taken from the 6-yard box might have a high **xG** value of 0.8 because it is more likely to go into the goal than a shot that is taken from outside of the penalty box, which might have an **xG** value of 0.4. A penalty kick has an **xG** value of 0.76. An **xG** value of 1 means a sure goal, which does not exist. Sometimes the presence of defenders is also considered when calculating **xG**, as the presence of defenders also affects the probability of a shot being converted into a goal. The **xG** value helps us a lot in assessing the qualities of a striker.

Let us consider two players, A and B, who have scored 10 goals each. We may think that they are equally good as they scored the same number of goals. Suppose player A has scored all the goals through **set-pieces** (free-

kicks, corner kicks), and player B has scored just **tap-ins** (a simple shot into the goal from close range, without opposition). Then player A's goal tally will be higher than his **xG** value, knowing that he has scored from difficult positions and tighter angles. Hence, we may conclude that player A is better in terms of scoring goals.

❑ **Expected Assists(xA):**

Expected assists(xA) measures the likelihood that a given pass will become an assist with the help of statistical models, like logistic regression models. It considers several factors such as the *type of pass, the pass endpoint, the length of the pass,* etc. Like the xG model, an **xA** model also uses historical information from thousands of passes in similar conditions to estimate the probability that the pass will convert into an assist. Every pass made by a player is assigned some **xA** value, though most of the time it is low as not all passes made by a player can generate chances. For example, a pass made by a player in the penalty box is more likely to be scored and will have a higher **xA** value, compared to a pass that is made outside the penalty area. The **xA** value is also affected by some other factors, such as the finishing location of the pass and the type of pass. **xA** helps us to understand how creative a player is, regardless of the assists he provides in a measured number of games. In the FIFA World Cup 2022, for example, German youngster Jamal Musiala had an xA value of 1.6 but no assists. This shows he had been let down by his teammates when it came to notching assists.

❑ **Big Chances:**

"Big Chances" in football refers to a situation where a player should reasonably be expected to score goals, usually from a very close range when the ball has a clear path to goal. Providing an opportunity where the receiving player is expected to score is defined as creating a **Big Chance.**

Let us look through an example in order to understand how **xA**, **xG**, **Big Chance created**, etc., provide information that helps in assessing a player's performance.

❖ **Was the Golden Ball Award in FIFA World Cup 2022 justified?**

Articles from Students

Father of modern football Johan Cruyff once said about Leo Messi, "Still, at only 20, he has a little bit to learn about the game: when to provide the right pass and when to dribble. When he fully understands that, he won't just get one Golden Ball (Ballon d'Or), he'll have an entire collection by the time his career ends." His prediction came true, as Messi has won seven Ballon d'Ors to date. Though our discussion is not that, we are focusing on whether Messi's statistics justify him receiving the Golden Ball award at the FIFA World Cup 2022.

To award the Golden Ball, a shortlist is drawn up by the FIFA technical committee, and the winner is voted for by representatives from the media. We do not know what measures FIFA uses to shortlist the players. We feel that **xG, xA, and Big Chance created** are good measures to assess one player's attacking performance. We have collected data on the players' **xG, xA, Goals, Assists, and Big Chance created** by the players in the tournament and plotted them graphically to understand their performance.

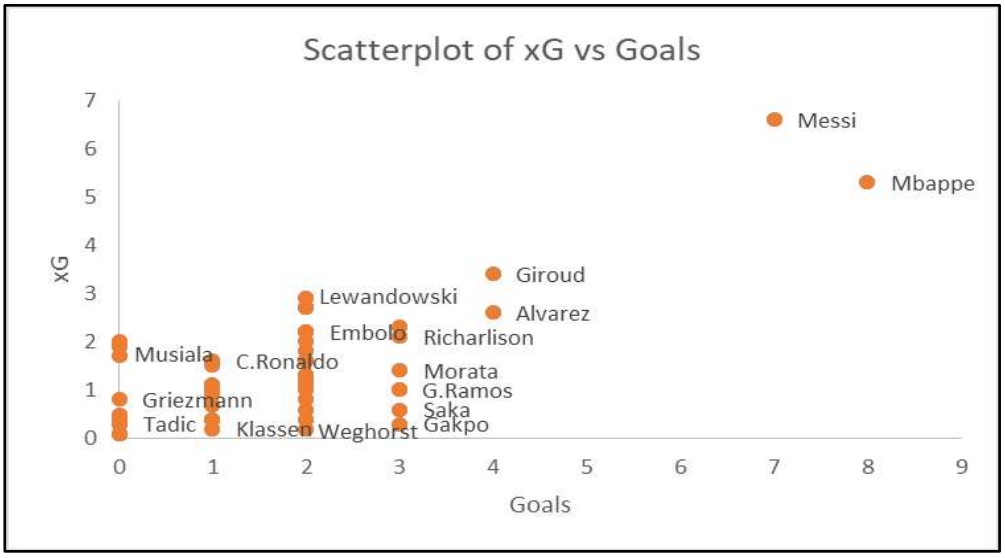


Figure-1: xG vs Goals

From Figure-1, we can see that the **xG** of Lionel Messi is 6.6, but after the tournament came to an end, Messi ended his World Cup goal tally of 2022 with 7 goals. As a result, Messi clearly outperformed his xG. On the other hand, Mbappe had scored a mammoth of 8 goals with only 5.3 **xG**. Despite having an **xG** of 1.9, Musiala had failed to score even a single goal. The Dutch footballer Weghorst scored 2 goals with a very low **xG** of 0.2 and forced the quarterfinal match between Argentina and the Netherlands to go to extra time.

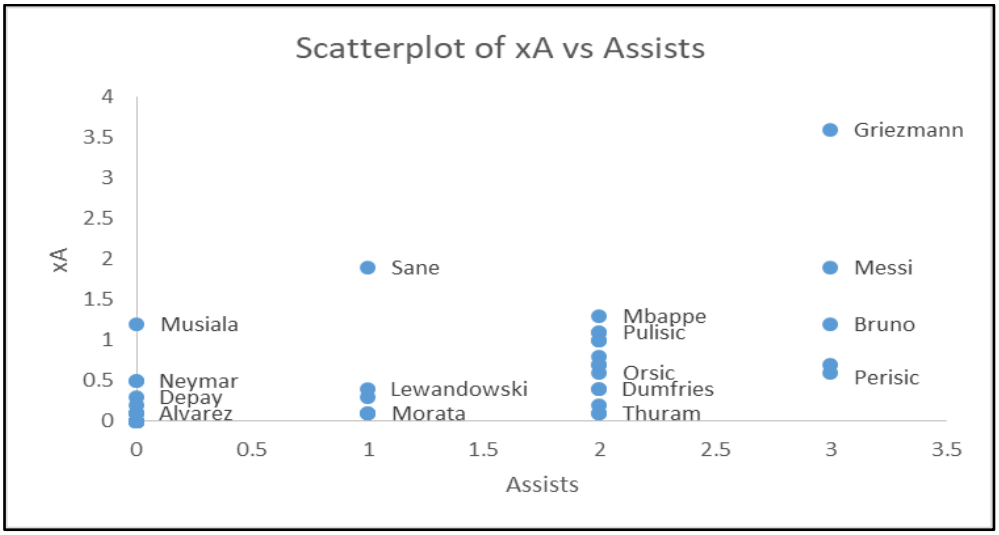


Figure-2: xA vs Assists.

From Figure-2, it can be seen that Perisic had provided 3 assists with just 0.6 **xA**, i.e., he overperformed his **xA**. And the Argentine superstar Lionel Messi, who had a very good tournament ended up with 3 assists with 1.9 **xA**, even though he had provided an outstanding assist with a low **xA** against Netherlands. On the other hand, Griezmann and Musiala provided 3 and 0 assists, respectively, with their corresponding **xA** of 3.6 and 1.6. So, they underperformed their **xA**.

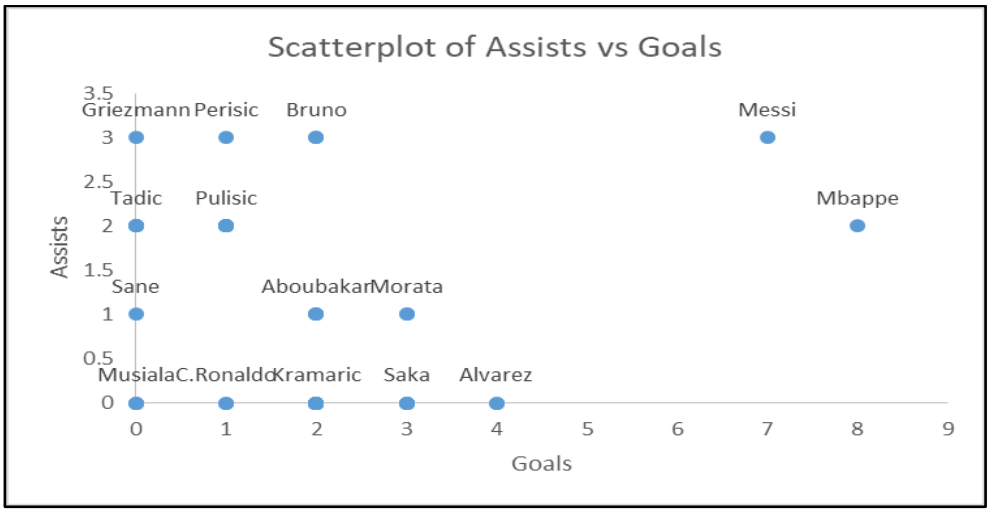


Figure-3: Assists vs Goals.

From Figure-3, it is seen that Mbappe and Messi had 10 G+A (Goals + Assists) which is a great contribution to their respective teams. On the other hand, Kane and Bruno Fernandes had provided 5 G+A. Despite having a good **xG** and **xA**, Musiala ended up with 0 G+A.

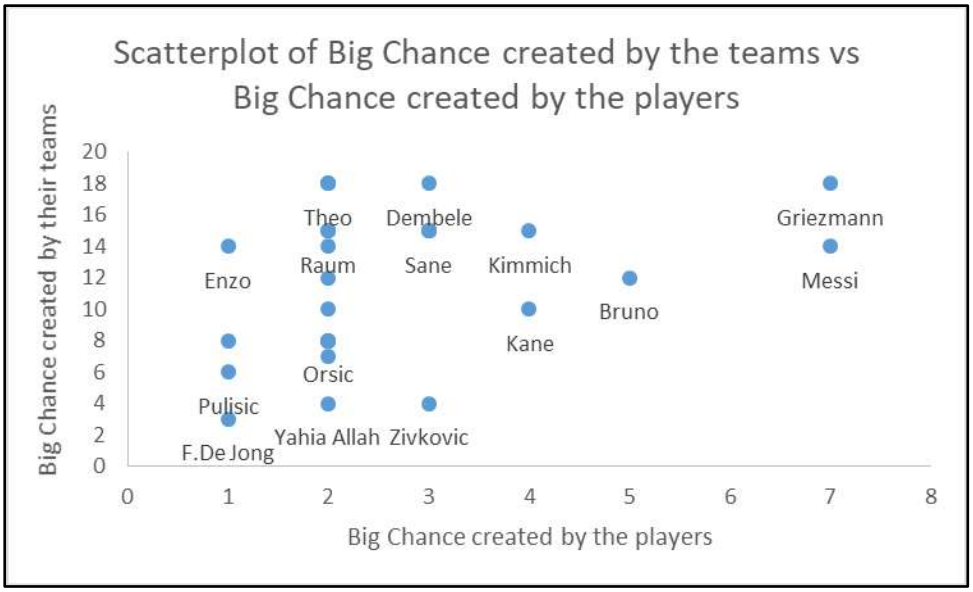


Figure-4: Big Chance created by their team vs Big Chance created by the players.

From Figure-4, Lionel Messi created the 7 big chances out of 14 big chances created by Argentina. And Griezmann and Bruno Fernandes had created 7 and 5 big chances, respectively, out of 18 and 12 big chances created by France and Portugal. Surely it can be said that Messi was the focal point of his team Argentina and played a massive role in Argentina's World Cup victory.

❖ Conclusion:

From the above 4 graphs, we can see that one name is always present in all the discussions of attacking output. He is the Argentine maestro, Lionel Messi. This demonstrates how a 35-year-old man leads his team Argentina to World Cup victory after 36 years of bringing out the best version of himself. This also shows that the FIFA authority did not make any wrong decisions by naming him the best player of the tournament, as the stats of Messi speak for themselves (he is the only player to win the World Cup Golden Ball twice).

Also, he managed to win the Silver Boot and 5 Man of the Match awards in a single edition of the World Cup, which is a record in the history of the FIFA World Cup. Also, this is the first time any player has contributed goals in all knockout games. On the other hand, the French youngsters Mbappe and Griezmann won the Golden Boot (for scoring the highest number of goals) and Silver Ball (for their massive performances in the attacking and defensive transition) respectively. In this World Cup, many superstars like De Bruyne and Cristiano Ronaldo did not fulfil the expectations, whereas players like Perisic, Ounahi, and Mac Allister had an impressive tournament.

❖ **References:**

1. https://www.statsperform.com/opta-event-definitions/
2. https://www.sportskeeda.com/football/what-are-expected-goals-xg-and-expected-assists-xa-why-are-they-a-good-measure-of-player-performances
3. https://en.m.wikipedia.org/wiki/FIFA_World_Cup_awards#:~:text=The%20Golden%20Ball%20award%20is,by%20representatives%20of%20the%20media
4. https://theanalyst.com/eu/2021/07/what-are-expected-goals-xg/
5. https://theanalyst.com/eu/2021/03/what-are-expected-assists-xa/
6. https://play.google.com/store/apps/details?id=com.mobilefootie.wc2010 (FotMob app)
7. https://www.football365.com/news/world-cup-2022-player-stats-golden-boot-shots-xg-key-passes-tackles-dribbles
8. All the graphs are generated using Excel.

JAMES-STEIN ESTIMATOR: A BREAKTHROUGH IN STATISTICAL INFERENCE

Yenisi Das, 3rd Year

The James-Stein estimator, also known through Stein's paradox, is a widely recognised concept in statistics that illustrates the benefits of shrinkage in reducing the mean squared error of a multivariate estimator. The estimator, named after Charles Stein, who first introduced it, is considered a significant contribution to the field of statistics. In 1961, when James and Stein first published it, it came as a huge surprise to the statistics community, which was firmly rooted in the principle of maximum likelihood estimation laid out by R.A. Fisher in 1922, which states that given some data, one should choose the parameters that maximise the probability of observing the data that you actually observed. The James-Stein estimator also performs better than the least-squares estimator. It can be formulated as follows:

Let $X_1, X_2, \dots, X_p$ be independent random variables, such that $X_i \sim N(\theta_i, 1)$ for each $i = 1(1)p$.

Now, we want to estimate the unknown parameters $\theta_i, i = 1(1)p$. Since we have only one sample for each θ_i, an obvious choice of estimator is $\hat{\theta}_i = X_i$ for each $i = 1(1)p$.

When we calculate the mean squared error (risk function) of this estimator, we get:

$$E\left[\|\hat{\theta} - \theta\|^2\right] = \sum_{i=1}^{p} \quad E\left[\left(\hat{\theta}_i - \theta\right)^2\right].$$

A risk function is used to quantify the average error of an estimator, whereas admissibility can be used to compare different estimators when they are used to estimate the same quantity. If Θ is the parameter space, we say that the estimator $\hat{\theta}$ dominates the estimator $\hat{\eta}$ if,

$$MSE\left(\theta, \hat{\theta}\right) \leq MSE(\theta, \hat{\eta})$$

for all $\theta \in \Theta$,

$$MSE\left(\theta_0, \hat{\theta}\right) \leq MSE(\theta_0, \hat{\eta})$$

for some $\theta_0 \in \Theta$. An estimator is admissible if it is not dominated by any other estimator.
It turns out that this estimator is inadmissible when $p \geq 3$. This means that we can find an estimator that always achieves a lower mean squared error irrespective of the value of θ.

In 1961, Stein derived the following explicit form of an estimator that strictly dominates $\hat{\theta}$ in terms of the mean squared error:

$$\hat{\theta}_{Js} = \left(1 - \frac{p-2}{||X||^2}\right)$$

Investigating the James-Stein estimator further, we see that it shrinks the initial estimate (X) towards the origin by multiplying it by a shrinkage factor proportional to the norm of X and the dimension p. It may seem surprising and perhaps paradoxical: given a set of noisy observations with means $\theta_1, \theta_2, \ldots, \theta_P$, we can apparently obtain a better estimate by relocating the observations toward some arbitrary point in the space, in this case the origin, rather than taking the individual observations as estimators of $\theta_1, \theta_2, \ldots, \theta_P$.

❖ **Bias-Variance Trade-off**

To make this estimator intuitive to understand, it is important to factor in the combined mean squared errors of all $\theta_i's$ i.e., $\sum_{i=1}^{p} \quad E\left[\left(\hat{\theta}_i - \theta_i\right)^2\right]$ when judging the quality of the estimator. No shrinkage estimator would be able to uniformly dominate $\hat{\theta}_i = X_i$ if we judged the quality of the estimator using the mean squared errors of each $\theta_i's$. However, since we focus on the mean squared error across all $\theta_i's$, it turns out we can do slightly better by reducing the variance of the estimator at the added cost of some bias. Stein's paradox is a great demonstration of how removing the "unbiased" condition allows one to achieve better estimators in terms of mean squared error. The mean squared error can be decomposed into, one, a squared bias term and, two, a variance term which can be shown using the linearity of the expectation:

$$\sum_{i=1}^{p} \quad E\left[\left(\hat{\theta}_i - \theta\right)^2\right] = \sum_{i=1}^{p} \quad \left(E\left(\hat{\theta}_i\right) - \theta\right)^2 + \sum_{i=1}^{p} \quad E\left[\left(\hat{\theta}_i - E\left(\hat{\theta}_i\right)\right)^2\right]$$

$$= p$$

The estimator $\hat{\theta}_i = X_i$ is unbiased so the first term is 0, and the second term is equal to p due to our assumption that $Var(X_i) = 1$ for each i. Now, we define a general shrinkage estimator of the form $\hat{\theta}_\beta = \beta X, \beta \in R$. We can find the MSE of this estimator as follows:

$$\sum_{i=1}^{p} E\left[\left(\hat{\theta}_{\beta,i} - \theta_i\right)^2\right] = (\beta - 1)^2 \|x\|^2 + \beta^2 p$$

The first term is the square of the bias, and the next term is the variance. We see that for any given β, the variance term only depends on the dimension p, and the bias term only depends on the norm (i.e., size) of θ. Note that on one end, if $\beta = 1$, we get back our original estimator $\hat{\theta}_1 = X$ which has 0 bias but maximum variance. On the other end $\hat{\theta}_0 = 0$ has 0 variance but arbitrarily large bias.

For $\beta = 1 - \frac{p-2}{\|x\|^2}$ i.e., the shrinkage factor of the James-Stein estimator,

$$MSE = E\left\{\|X - \theta - \frac{(p-2)\cdot X}{\|X\|^2}\|^2\right\}$$

$$= E\left|\|X - \theta\|\right|^2 + E\|\frac{(p-2)\cdot X}{\|X\|^2}\|^2$$

$$- E\|2(p-2)\sum \frac{(X_i - \theta_i)\, X_i}{\|X\|^2}\|^2$$

$$= p - (p-2)^2 E\left[\frac{1}{\|x\|^2}\right] + 2(p-2)\sum_{i=1}^{p} E\left[\frac{(X_i - \theta_i)X_i}{\|X\|^2}\right]$$

Now the last term for i=1 can be calculated as:

$$E\left[\frac{(x_1 - \theta_1)x_1}{\|x\|^2}\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{(x_1 - \theta_1)x_1}{\|x\|^2} \cdot \frac{e^{-\frac{\|x-\theta\|^2}{2}}}{(2\pi)^{\frac{p}{2}}} dx_1 \ldots dx_p$$

Using integration by parts, we get $\int_{-\infty}^{\infty} \cdots \left(\left[-\frac{x_1}{\|x\|^2} e^{-\frac{\|x-\theta\|^2}{2}} \right]_{\infty}^{-\infty} + \int_{-\infty}^{\infty} e^{-\frac{\|x-\theta\|^2}{2}} d\left(\frac{x_1}{\|x\|^2}\right) \right) \dots dx_p = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\|x\|^2 - 2x_1^2}{\|x\|^4} \cdot \frac{e^{-\frac{\|x-\theta\|^2}{2}}}{(2\pi)^{\frac{p}{2}}} dx_1 \dots dx_p = E\left[\frac{\|x\|^2 - 2x_1^2}{\|x\|^4}\right]$

Therefore $MSE = p - (p-2)^2 E\left(\frac{1}{\|x\|^2}\right) - 2(p-2)E\left[\sum_{i=1}^{p} \frac{\|x\|^2 - 2x_i^2}{\|x\|^4}\right]$

$$= p + (p-2)^2 E\left(\frac{1}{\|x\|^2}\right) - 2(p-2) \cdot (p-2)E\left(\frac{1}{\|x\|^2}\right)$$

$$= p - (p-2)^2 E\left(\frac{1}{\|x\|^2}\right)$$

Observe that $E\left(\frac{1}{\|x\|^2}\right)$ does not converge for 1 or 2 dimensions. However, in 3 or higher dimensions after transforming the integral into polar coordinates we get the following integral which converges since the r^2 term in the denominator cancels out. Here f denotes the pdf of the multivariate Normal distribution.

$E\left[\frac{1}{\|x\|^2}\right] = \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} \frac{1}{r^2} f(r,\theta,\varphi) r^2 \sin\sin\theta \, dr \, d\theta \, d\varphi = \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} f(r,\theta,\varphi) \sin\sin\theta \, dr \, d\theta \, d\varphi$

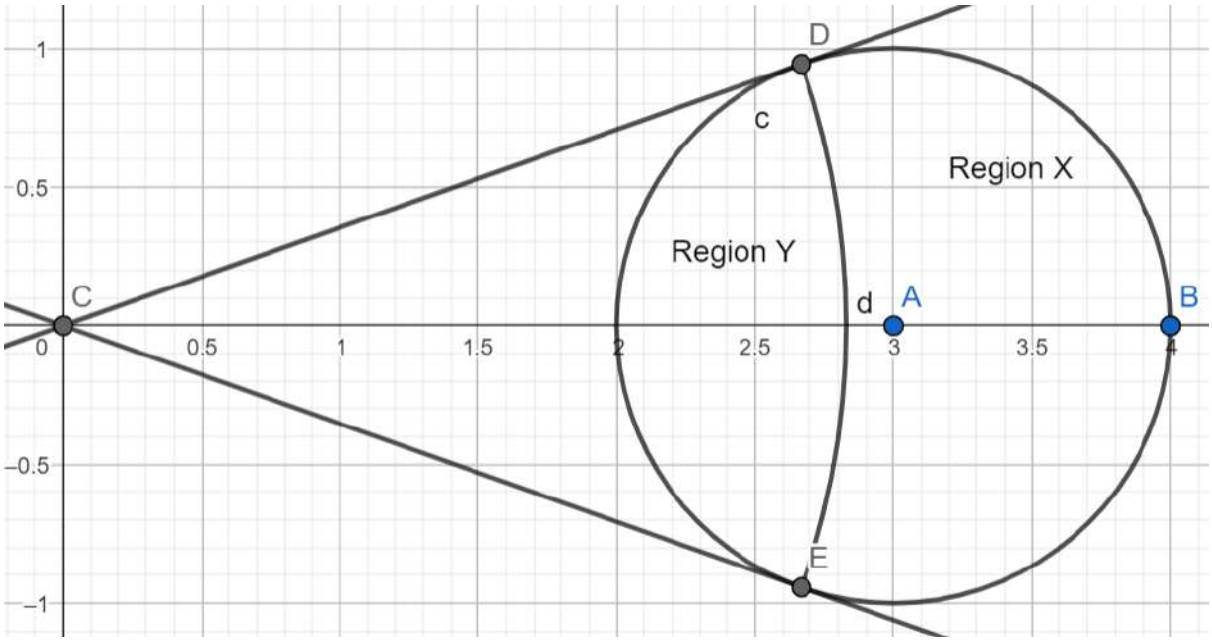
In general, the Jacobian for p dimensions is given by $|J| = r^{p-1}\theta_1 \dots \sin\sin\theta_{P-2}$ and hence for p greater than or equal to 3 the r^2 the term in the denominator cancels out which makes the integral converge. Since $p > p - (p-2)^2 E\left(\frac{1}{\|x\|^2}\right)$, $\hat{\theta}_{Js}$ *is a better estimator than* $\hat{\theta}$.

❖ **Visually intuitive explanation:**

Assume there is a circle on the XY plane, and we must guess the location of the circle's centre based on the location of a randomly selected point within the circle. Let's call this position P, and we'll call the true, unknown centre of the circle A.

Let's suppose the circle has radius 1 and its true centre is at $(3,0)$. For simplicity, let the point we shrink towards be the origin (C). Now, if I randomly draw a point from this circle, we want to find what proportion of the time my guess will actually be better if I shrink it slightly towards the point C, i.e., what proportion of the circle could get closer to the centre if it were shrunk towards C a little.

The answer to that question is region X on the figure. Using geometry, we can say it's about 61% of the circle. In the actual scenario, i.e., in the case of a bivariate normal distribution, 39% of the circle that moves farther from the mean collectively moves more far away than that distance moved closer by the other 61%, and that is why the James-Stein estimator works only in dimension 3 or higher.



Considering the problem in three dimensions, we now have a sphere rather than a circle, but everything else remains the same. In three dimensions, the region X covers just over 79% of the sphere, so about four times out of five, our shrinkage estimator does better than estimating the point A. Region X covers 87% of the sphere in four dimensions, about 98% of the sphere in ten dimensions, and 99.99999% in one hundred dimensions.

There is a strong connection between random walks and the admissibility of the naive estimator $\hat{\theta}_i = X_i$, discovered by Lawrence Brown. Brown found that the naive estimator is considered admissible only if a random walk

returns to the origin an infinite number of times. In one or two dimensions, random walks do return to the origin, but in three or more dimensions, they do not. This is because as the dimensionality of the space increases, the random walk drifts away from the origin at a linear rate, while the volume subtended by the origin at a fixed distance decreases exponentially. In one or two dimensions, the volume subtended by the origin is large enough that it is likely to eventually return to it, but in higher dimensions, the volume is so small that the probability of returning to the origin approaches zero. It is difficult to demonstrate this correspondence, but an intuitive method for solving both problems involves comparing the distance between a point and the origin and the volume of the unit sphere in space.

❖ **Some applications**

The best thing about being a statistician is that you get to play in everyone's backyard.

-- John Tukey

- Modern machine learning algorithms, such as ridge and Lasso regression, are underpinned by these ideas of shrinkage. The idea of shrinkage (sometimes to even zero) is known as regularization. James Stein estimator is not admissible either, there are some estimators that dominate it, like the positive part James-Stein estimator.
- Many modern statistical models may involve thousands or even millions of parameters (e.g., in microarray experiments in genetics, or fMRI studies in neuroimaging); in such circumstances, we would almost certainly want estimators to set some of the parameters to zero, not only to improve performance but also to ensure the interpretability of the fitted model which is the main idea behind the estimator to shrink.
- The extensions of the James-Stein estimator have been applied to adaptive statistical signal processing problems. The James–Stein state filter (JSSF), which is a robust version of the Kalman filter, has been derived.
- Using the James-Stein estimators and Pinsker's theorem, a novel deep learning architecture that combines nonparametric regression for feature extraction with a deep neural network has been used for classification in the problem of decoding eye movement intentions from Local Field Potentials collected in macaque cortex at various cortical depths.

❖ **References:**

Articles from Students

1. James, W.; Stein, C. (1961), "Estimation with quadratic loss", *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, vol. 1, pp. 361–379
2. https://joe-antognini.github.io/machine-learning/steins-paradox, last accessed on 12.01.2023.
3. Lehmann, E. L.; Casella, G. (1998), *Theory of Point Estimation* (2nd ed.), Springer
4. C. Stein, Inadmissibility of the usual estimator of the mean of a multivariate normal distribution, Third Berkeley Symposium, 197–206, Univ. California Press (1956)
5. Samworth, RJ; Stein's Paradox, Statslab Cambridge

BAYES' THEOREM: THE TOOL FOR PREDICTING THE FUTURE

Swagata Kar, 2nd Year

❖ Introduction:

Suppose you are not feeling well for a few days but don't have any particular symptoms. The next day you decide to visit the doctor, and they suggest a series of tests. After some days when you get all the reports, it turns out that you have tested positive for a rare disease. To learn more about it, you go through Google and find that it is a serious one that affects only about 2% of the population and has negative consequences. When you ask your doctor – "how certain is it that I have this disease?", they say – "the test is 90% sensitive (true positive rate of the test) and also correctly identifies 85% of people who are not affected by the disease (true negative rate)". This means the chance that you have the disease is 90%. This sounds pretty bad, right? However, that is not correct! You need Bayes' Theorem to get some mathematical perspective.

❖ Overview:

❑ Changing beliefs on Bayes' Theorem:

At the core of Bayes' rule is the idea that we update our beliefs every time we get new evidence. The Bayesian school of thought allows us to adapt our thinking reactively to new evidence as it arises and adjust our action to optimize the odds of success as the probabilities evolve in real-time.

The probabilities involved in the theorem may have different probability interpretations. With Bayesian probability interpretation, the theorem expresses how a degree of belief, expressed as a probability, should rationally change to account for the availability of related evidence.

❖ Bayes' Theorem:

Bayes' theorem (alternatively known as Bayes' Law), named after Thomas Bayes, describes the probability of an event based on prior knowledge of conditions that might be related to the event. Expressed below is the mathematical equation that illustrates Bayes' theorem:

For any two events A and B,

$$P\ (B) = \frac{P(A) * \ P(A)}{P(B)}$$

where P(B) ≠ 0.

- P (A | B) is the conditional probability of occurrence of event A given that B is true. It is also called the posterior probability of A given B.
- P (B | A) is the conditional probability of occurrence of event B given that A is true. It can also be interpreted as the likelihood of A given a fixed B.
- P(A) and P(B) are the marginal probabilities of observing events A and B respectively.

We can write,

$P(B) = P(A) * P(A) + P(A^C) * P(A^C)$, where A^C is the complement of the event A.

Thus, using this equation, Bayes' theorem can be derived as:

$$P(B) = \frac{P(A) * P(A)}{P(B)} = \frac{P(A) * P(A)}{P(A) * P(A) + P(A^C) * P(A^C)}$$

❖ **Diagnostic Test Scenario:**

Let A be the event that a person is affected by a disease and B be the event that the result of the test of that particular person is positive. Let,

- P (A): Probability that the person is affected by the disease.
- P (B): Probability that the test result is positive.
- P (A | B): Probability that the person is affected when the test result is positive.
- P (B | A): Probability that the test result is positive when the person is affected.

By Bayes' Theorem we can write,

$$P\ (B) = \frac{P(A) * P(A)}{P(B)}$$

P (Disease=True | Test=Positive)
$= \frac{P(Test=Positive \mid Disease=True) * P(Disease=True)}{P(Test=Positive)}$

As discussed earlier, the accuracy of the medical diagnostic test is not perfect; they have errors.
Sometimes a patient will have the disease, but the test will not detect it. This capability of the test to detect the disease is referred to as the **sensitivity** or the true positive rate.
In this case, the test is good, but not great, with a true positive rate or sensitivity of 90% i.e., of all the people who have the disease and are tested, 90% of them will get a positive result from the test.

P (Test = Positive | Disease = True) = 0.90

Given this information, our intuition would suggest that there is a 90% probability that the patient has the disease.
But our intuitions of probability are wrong!
This type of error in interpreting the probabilities is called the **Base Rate Fallacy**.

We have already assumed that the probability of having this disease is very low and the base value is 2%.
P (Disease = True) = 0.02

Thus, we can correctly calculate the probability of a patient having the disease given a positive test result using Bayes' Theorem.

P (Disease=True | Test=Positive) $= \frac{0.90 * 0.02}{P(Test=Positive)}$
To find the value of P(B) i.e., P (Test=Positive), we know that,

$$P(B) = P(A) * P(A) + P(A^C) * P(A^C)$$

i.e.,

P (Test = Positive) = P (Test = Positive | Disease = True) * P (Disease = True) +

P (Test = Positive | Disease = False) * P (Disease = False)

First, we will calculate P (Disease = False) as the complement of P (Disease = True), which we already know.

P (Disease =False) = 1 – P(Disease=True) = 1 – 0.02= 0.98

We still do not know the probability of a positive test result given no disease. To calculate this, we should know the probability of getting a negative result (Test=Negative) when the patient does not have the disease (Disease = False). It is called the true negative rate or the **specificity**.

P (Test = Negative | Disease = False) = 0.85

P (Test = Positive | Disease = False) = 1 – P (Test = Negative | Disease = False)

= 1 – 0.85
= 0.15

We can plug this false positive value into our calculation of P(Test=Positive) as follows:

P(Test=Positive) = 0.90 * 0.02 + 0.15 * 0.98
= 0.018 + 0.147
= 0.165

Here is the table of data:

Test / **Disease**	Negative	Positive
True	0.10	0.90
False	0.85	0.15

Now we will estimate the probability of a randomly selected person having the disease if they get a positive test result.

P (Disease=True | Test=Positive)

$$= \frac{P(Test=Positive \mid Disease=True) * P(Disease=True)}{P(Test=Positive)}$$

$$= \frac{0.90 * 0.02}{0.165}$$

$$= 0.109090909$$

$$= 0.11 \text{ (approximately)}$$

Thus, the above calculation suggests that if the patient is tested and the result is positive, there is only an 11% chance that they have the disease.

It is a terrible diagnostic test!

Thus, using Bayes' Theorem we get a real-life perspective of the chances of having the disease.

Let us discuss practical implementation and some real-life examples.

❖ **Medical Test Results and Bayes' Theorem:**

Nowadays, it is becoming a trend that doctors are prescribing various types of tests, especially Full Body Checkup(s) which include 30 – 40 different kinds of tests altogether.

To get the test report, we need to count and calculate the number of blood cells and compounds in the sample taken from the user. We can use this data to predict the possibilities of upcoming problems (due to deficiency or surplus of elements) using Bayes' Theorem.
Here is an example:

Test Name	Value	Unit	Bio Ref. Interval
Vitamin B-12	156.47	pg / mL	120 - 914

Comments:

Increased Levels:

- Renal Failure
- Liver Disease

- Myeloproliferative Disorder.

Decreased Levels:

- Pernicious Anemia
- Megaloblastic Anemia
- Iron Deficiency
- Dizziness

After getting the count of the components from the sample, we need to consider the count in relation to the standard bio-reference interval. There can be two cases –

1) The count doesn't lie within the standard interval.

I. The count lies below the lower limit of the interval.
II. The count lies above the upper limit of the interval.

Here I. causes the diseases mentioned as "decreased levels" and II. causes the diseases mentioned as "increased levels" in the above table respectively.

2) The count lies within the standard interval.

I. The count lies more or less in the middle of the interval.
II. The count lies near the lower limit of the interval.
III. The count lies near the upper limit of the interval.

Here I. denotes that the test result is good, and the person need not to worry about that particular element. II. and III. denote the possibilities that the person can be affected by the "decreased level" diseases and "increased level" diseases respectively in near future. This should be taken care of.

❖ **Use of Bayes' Theorem:**

Suppose someone's Vitamin B12 level is 156.47 pg/mL (as mentioned in the above table) which is near the lower limit of the given interval. This means that the person has a high possibility of having "decreased levels" of diseases in the near future. To know the chances of being affected by one of the diseases (e.g., Anaemia), we use Bayes' Theorem.

We can write,

$$P\ (B) = \frac{P(A) * P(A)}{P(B)}$$

where A is the event that the person can be affected by the disease and B denotes the amount of Vitamin B12 in that person's body.
P (A | B) is the probability that the person can be affected by the disease when the amount of Vitamin B12 in their body is at 156.47 pg/mL.
P (B | A) is the probability that the amount of Vitamin B12 in the person's body is 156.47 pg/mL when they have the chance of being affected by the disease in the future.

From the previous dataset, we can get the above-required values for 155-160 pg/mL level of Vitamin B12 in patients' bodies who got affected and also who didn't get affected by the disease in the future.
Putting these values in Bayes' Theorem, we calculate the chances of having the disease in the future which gives the patient more reliable information.

❖ **Conclusion:**

Bayes' theorem is a framework for decision-making based not on predictions or opinions but on evolving statistical analysis of incoming data filtered through an existing model. Applying Bayes' Theorem on prior knowledge of conditions or previously stored data sets, which is related to the event, can give us desired results to make proper decisions.
We cannot deny the uncertainty of the upcoming future, but using Bayes' Theorem along with other statistical tools we can surely have a brief idea of future predictions which will help mankind to lower the risk factors while making decisions and taking precautionary measures.

❖ **Future Scope:**

In intricate situations, the utilization of Artificial Intelligence in decision-making will reduce the chances of human error while making strategic planning more effective with higher efficiency.

❖ **Bibliography:**

Articles from Students

1. Bayes' Theorem – Wikipedia
2. "Bayes' Theorem, the Geometry of Changing Beliefs" – Steve Burns.
3. "How to update your beliefs systematically – Bayes' Theorem" – Veritasium.
4. "A Gentle Introduction to Bayes' Theorem for Machine Learning" – Jason Brownlee.

WATER STATISTICS AND STATISTICAL METHODS IN WATER CONSERVATION

Shreya Saha, 2nd Year
Pritam Saha, 2nd Year
Ahan Prodhan, 2nd Year

Water is all around us, isn't it? Rivers, Lakes, the Sea you swim in on holidays or live nearby. There seems to be an endless supply. So, why do people keep telling us we need to save water as much as possible?

According to SES Water, 1 in 4 people admit to taking water availability for granted, with no idea how much water they use daily. 67% feel they can't use less water than they already do. An average household uses 350 litres of water a day, yet people estimate they use only 157 litres a day.

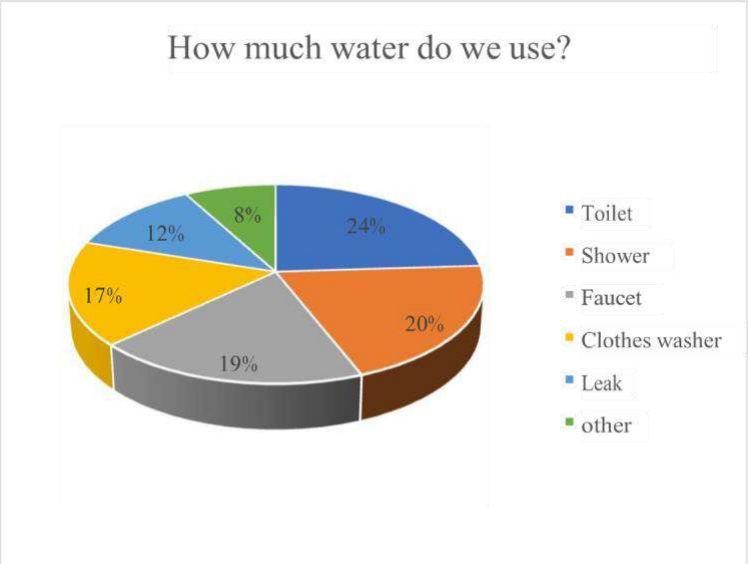
The problem is, we take things for granted. To be completely honest, I did not realize the exact extent of this issue until I started my research on what is mentioned below and it certainly opened my eyes. The information has led me to look at my own habits when using water.

We all know water is an essential part of our lives. We need it daily to live and perform vital functions of life. Plants need water to grow and in turn, we get to eat the plant or the organisms which eat the plants. It is an extremely important substance.

- The average family can waste 180 gallons per week or, 9,400 gallons of water annually, from household leaks.
- Household leaks can waste approximately nearly 900 billion gallons of water annually nationwide. That's equal to the annual household water use of nearly 11 million homes.
- Running the dishwasher only when it's full can eliminate one load of dishes per creek and can save the average family use to nearly 320 gallons of water annually.
- Turning off the tap while brushing your teeth can save 8 gallons of water, while shaving can save 10 gallons of water per shave. Assuming you brush your teeth twice daily and shave five times per week, you could save 5,700 gallons per year.

Articles from Students

- Letting your faucet run for five minutes while washing dishes can waste 10 gallons of water that can produce enough energy to power a 60-watt light bulb for 18 hours.
- On average, outdoor water use accounts for more than 30 percent of total household water use but can go up to 60 percent of total household water use in arid regions.
- As much as 50 percent of the water we use outdoors, is lost due to wind, evaporation and runoff caused by inefficient irrigation methods and systems. A household with an automatic landscape irrigation system that isn't properly maintained and operated, can waste up to 25000 gallons of water annually.



From the above pie chart, we can show that we use significantly more water in different working sectors than we need.

So, here I am going to discuss how we can save water or use water in the right measurements so that we can waste water as little as possible.

Now the question is how statistics can help in water conservation. Basically, in water conservation, everything is related to data, and data means statistics. Initially, through surveys, we can collect broad data and based on that we can adopt different ways and technologies to conserve water.

Articles from Students

- **Surveys:**

  Through organizing surveys, we can collect a broad amount of data. We take India as our area of study. So, firstly we can organize surveys locally, and many local surveys together result in a broad survey. Through these surveys, we can get a huge amount of data. Summarizing these data, we can get the amount of water needed per day. Now, the government should assign a certain limit to the amount of water a person could use daily. I know, in India it is not easy to implement, but to reserve fresh water for our future generations, it's high time we should think about it.

  For example, let an average Indian need 100 Litres of water per day and in a house, there are 4 members. So, they get 400 litres of water per day. If they need more than that they can buy using money. And with that money the government can install necessary technologies.

- **Spread Awareness:**

  There is a well-known method called the interview method in data collection, where our interviewers talk to the Indian households and try to know about their habits and how much water they waste. Using this data, we can make them aware of their bad habits and how they can save more water.

- **Some ways how we can spread awareness –**

  • Check your toilet for leaks. (Leaking can waste more than one hundred gallons of water a day.)

  • Stop using your toilet as an astray or wastebasket. (Every cigarette butt or tissue you flush away also flushes away five to seven gallons of water.)

  • Take Shorter Showers - (A typical shower uses five to ten gallons of water a minute. Limit your showers to the time it takes to soap up, wash down and rinse off.)

  • Take a bath in tubs. (A partially filled tub uses less water than all but the shortest showers.)

• Turn off the water faucet while brushing your teeth. (Before brushing wet your brush and fill a glass for rinsing your mouth.)

• Turn off the water faucet while shaving. (Fill the bottom of the sink with a few inches of warm water to rinse your razor.)

• Check faucets and pipes for leaks. (Even a small drop can waste 50 or more gallons of water a day.)

• Use your automatic dishwasher and washing machine only for full loads.

• Tell your children not to play with the hose and sprinklers. (Children love to play under a hose or sprinkler on a hot day. Unfortunately, this practice leads to the wastage of precious water and should be discouraged.)

• Don't run the hose while washing your car. (Soap down your car using a pail of soapy water. Use a hose only to rinse it off.)

And many more.

- **Water Saving Technologies –**

In the USA, many water-saving technologies are developed. But in India, we are still following those typical old techniques. Some water-saving technologies need to be developed in India as soon as possible.

• WaterSense Labelled Irrigation controllers. (WaterSense labels weather-based irrigation controllers, a type of "smart" irrigation control technology that uses local weather data to determine when and how much to water. WaterSense Labelled irrigation controllers can save you water, time, and money when compared to standard models.)

• Soil Moisture Sensors. (Soil moisture-based control technologies water plants based on their needs by measuring the amount of moisture in the soil and tailoring the irrigation schedule accordingly.)

• Rain Sensors. (Rain Sensors can help decrease water wasted in the landscape by turning off the irrigation system when it is raining.)

• Rainfall Shut off devices. (Rainfall shut-off devices turn off your system in rainy weather and help compensate for the same. This inexpensive device can be retrofitted to almost any system.)

• Sprinkler Heads. (Certain types of sprinkler heads apply water more efficiently than others. Rotary spray heads deliver water in a thicker stream than mist spray heads.)

- **What issues are there with water availability?**

• When rain falls, 50% goes back to the atmosphere through evaporation or gets used by plants, a process known as evapotranspiration. The rest (effective rainfall) is available as surface water or groundwater.

• Climate change. (Climate change is at critical levels. We've noticed the changes in weather ourselves over the years. We are getting dryer summers which can lead to droughts. There are also wetter monsoons that bring about the potential for flooding.)

• Population growth. (The population is growing and with a growing population comes more demand for resources. The rate of abstraction of water is already at levels which are not sustainable. A growing population demanding more water in an instant, will only worsen this.)

• Energy generation. (Generating energy is a major use of water. Not only this, if we use more water than needed (i.e., we waste water) there is a need for heating water that isn't required. This not only increases the bills but creates more carbon to produce the energy required to heat the water. The increasing population means an increase in energy demands. This can be offset by using more eco-friendly methods of energy production, but again is not a quick-fix option.)

• Changes in lands, rivers and wetlands use. (The new generation will need to live and work somewhere. Therefore, ground that was available to abstract water from is being covered with bricks, mortar and tarmac.)

• Water in our lakes, wetlands and watercourses protects the environment. This means that increasing the abstraction of water would have a negative effect.)

Articles from Students

Nowadays water conservation is an important issue. Proper Water Management is important for several reasons. Some of those reasons include:

• Water is a resource. The current water supply on earth comes from surface water, groundwater, and snow. This supply comes from the same sources that have been used for thousands of years, which are now being threatened by overuse, pollution, and global warming. Only three percent of the earth's water supply is made up of fresh water, with only half a percent of that available for human consumption.

• Conservation alleviates droughts. Dry areas like deserts experience drought regularly, in which the rainfall and snowfall aren't adequate and might cause water shortages. Conserving water can help alleviate the effects of water shortage in any given community.

• Using water drains other resources. Using in-house water resources requires energy to deliver the water to your home. This energy use increases when you use hot water since a lot of energy goes into heating. Reducing the use of hot and cold water can help conserve both water and energy, cutting down on energy pollution which harms the environment.

References:

1. epa.gov/watersense/statistics-and-facts
2. intelligenthanddryers.com
3. masterclass.com/articles/
4. https://blog.ferrovial.com/en

STSA 2022-2025

STSA 2021-2024

STSA 2020-2023

MDTS 2022-2024

Xavier Rozario
Convenor,
ED ‘23

Srijan Sen
Editor-in-Chief,
Prakarsho Vol XV

Rachit Yadav
Co-Convenor,
ED ‘23

Sneha Maheshwari
Associate Editor-in-Chief,
Prakarsho Vol XV

Yuvraj Dutta
Cultural Head,
ED ‘23

Mehuli Bhandari
Events Head,
ED ‘23

Anuroop Roy
Finance Head,
ED ‘23

Shamie Dasgupta
Designing Head,
ED ‘23

Art, Cover & Illustrations by
Arka Roy