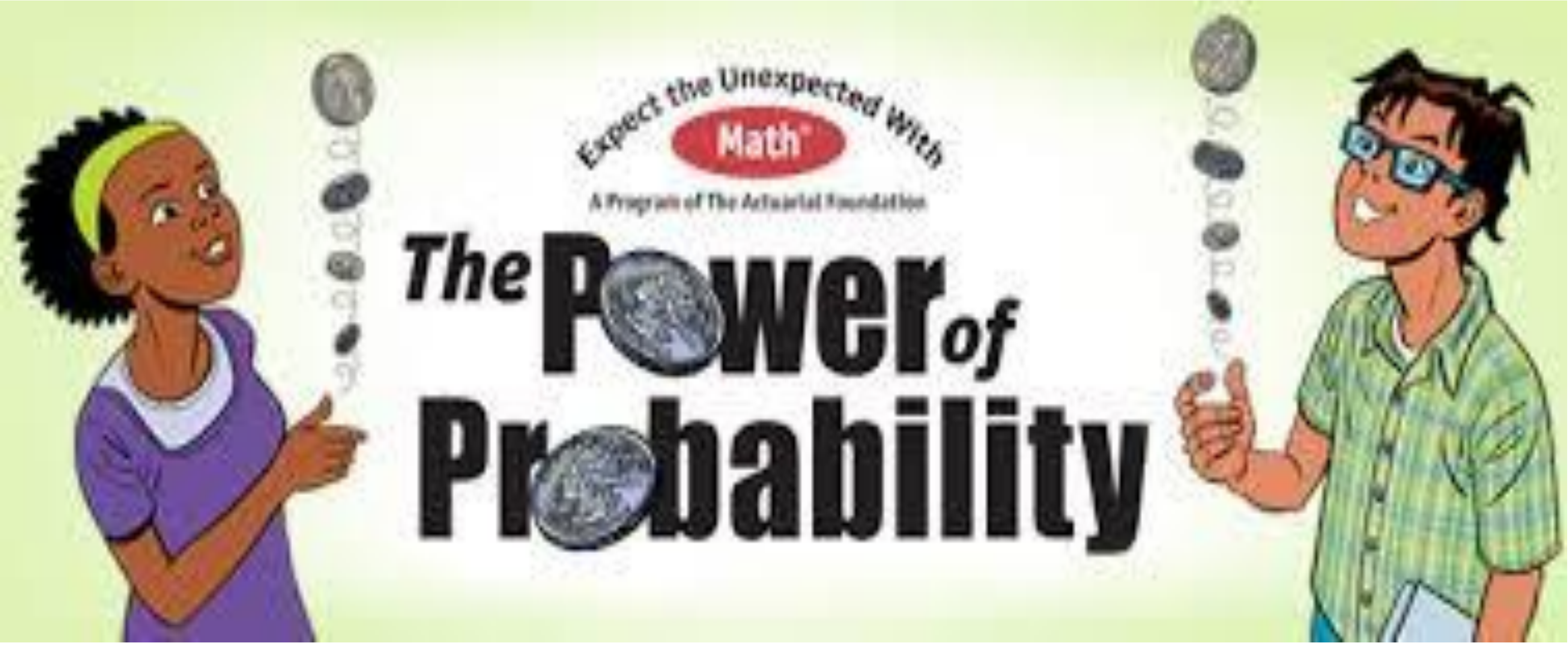
Probability theory
The foundation of Statistics

- Dr. Sarita Azad
Associate Prof.
School of Mathematical and Statistical Sciences.

Probability

How likely something is to happen.

Many events can't be predicted with total certainty. The best we can say is how likely they are to happen, using the idea of probability.



Basic Idea

- Future: what is the likelihood that a student will get a CS job given his grades?
- Current: what is the likelihood that a person has cancer given his symptoms?
- Past: what is the likelihood that Marilyn Monroe committed suicide?
- Combining evidence.
- Always: Representation & Inference

History

- Games of chance: 300 BC
- 1565: first formalizations
- 1654: Fermat & Pascal, conditional probability
- Reverend Bayes: 1750's
- 1950: Kolmogorov: axiomatic approach
- Objectivists vs subjectivists
  - (frequentists vs Bayesians)
- Frequentist build one model
- Bayesians use all possible models, with priors

Tossing a Coin

When a coin is tossed, there are two possible outcomes:

heads (H) or
tails (T)

We say that the probability of the coin landing H is ½.

And the probability of the coin landing T is ½.

Throwing Dice

When a single die is thrown, there are

six possible outcomes: 1, 2, 3, 4, 5, 6.

The probability of any one of them is 1/6.

Probability

In general:

$$\text{Probability of an event happening} = \frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}}$$

Example: **the chances of rolling a "4" with a die**

Number of ways it can happen: 1 (there is only 1 face with a "4" on it)

Total number of outcomes: 6 (there are 6 faces altogether)

So the probability = 1/6

Example: **there are 5 marbles in a bag: 4 are blue, and 1 is red. What is the probability that a blue marble will be picked?**

Number of ways it can happen: 4 (there are 4 blues)

Total number of outcomes: 5 (there are 5 marbles in total)

So the probability = 4/5

Probability is Just a Guide

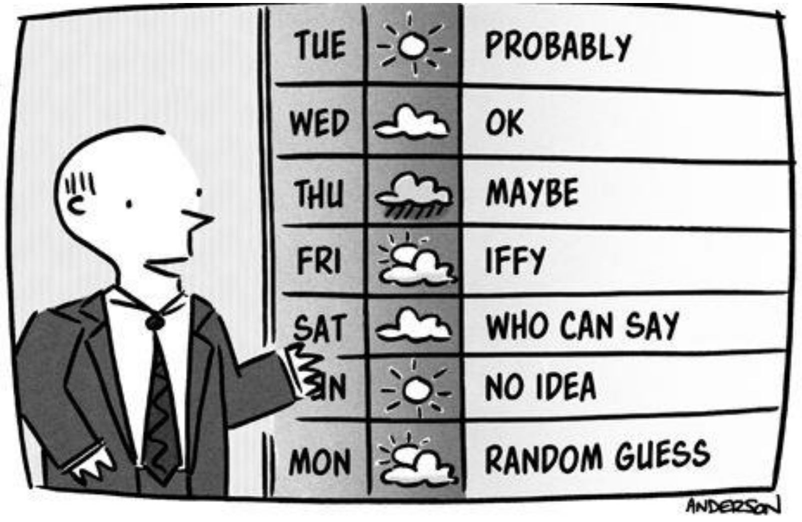
Probability does not tell us exactly what will happen, it is just a guide

Example: **toss a coin 100 times, how many Heads will come up?**

Probability says that heads have a ½ chance, so we would expect 50 Heads. But when you actually try it out you might get 48 heads, or 55 heads ... or anything really, but in most cases it will be a number near 50.

Weather Forecasting

Real Life Examples Of Probability

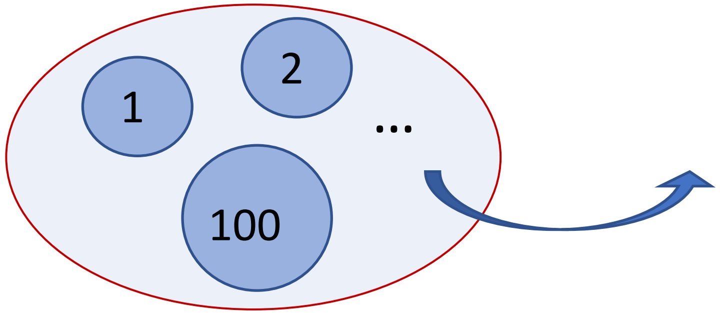


"And now the 7-day forecast..."

Before planning for an outing or a picnic, we always check the weather forecast. Suppose it says that there is a 60% chance that rain may occur. Do you ever wonder from where this 60% come from? Meteorologists use a specific tool and technique to predict the weather forecast. They look at all the other historical database of the days, which have similar characteristics of temperature, humidity, and pressure, etc. And determine that on 60 out of 100 similar days in the past, it had rained.

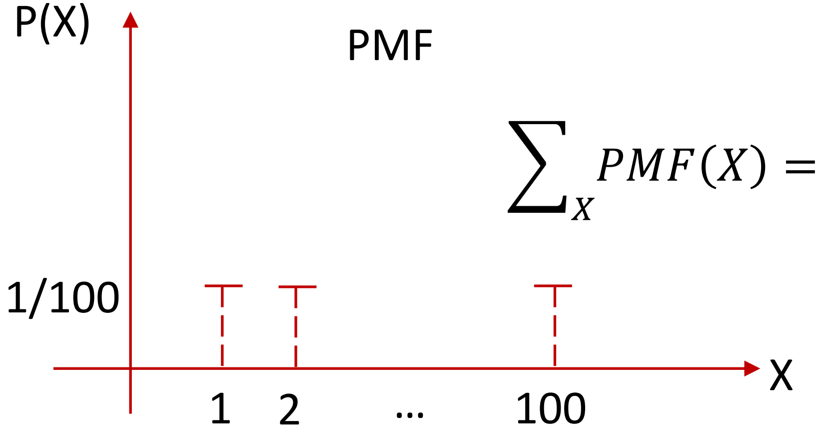
Probability distribution

Discrete



X	P(X)
1	1/100
2	1/100
...	...
100	1/100

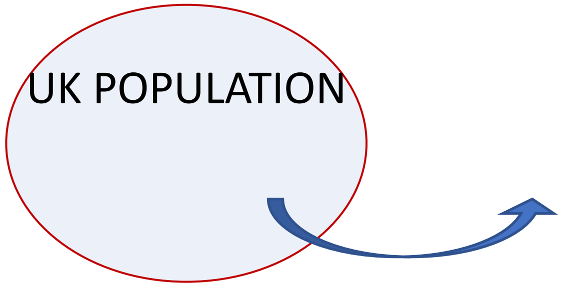
PMF



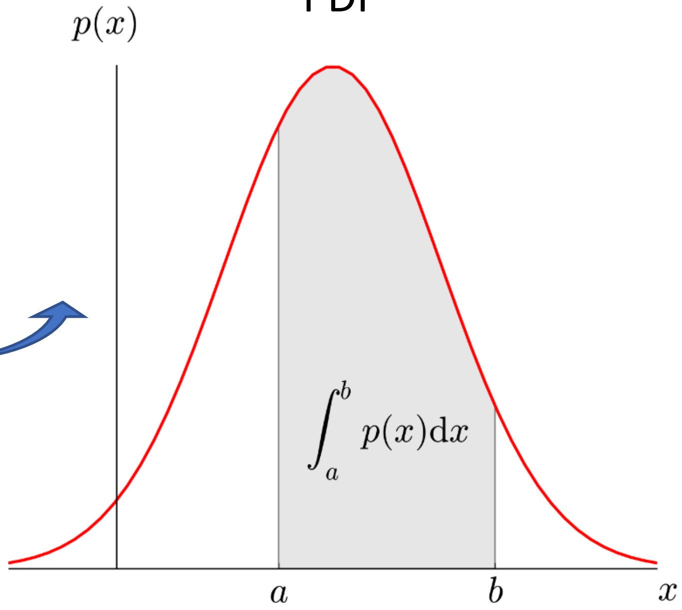
$$\sum_X PMF(X) = 1$$

Continuous

Height



PDF



X	P(X)
1.8 m	0

P given by the area

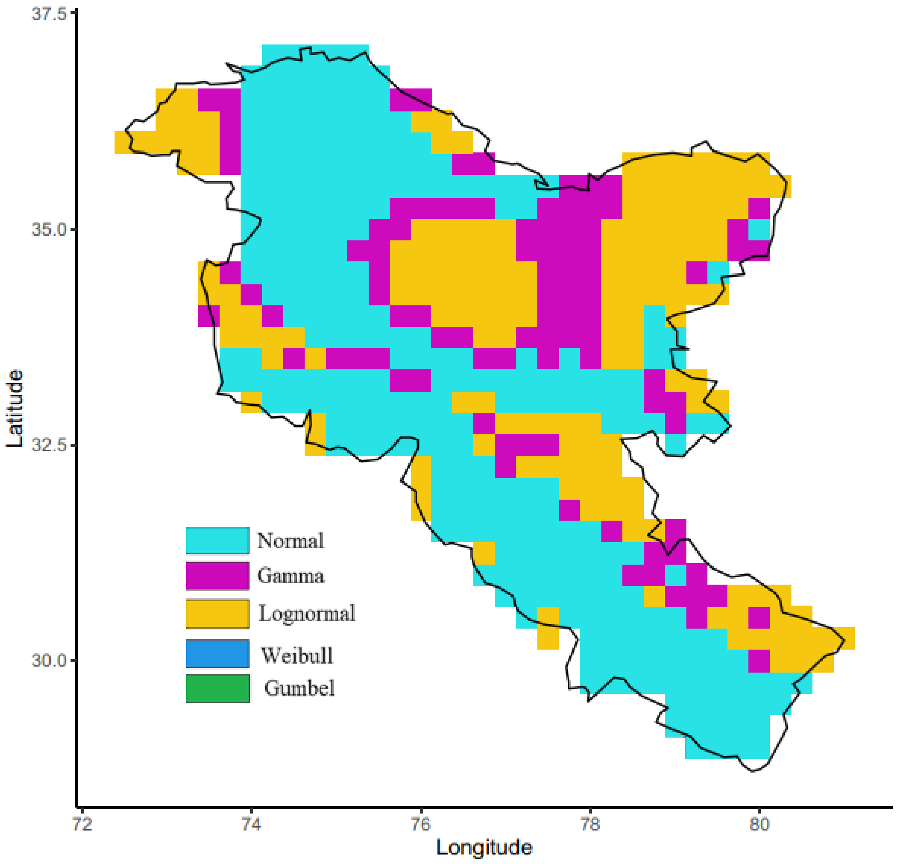
$$1.75 \le X \le 1.85$$

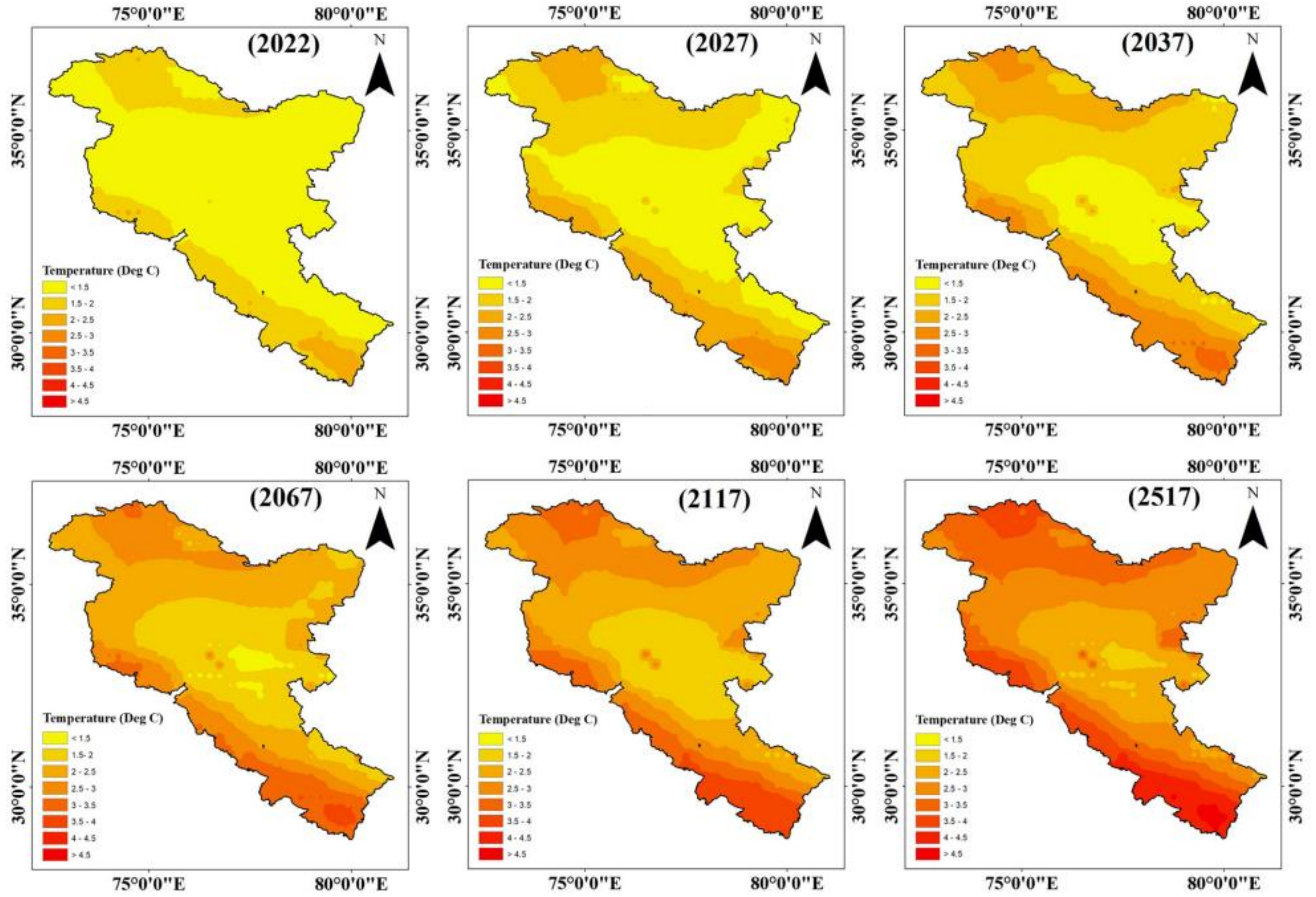
Fitting a probability distribution model

The Normal Distribution:
as mathematical function (pdf)

$$f(x) = \frac{1}{\boxed{\sigma}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\boxed{\mu}}{\sigma}\right)^2}$$

This is a bell shaped curve with different centers and spreads depending on μ and σ




(2022)
(2027)
(2037)
(2067)
(2117)
(2517)
N
75°0'0"E
80°0'0"E
35°0'0"N
30°0'0"N
Temperature (Deg C)
< 1.5
1.5 - 2
2 - 2.5
2.5 - 3
3 - 3.5
3.5 - 4
4 - 4.5
> 4.5

Data visualization

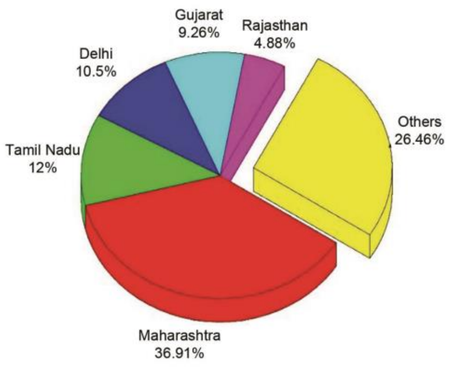


Figure 1. Distribution of COVID-19 confirmed cases in India as on 31 May 2020.

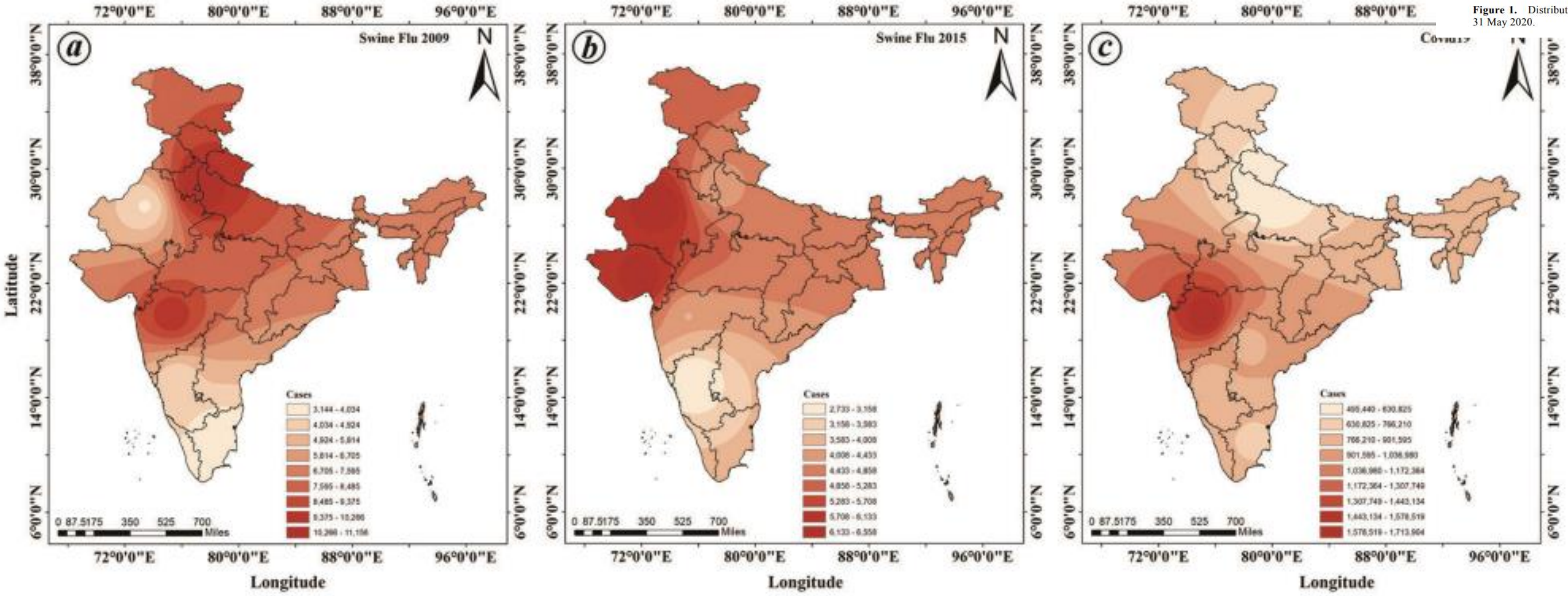


Figure 2. Hotspots of pandemics in the Indian states. ***a***, Swine flu (2009–10); ***b***, Influenza H1N1 (2014–15); ***c***, COVID-19 (2019–20).

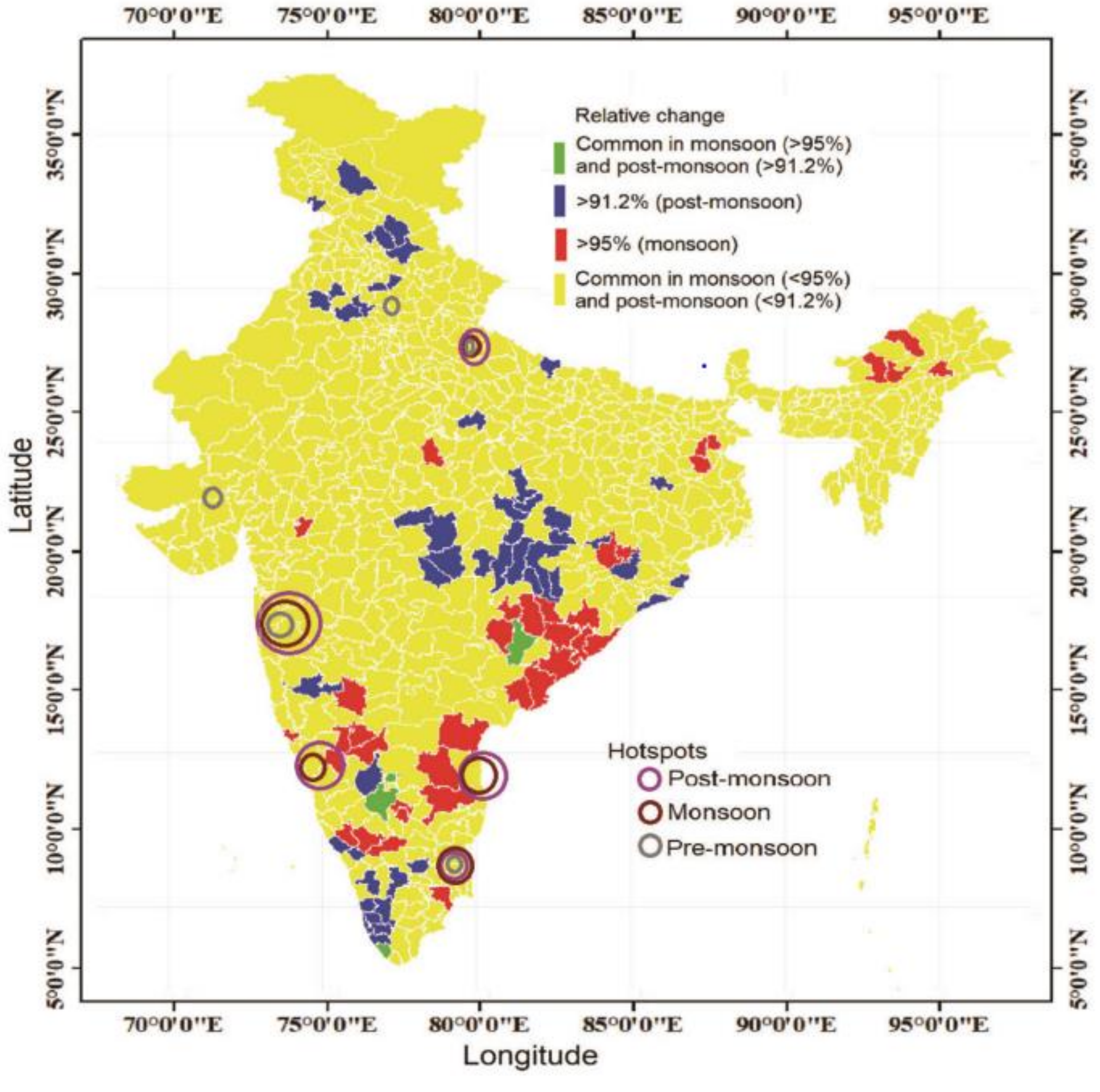


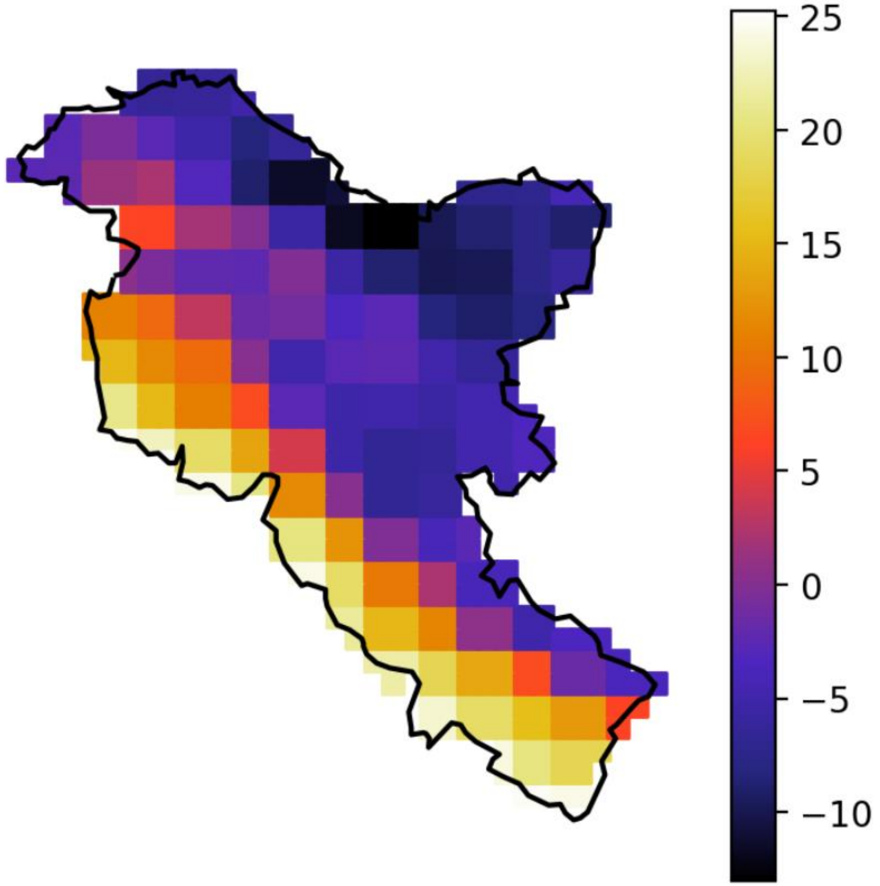
Figure 4. Trajectory of COVID-19 in India across different seasons.

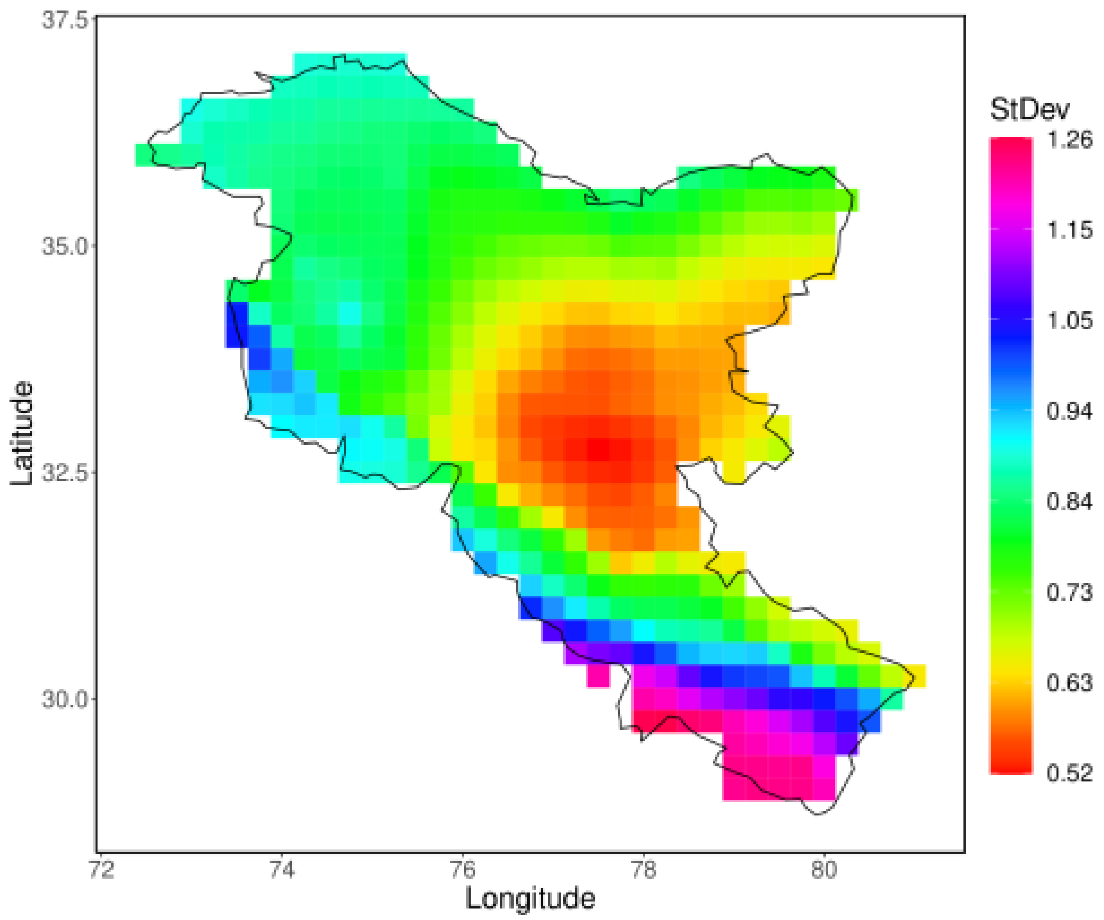
Statistics as a powerful tool of Data Analysis

Red districts showed a relative increase of 800% during monsoon at 95% confidence level .

Reason for such an increase is that these districts are closer to a large water body.

Simple Mean and Standard Deviations of Temperature over Himalayas





Regression

An optimal vaccination strategy for pandemic management and its impact on economic recovery

Vansh Kodesia[1], Ankur Suri[2] and Sarita Azad[3,*]

[1]School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi 175 005, India
[2]School of Engineering and Technology, University of Washington, Tacoma 98402, USA
[3]School of Mathematical and Statistical Sciences, Indian Institute of Technology, Mandi 175 005, India

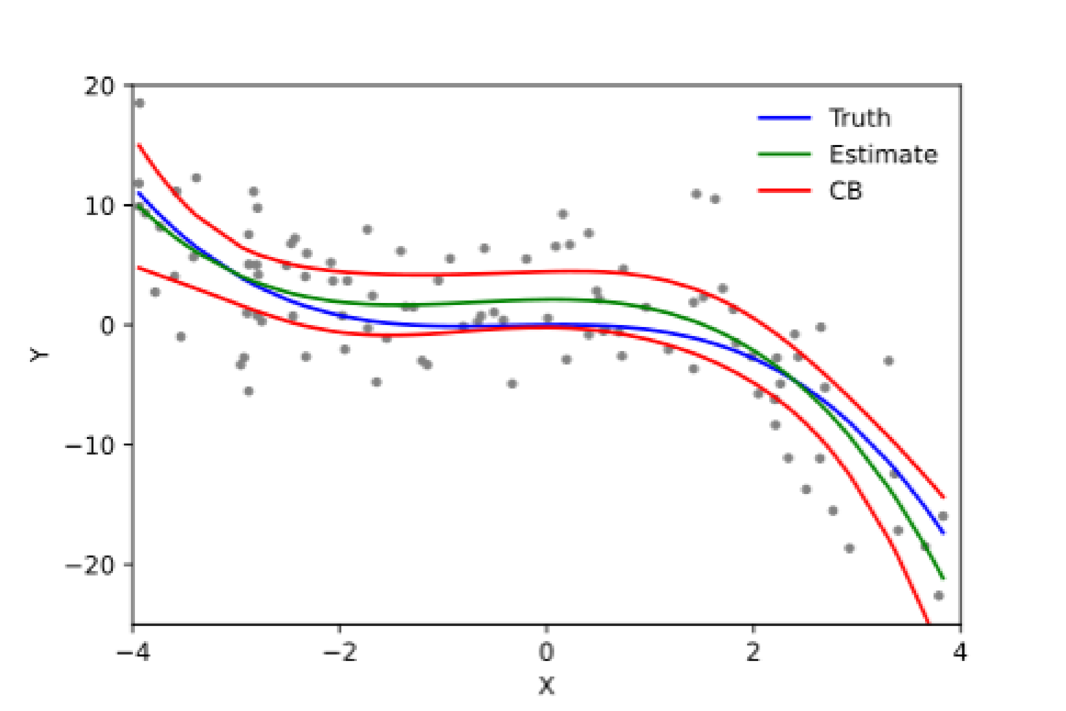
Regression

- In statistics, **polynomial regression** is a form of regression analysis in which the relationship between the independent variable *x* and the dependent variable *y* is modelled as an *n*th degree polynomial in *x*.

In general, we can model the expected value of *y* as an *n*th degree polynomial, yielding the general polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

A nonlinear regression model was developed to explore the regressors/predictors that affect vaccination coverage. The regressors (input variables) were taken as sites/area (sq. km, deaths%, confirmed cases% and recovered cases% and pharmaceutical mobility (mobility trends for places like pharmacies, grocery markets and drug stores)



A cubic polynomial regression fit to a simulated data set. The confidence band is a 95% simultaneous confidence band constructed using the Scheffé approach.

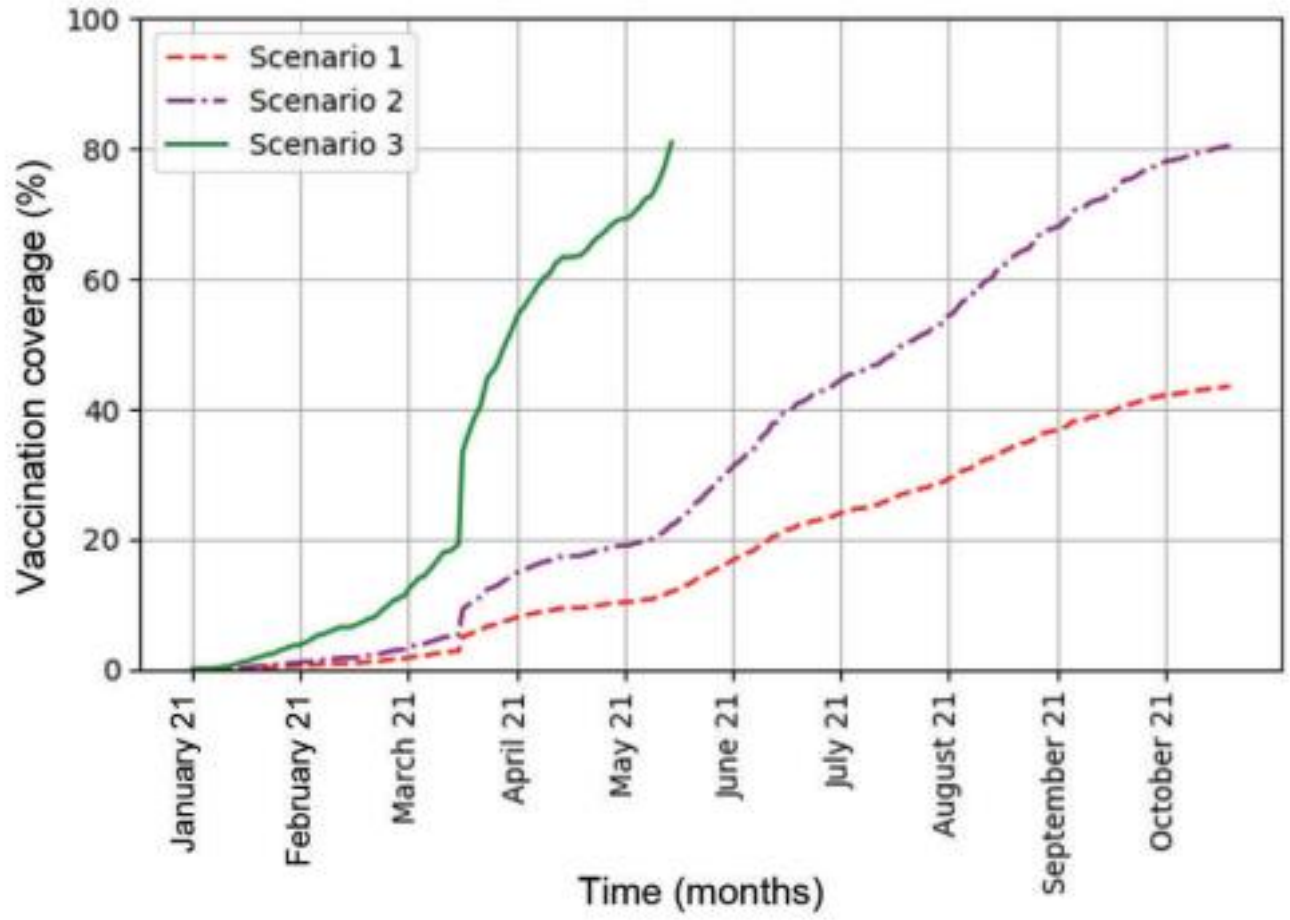
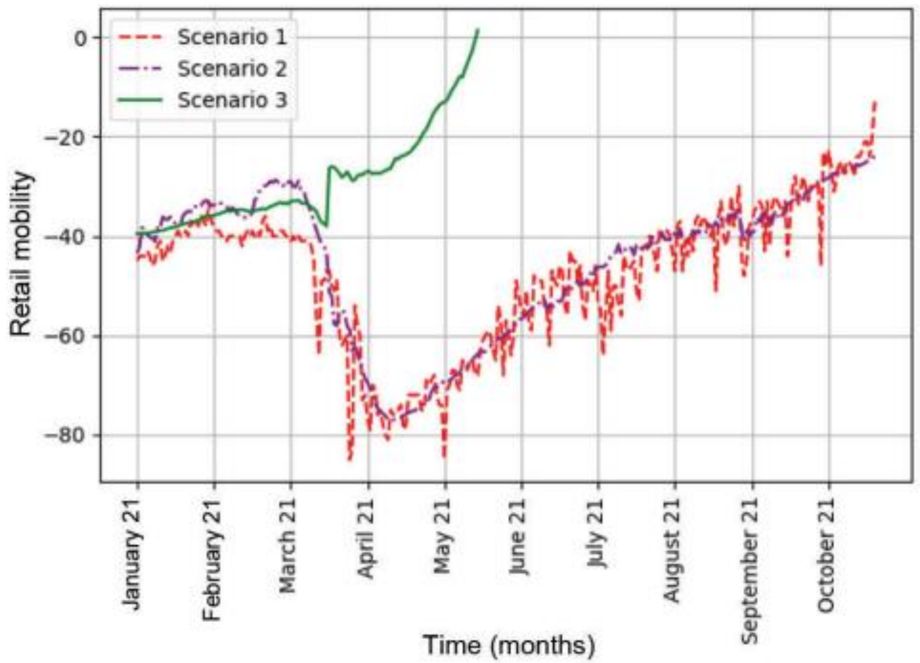
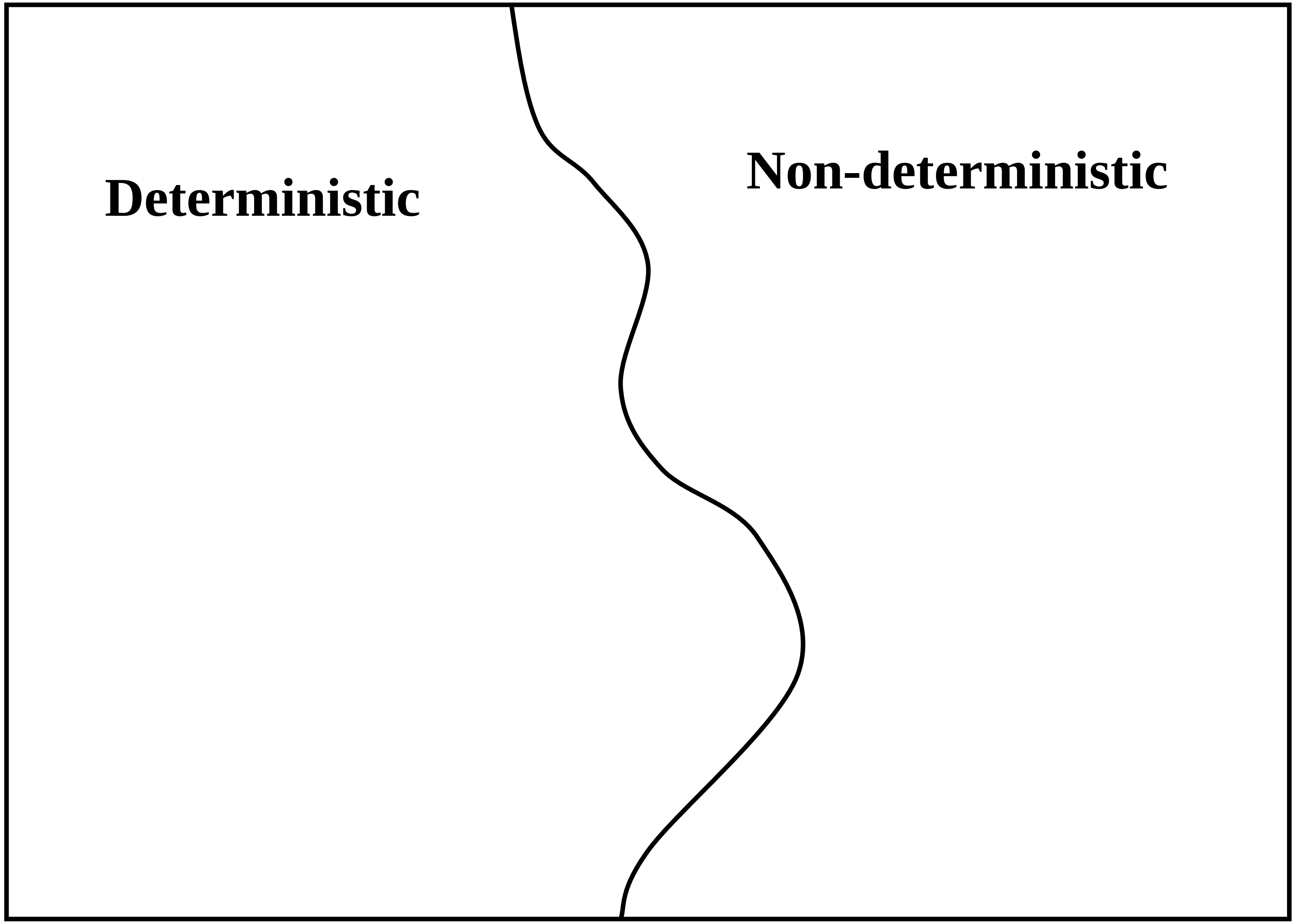


Figure 3. Vaccination coverage (OVC) for three scenarios in Mumbai city.

For scenario 3, using OVC 80% as input the recovery rate would have increased on an average by 5%. Based on this, retail mobility would have resulted in a 46% increase over the city's average. Similarly, transport mobility would have increased by 58% for Mumbai. On comparing the three scenarios, scenario 3 was found to be the most favourable. In Figure, the green curves with an exceptionally high slope represent scenario 3



Phenomena



Deterministic Phenomena

- There exists a mathematical model that allows "***perfect***" prediction the phenomena's outcome.
- Many examples exist in Physics, Chemistry (the exact sciences).

Non-deterministic Phenomena

- **No** mathematical model exists that allows "***perfect***" prediction the phenomena's outcome.

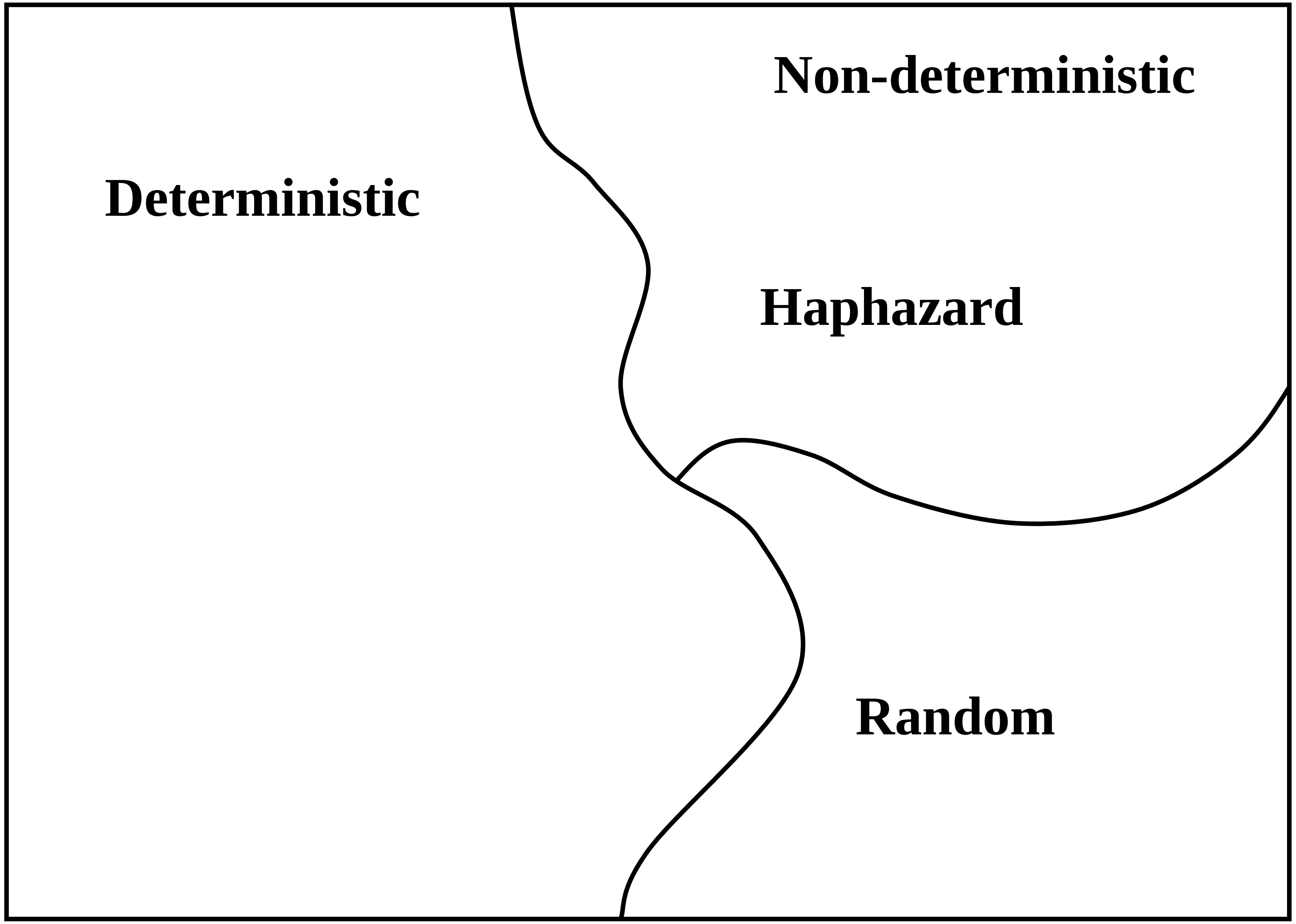
Non-deterministic Phenomena

- may be divided into two groups.

1. **Random phenomena**
   - Unable to predict the outcomes, but in the long-run, the outcomes exhibit statistical regularity.

2. **Haphazard phenomena**
   - unpredictable outcomes, but no long-run, exhibition of statistical regularity in the outcomes.

Phenomena



Random phenomena

– Unable to predict the outcomes, but in the long-run, the outcomes exhibit statistical regularity.

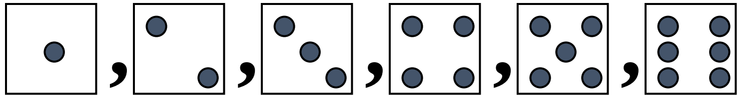
Examples

1. Tossing a coin – outcomes S ={**Head, Tail**}

Unable to predict on each toss whether is Head or Tail.

In the long run can predict that 50% of the time heads will occur and 50% of the time tails will occur

2. Rolling a die – outcomes

$S = \{$  $\}$

Unable to predict outcome but in the long run can one can determine that each outcome will occur 1/6 of the time.

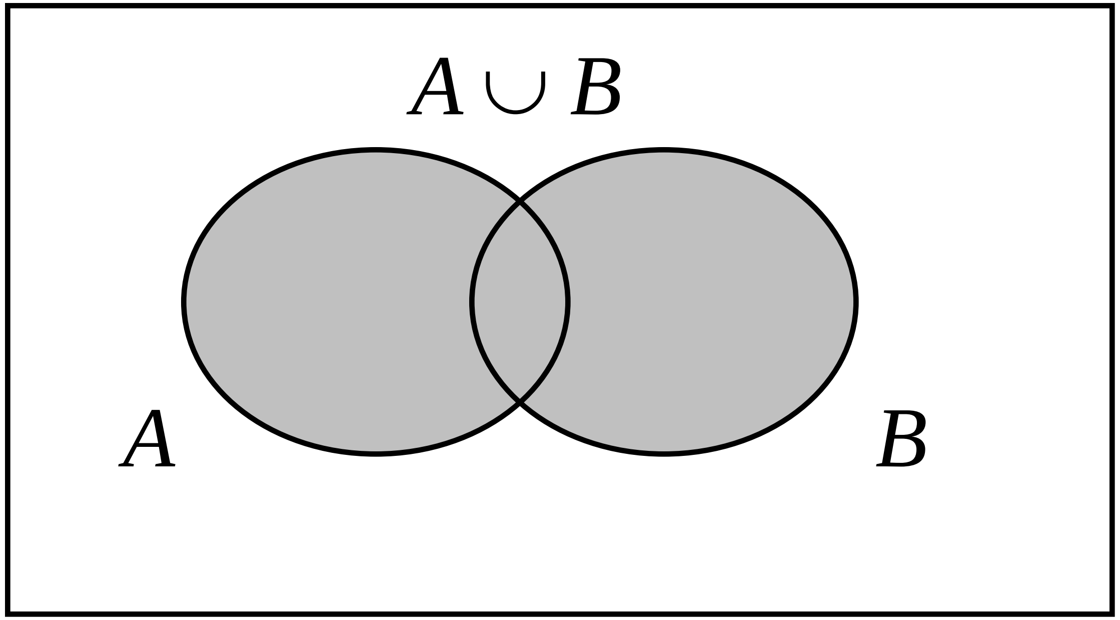
Use symmetry. Each side is the same. One side should not occur more frequently than another side in the long run. If the die is not balanced this may not be true.

Set operations on Events

Union

Let A and B be two events, then the **union** of A and B is the event (denoted by $A \cup B$) defined by:

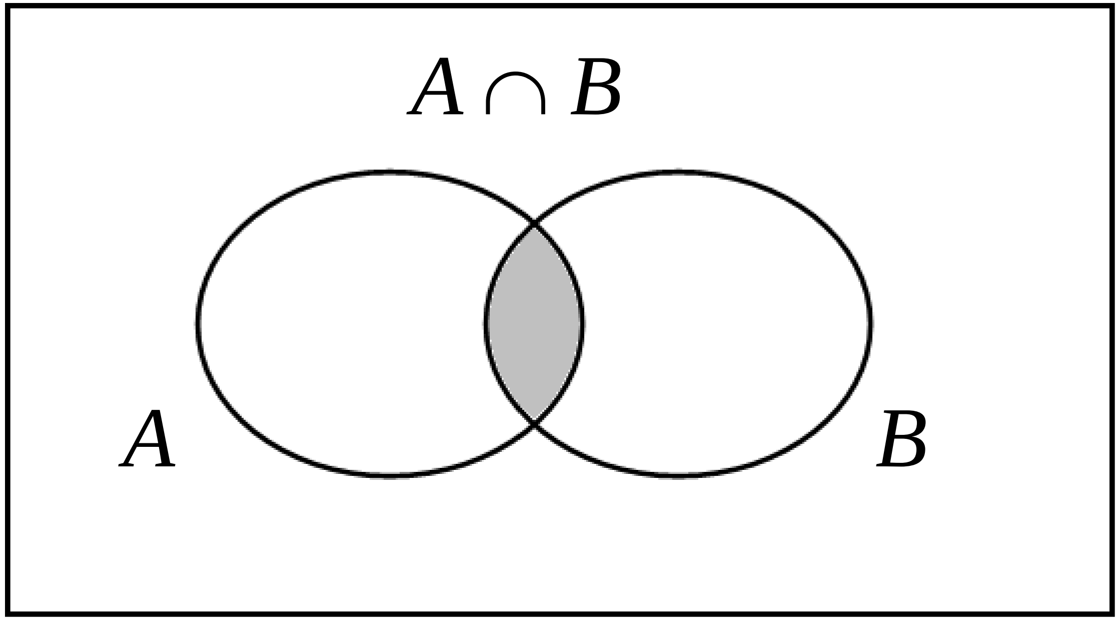
$A \cup B$ = {x| x belongs to A **or** e belongs to B}



Intersection

Let A and B be two events, then the **intersection** of A and B is the event (denoted by $A \cap B$) defined by:

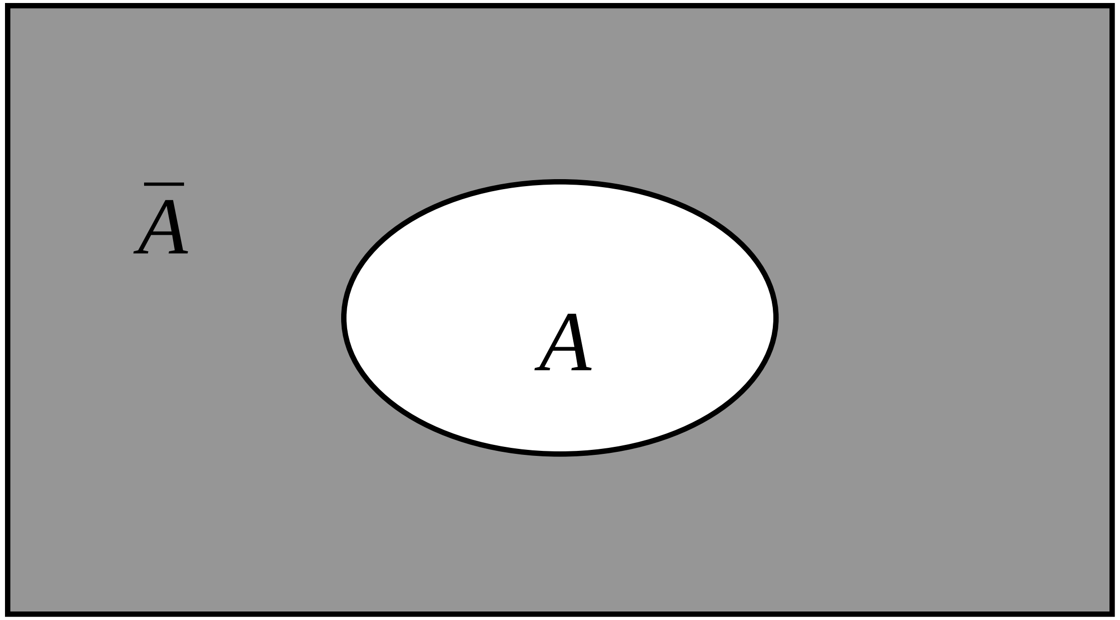
$$A \cap B = \{x |\ x \text{ belongs to } A \textbf{ and } x \text{ belongs to } B\}$$



Complement

Let A be any event, then the **complement** of A (denoted by $\bar{A}$) defined by:

$\bar{A}$ = {x| x **does not** belongs to A}

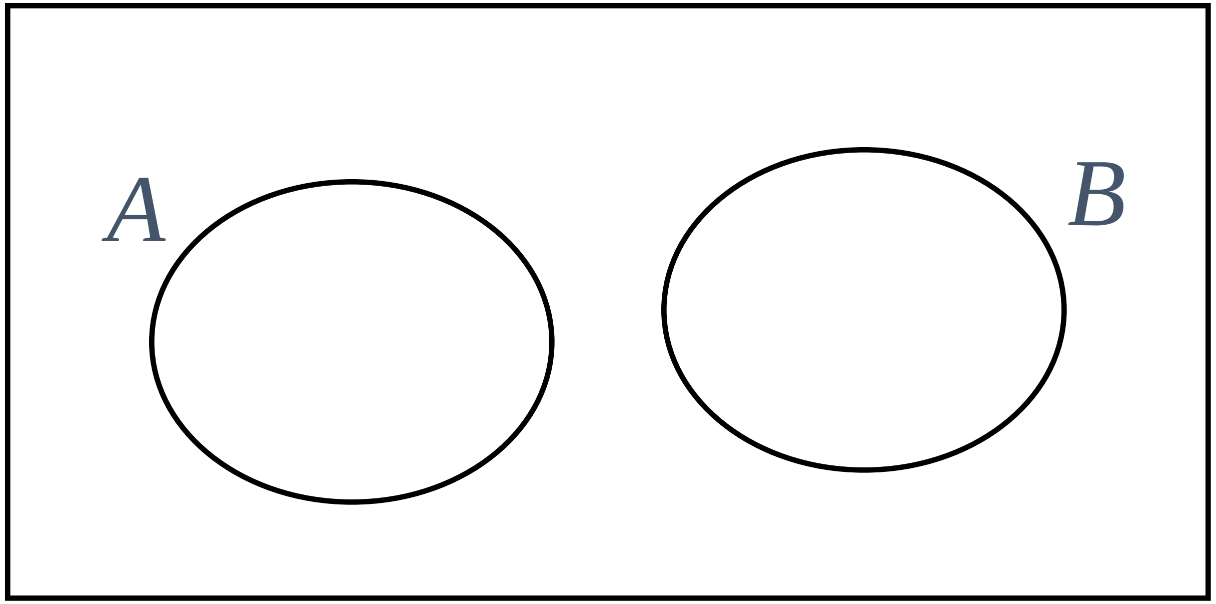


In problems you will recognize that you are working with:

1. **Union** if you see the word **or**,
2. **Intersection** if you see the word **and**,
3. **Complement** if you see the word **not**.

If two events *A* and *B* are are **mutually exclusive** then:

1. They have no outcomes in common.
   They can't occur at the same time. The outcome of the random experiment can not belong to both *A* and *B*.



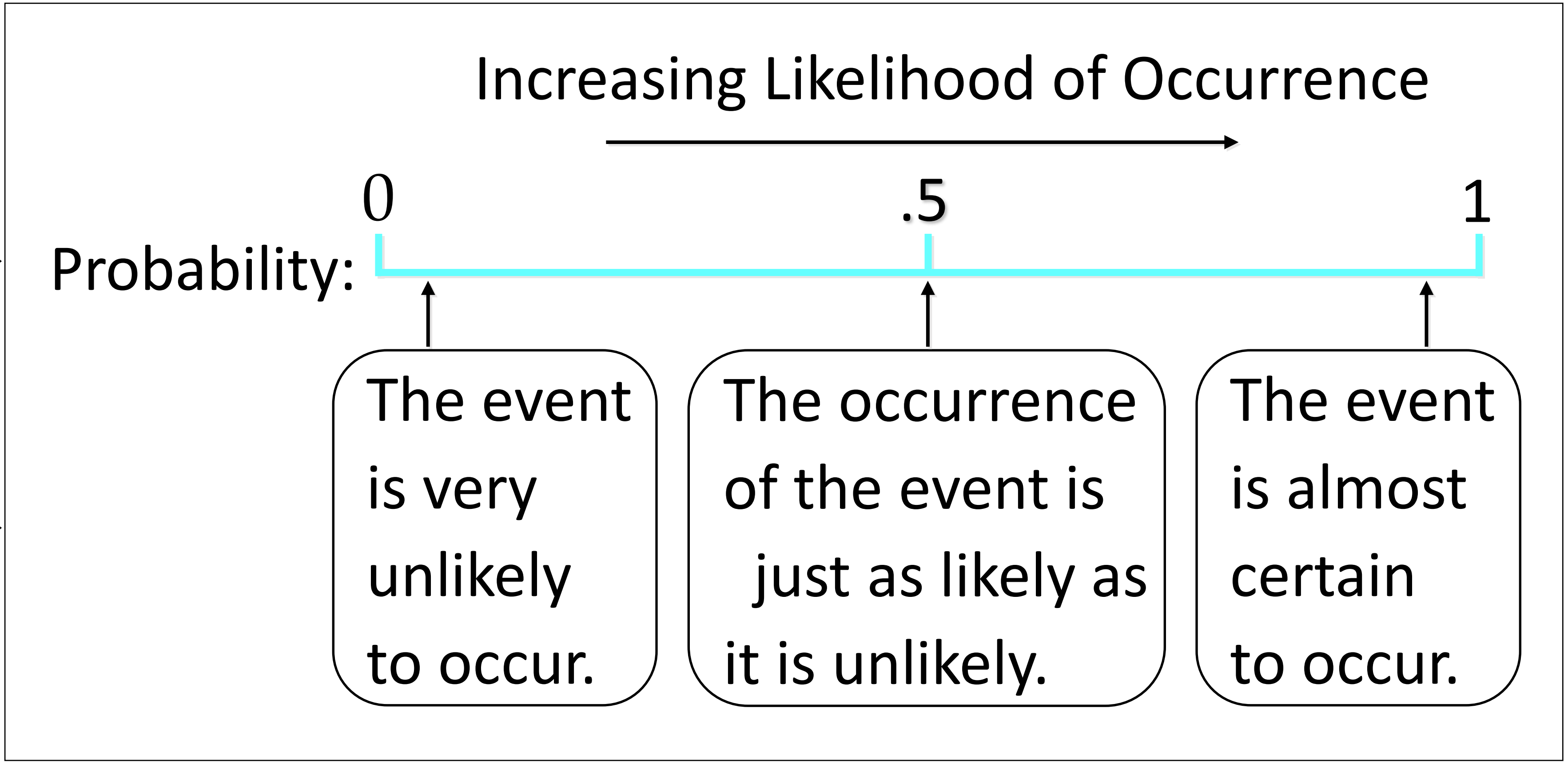
Probability

- By the very term probability we are trying to answer something about an uncertain situation.

- Idea of probability comes in context of carrying out a random experiment

- In a random experiment, when certain averages are taken over sufficient large outcomes, it approaches a steady value.

Probability as a Numerical Measure of the Likelihood of Occurrence



Assigning Probabilities

- Classical Method

  Assigning probabilities based on the assumption of equally likely outcomes

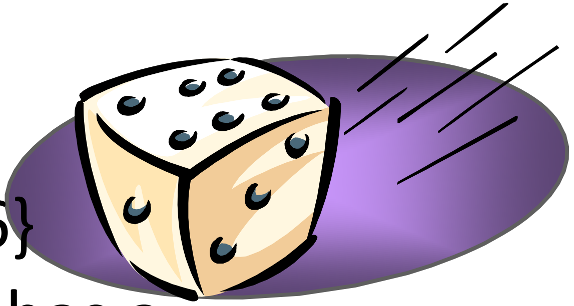
- Relative Frequency Method

  Assigning probabilities based on experimentation or historical data

Classical Method

If an experiment has n possible outcomes, this method would assign a probability of $1/n$ to each outcome.

Example



- Experiment: Rolling a die
- Sample Space: $S = \{1, 2, 3, 4, 5, 6\}$
- Probabilities: Each sample point has a 1/6 chance of occurring

Relative Frequency Method

- Example: Lucas Tool Rental

Lucas Tool Rental would like to assign probabilities to the number of car polishers it rents each day. Office records show the following frequencies of daily rentals for the last 40 days.

Number of Polishers Rented	Number of Days
0	4
1	6
2	18
3	10
4	2

Relative Frequency Method

Each probability assignment is given by dividing the frequency (number of days) by the total frequency (total number of days)

Number of Polishers Rented	Number of Days	Probability
0	4	.10
1	6	.15
2	18	.45
3	10	.25
4	2	.05
	40	1.00

4/40

Events

- An **event** is an outcome or a set of outcomes of a random process

Example: Tossing a coin three times

Event A = getting exactly two heads = {HTH, HHT, THH}

Example: Picking real number X between 1 and 20

Event A = chosen number is at most 8.23 = $\{X \leq 8.23\}$

Example: Tossing a fair dice

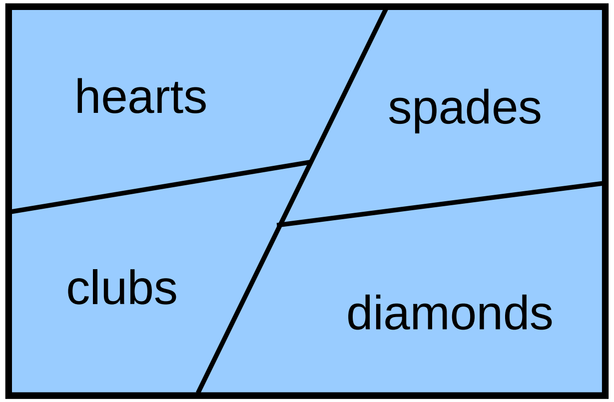
Event A = result is an even number = {2, 4, 6}

- Notation: P(A) = Probability of event A
- **Probability Rule 1:**

$$0 \leq P(A) \leq 1 \text{ for any event } A$$

Events

Two events A and B are ***mutually exclusive*** if $A \cap B = \emptyset$.



Definition

If $A_1, A_2, \ldots$ are mutually exclusive **and** $A_1 \cup A_2 \cup \ldots = S$, then the collection $A_1, A_2, \ldots$ forms a ***partition*** of S.

Sample Space

- The **sample space** S of a random process is the set of all possible outcomes

  Example: one coin toss

  S = {H,T}

  Example: three coin tosses

  S = {HHH, HTH, HHT, TTT, HTT, THT, TTH, THH}

  Example: roll a six-sided dice

  S = {1, 2, 3, 4, 5, 6}

  Example: Pick a real number X between 1 and 20

  S = all real numbers between 1 and 20

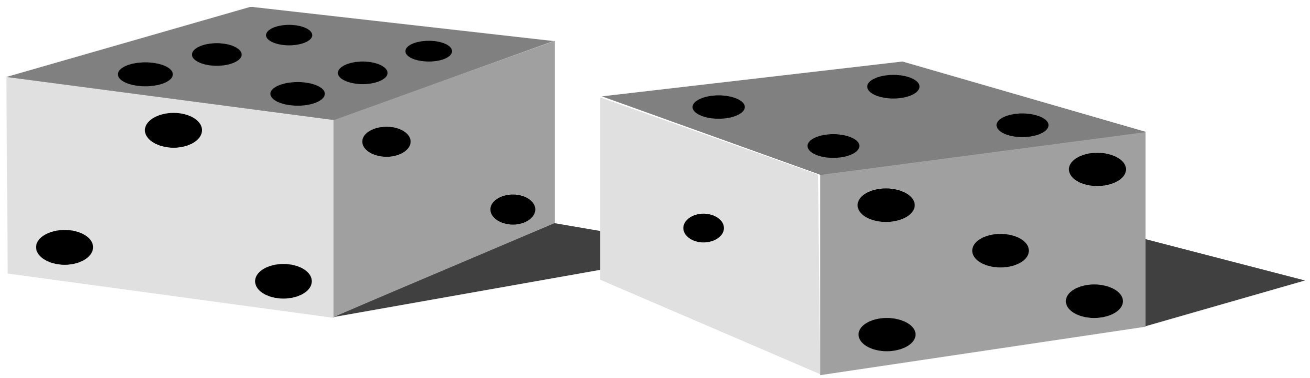
- **Probability Rule 2: The probability of the whole sample space is 1**

$$P(S) = 1$$

Example:
THE GAME Of CRAPS

In Craps one rolls two fair dice.

What is the probability of the sum of the two dice showing 7?



(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

					(1,6)
				(2,5)	
			(3,4)		
		(4,3)			
	(5,2)				
(6,1)					