Data, Statistics and Probability

Week 0: Introduction

Overview

- What you have learnt so far?
  - Statistics 1: handle data, summarise it in different ways, handle simple probabilistic scenarios
  - Computational Thinking: perform tasks with data like a computer
  - Mathematics 1, English 1: basic foundations
- Statistics 2
  - Continue your study of data, statistics and probability
  - Handle larger quantity of data from real-life situations
  - Consider more complex probabilistic scenarios
  - Study some classical statistical questions and procedures
- Grading (note a minor difference from other courses)
  - Quizzes and exams: 100%
  - Activities: 10% bonus

Scientific study of phenomena

Image: wikipedia commons

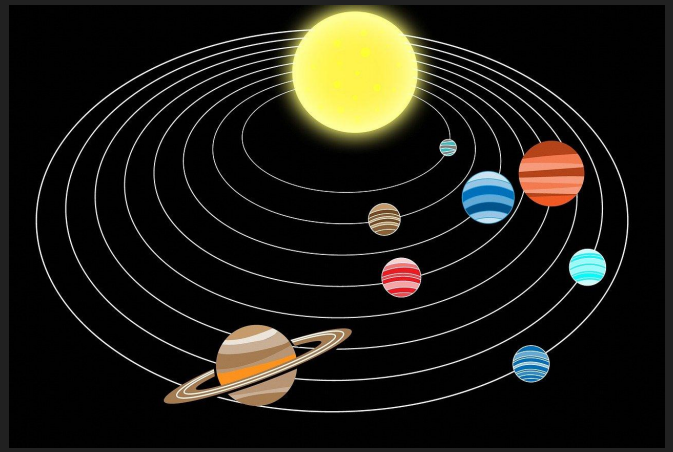
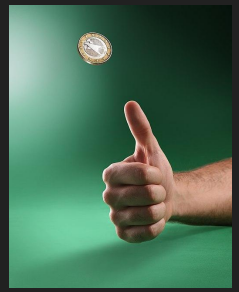


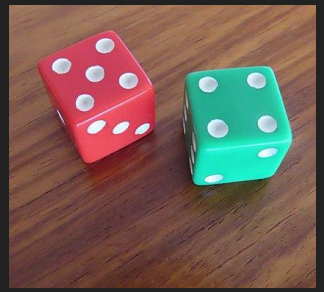
Image: pixaby

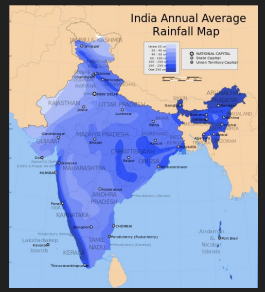
- Observations: provide data (location of objects at different times)
  - Patterns in the data - apple falls to the ground, planets revolve around the sun
- Theory: consistent explanation for observed patterns in data (Newton’s laws)
- Applications: plentiful….

Deterministic vs random-like patterns

- Many patterns are “deterministic”
  - Example: current position = initial position + velocity x time
- What happens when the situation is more complex?







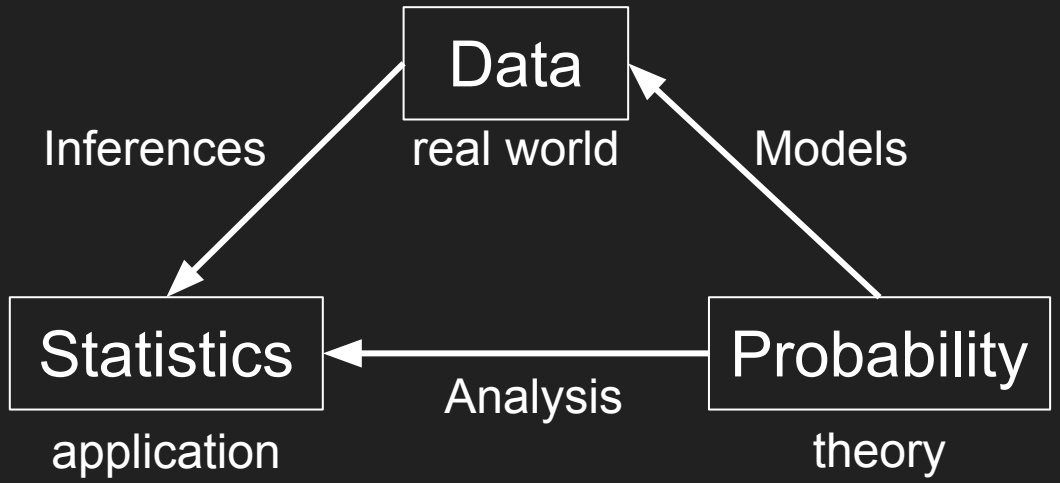
Images: wikipedia commons

- No “deterministic” equation (not enough information or processing ability)
  - There are still many useful, observed patterns with lots of applications
- How to study patterns in random-looking phenomena and use them?

Nature of patterns in random-like phenomena

- When phenomena are unpredictably complex, we cannot expect exact models
  - Toss a coin: cannot exactly say whether result will be heads or tails
  - Rainfall: cannot exactly say rainfall quantity in a year
  - Stock market: cannot exactly say how high a stock's value will be tomorrow
- Intuitively, we expect some patterns when the random phenomenon repeats
  - Toss a "fair" coin: half the time heads, half the time tails
  - Rainfall: average/above average/below average predictions, 70% chance of rain tomorrow
  - Stock market: bullish trends or bearish trends are predicted by traders
- This course: Statistical study
  - Collect/observe data and use probabilistic analysis to make inferences
  - Has been very successful in a wide range of applications over the last 100 years or so

Statistical study of phenomena



Activities in this course

- This course will involve several activities, which will account for 10 percent of your marks
- Depending on your work, you will earn 10, 7.5, 5, 2.5 or 0 marks for these activities
- How will grading happen?
  - Complete >= 75% of activities - 7.5 marks
  - Complete >= 50% and < 75% of activities - 5 marks
  - Complete >= 25% and < 50% of activities - 2.5 marks
  - Complete < 25% activities - 0 marks
  - Bonus - 2.5 marks (awarded at discretion of instructors and faculty)

Week 0: Activity 1

- Login to Google Sites (sites.google.com) with your onlinedegree id
- Create a site using the “Student Portfolio” template
- Add “Statistics 2” as one of the classes
  - Add a short description of the course
  - Add other material in the pages and change settings as you see fit. This is your page!
- All other activities in this course will be added to this Google Sites page
- Publish the page so that it is visible only to those within the IIT Madras organization
  - Do not make it public
- Week 0: Subjective activity assignment 1
  - Submit the url of your sites page

IPL data - Running example

Indian Premier League

- Twenty-twenty cricket tournament
- 2008-2020
  - Data on all matches available from various sources on the Internet
  - Match data
    - Venue, teams, toss
  - Ball-by-ball data for every over
    - Runs scored in every delivery - extras
    - Batsman, bowler, non-striker
    - Wickets, how out
- Are there patterns in the data? How to study it statistically?
  - This will be a running example throughout the course

Data and parsing

https://cricsheet.org/

- Download zip file for IPL matches
  - 816 matches from 2008 - 2020
  - Unzip the file into a folder
- For each match, a “yaml” file is provided
  - “yaml”: human-readable format
  - Open and see using a text editor
- To study the data, we need to process it
  - What data is needed? Runs scored in every ball? Wickets taken by a bowler?
  - Descriptive statistics of data: useful to parse the necessary data into a spreadsheet
  - How to go from yaml to spreadsheet?

Sample yaml file

```yaml
info:
  city: Bangalore
  competition: IPL
  dates:
    - 2013-04-09
  gender: male
  match_type: T20
  outcome:
    by:
      wickets: 7
    winner: Royal Challengers Bangalore
  overs: 20
  player_of_match:
    - V Kohli
  teams:
    - Royal Challengers Bangalore
    - Sunrisers Hyderabad
  toss:
    decision: bat
    winner: Sunrisers Hyderabad
  umpires:
    - S Ravi
    - SJA Taufel
  venue: M Chinnaswamy Stadium
innings:
  - 1st innings:
      team: Sunrisers Hyderabad
      deliveries:
        - 0.1:
            batsman: PA Reddy
            bowler: RP Singh
            non_striker: PA Patel
            runs:
              batsman: 0
              extras: 0
              total: 0
        - 0.2:
            batsman: PA Reddy
            bowler: RP Singh
```

Open Activity 1 (ungraded)

- Write a python script
  - To read all IPL yaml files and collect the actual data you need
  - Write the data into a spreadsheet
  - Several python data structures (lists, dictionaries) and modules (os, yaml, pandas) will be useful

When you study python, remember to focus on topics such as above!

This course - IPL Powerplay Overs

- Running example will be “IPL Powerplay Overs”
  - What is powerplay?
    - Overs 1-6 of an IPL innings
    - Only 2 fielders allowed on the boundary
- Data
  - Collect data about runs scored and wickets that fell during Overs 1 - 6 of an IPL game
    - 1st and 2nd innings
  - Are there patterns?

IPL Powerplay Data - sample from spreadsheet

date	venue	innings	target	team	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	d1	r1	w1
2008-04-18	M Chinnaswamy Stadium	1		Kolkata Knight Riders	1	0	1	0	0	0	1		7	3	0
2008-04-18	M Chinnaswamy Stadium	2	222	Royal Challengers Bangalore	1	1	0	1	1	0	0		7	4	0
2008-04-19	Punjab Cricket Association Stadium, Mohali	1		Chennai Super Kings	0	0	1	0	4	0			6	5	0
2008-04-19	Punjab Cricket Association Stadium, Mohali	2	240	Kings XI Punjab	4	0	0	1	0	4			6	9	0

Descriptive statistics - search for patterns

- Average runs by over
  - Over 1: 6.00, Over 2: 7.15, Over 3: 7.97, Over 4: 8.20, Over 5: 8.24, Over 6: 8.30
- Standard deviation of runs by over
  - Over 1: 3.85, Over 2: 4.23, Over 3: 4.53, Over 4: 4.47, Over 5: 4.57, Over 6: 4.79
- Average runs in Over 6 by team
  - Delhi Capitals - 8.94, Delhi Daredevils - 7.48
  - Sunrisers - 8.84, CSK - 8.81, Royals - 7.88, RCB - 7.81
- Wickets in Over 5
  - 0: 1216 innings, 1: 352 innings, 2: 30 innings
- Average runs in Over 6 by Wickets in Over 5
  - 0: 8.77, 1: 6.71, 2: 7.6

What will we try to do with the powerplay data?

- In this course
  - We will study basics of probability and statistics
  - For many concepts, we will try illustrations with an IPL powerplay example
- Probability
  - Are there models that explain the observed patterns in the data?
  - How will you “simulate” powerplay overs in an IPL innings?
- Statistics
  - Can you draw inferences from the data? Make predictions?

Week 0: Activity 2

- Find one other example of data from an interesting source
  - Data may be available in different formats
  - You might have to learn about the format
- Collect it into a Google Sheet
  - Spreadsheet must have at least 100 rows and 5 columns
- Google Sites portfolio page
  - Add a section in your “Statistics 2” class page for “Week 0: Activity 2”
  - Provide a link to the Google Sheet above
  - Describe the rows and columns of the data in your own words
  - Add anything interesting about the data
- Week 0: Subjective activity assignment 2
  - Submit the url of your “Statistics 2” class page