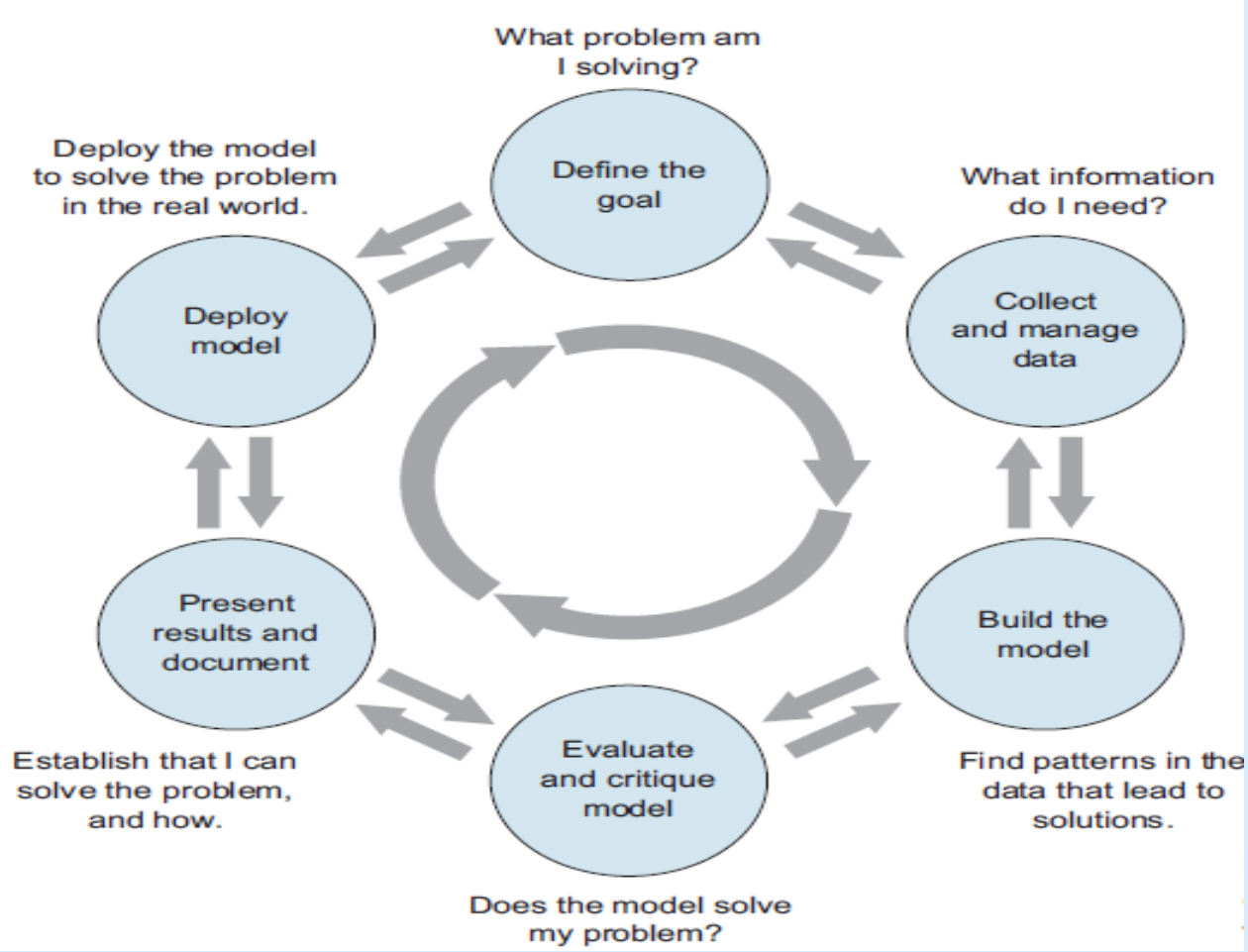
Probability & Statistics

LET US TRY TO ANSWER ….

- *You and a friend are at a cricket match, and out of the blue he offers you a bet that neither player will hit a century in that game. Should you take the bet?*
- *Your company is launching personalized marketing campaign to millions of potential customers. To which customer should you offer what type of product.*
- *A widget maker in your factory that normally breaks 4 widgets for every 100 it produces has recently started breaking 5 widgets for every 100. When is it time to buy a new widget maker?*
- *You are conducting poll on national election for a big media house. How many people do you have to poll? How do you ensure that your poll is free of bias? How do you interpret your results?*

Data Science Project Lifecycle



Why Stats for business analytics???

- Data Collection – What kind of data, sampling, are there any biases
- Data Cleaning & Visualization – Distribution of data, how to summarize
- Data Analysis – Which algorithm? Does it fit data? Assumptions
- Communication of results- what does p value mean, am I confidence statistically about results?

Data Science and Statistics

- A successful data scientist is one who knows more programming than a statistician and more statistics than a programmer.

- “Statistics is a crucial component of data science. At Twitch, our data science team brings together three things: statistics, programming, and product knowledge. And we would never hire someone who wasn’t strong in stats. You can be a great programmer, but if you don’t know what Bayes Rule is, then we have an engineering department I can point you to.”

Let us begin with the fundamental building blocks….Probability

The most essential skill you need to make your case as a data scientist

What is probability?

- Measure of likeliness of something happening
  - Strength of belief that something is true
  - Mathematical way of expressing uncertainty
- Given n observations of an event, it denotes the proportion of observations where a given event occurs
- Prob = (number of desired outcomes)/(total outcomes)
- Prob of a single event is always between 0 and 1
- Prob of all possible outcomes always sums to 1

Examples

- In a coin toss, prob of a head appearing?
- In a roll of dice, prob of 3 appearing?
  - Six possible outcomes: {1,2,3,4,5,6}
  - Each outcome equally likely, therefore prob of an outcome:1/6
  - Prob of an odd number appearing?
- Prob of amount of hailstorm in Delhi in March?
- Prob of RCB winning IPL 2020?

Technical Notations

- Experiment
  - Deterministic: Outcome always same and determined
  - Random: Many possible outcomes from a range of value
- Sample Space
  - Given a random experiment K, set of all possible outcomes for K
  - Denoted by S
  - Eg. For coin toss K, S = {H,T}
  - R package for common event prob is prob
  - library(prob)
  - tosscoin(2)
  - rolldie(2)
- Event
  - Subset of sample space for which the outcome is true

Example

- Imagine rolling two dice simultaneously.
  - What is the sample space of this experiment?
  - What is the probability of sum of 12 appearing?
  - What is the probability of 5 appearing on either of the dice?

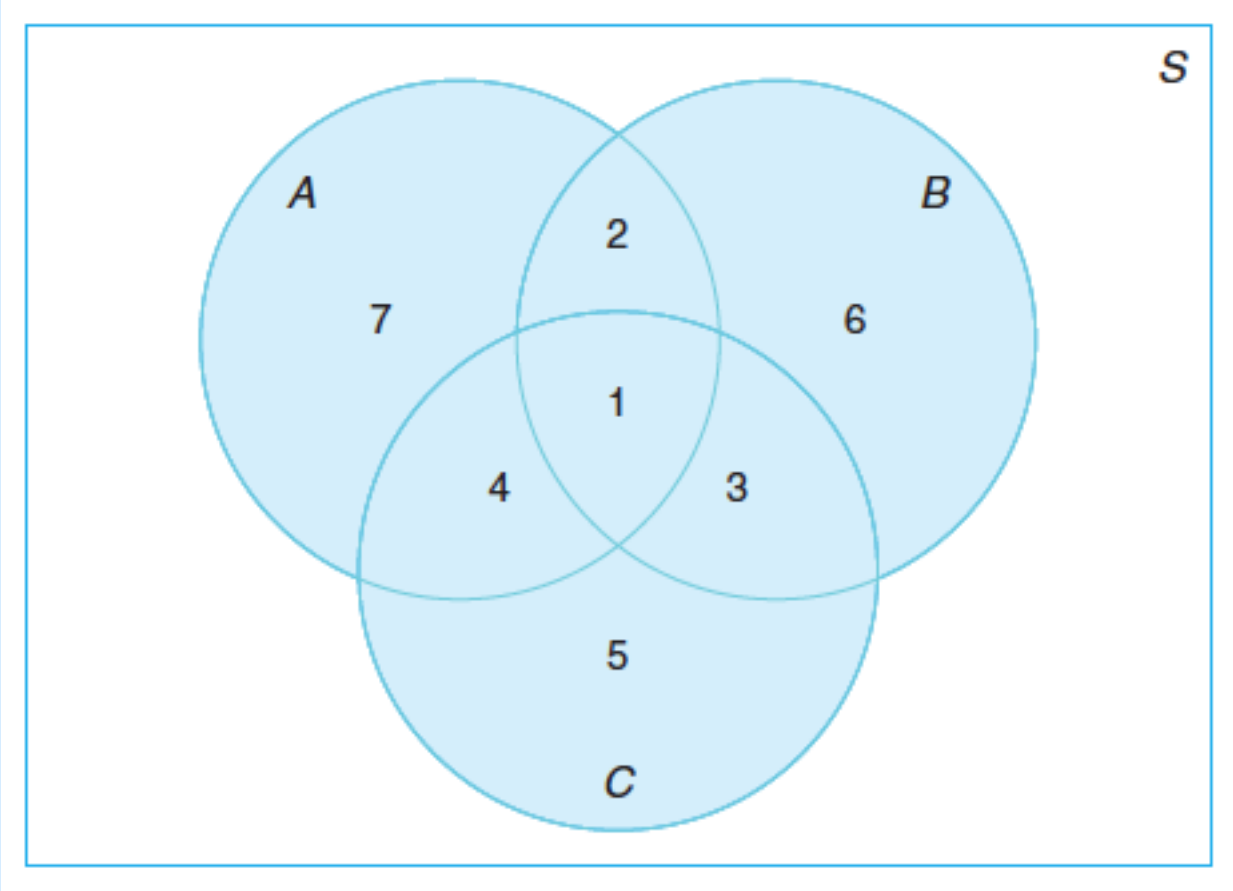
More on Events

- Event
  - Subset of sample space for which the outcome is true
  - Formally, prob of an event A is written as P(A)
  - $0 <= P(A) <= 1$
  - Complement (A'): Set of all elements in S that are not in A.
- Consider two events A and B
  - Intersection ($A \cap B$): Set of all elements common to A and B
  - Union (A U B): Set of all elements belonging to either A or B
  - Difference (A-B): Elements belonging to A but not to B
- A and B are mutually exclusive or disjoint if $A \cap B = \emptyset$

Probability Axioms

- For an experiment K, and event A:
  - $0 <= P(A) <= 1$
  - $P(\emptyset) = 0$
  - $P(S) = 1$
  - If A1, A2, A3…are mutually exclusive then $P(A1 \cup A2 \cup A3\ldots) = P(A1)+P(A2)+P(A3)\ldots$

Venn Diagrams



Example

- Example: Toss a coin twice. What is probability of
  - Heads appearing on first roll
  - Heads appearing at least once
- Roll a dice and flip a coin. What is probability of
  - A heads and a 5
  - A heads or a 5
- CBA has 120 participants. 60 are from IT, 30 from Finance, 10 Doctors, and 20 are CA. I select a person randomly. What is prob that person is
  - From Finance
  - From either CA or IT

Additive Rules

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$
- $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

Experimental Probabilities

- Cases we looked till now are subjective or theoretical probabilities
- More often, we deal with experimental probabilities
  - Especially in Analytics and Data Science
  - Prob arising from an experiment
  - (number of outcomes)/(number of trials)
- Example: flip a coin five times and report prob of head

Example

- Load up file sms_spam.csv in R
- What is structure of data?
- Count number of observations for ham and spam
- Store ham and spam observations into separate datasets
- Compute probability of a spam message in the dataset

To sum up

- Why do we need probability
  - Formal way to make sense of the world
  - Express uncertain outcomes
  - Organize data meaningfully
- Sample Space
- Event
  - Independent and Dependent