Quick Intro to Probability

Data 3402- Lecture 6

Amir Farbin

Some terms

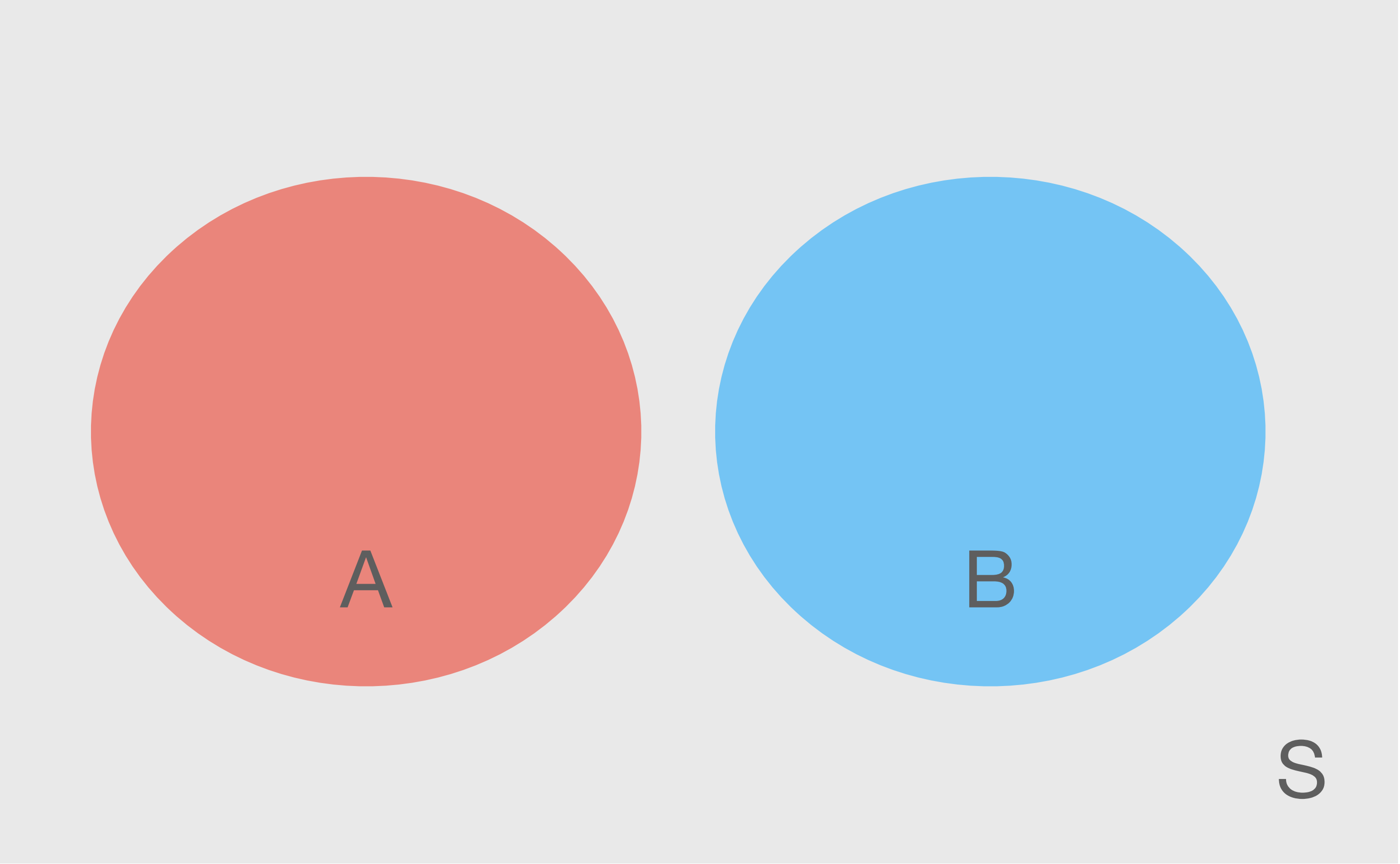
- **Probability**: Mathematical framework for quantifying likelihood of future events.
- **Statistics**: Analysis of data using language of probability. Why? Answer questions, quantifying uncertainty, predict future events.
- **Data**: Observations / Measurements of past events.
    - **Experiment**: Controlled setup to test specific hypothesis.
    - Data Collection can be seen as type of an experiment.
- **Modeling**:
    - **Theoretical Model**: Captures assumptions or best understanding, possibly with no/little quantitive input from data.
    - **Empirical model**: Based purely on pervious observations.
    - **Statistical modeling**: Predicting probability of future events using pervious data.
    - **Generative model**: Create new data based on a model.
        - **Simulation**: Using theory to create new data.
        - **Monte Carlo Simulation**: Use probabilistic model to create new data. Can be statistical or first principles.
- **Science**: At core, using data to test / enhance theoretical models -> improve our best understanding of phenomena.
    - **Scientific Process**: Propose model. Devise experiment to test model (focus on weakness). Make a prediction (hypothesis). Perform statistical analysis to gleam answer to test from data with uncertainties. Update model. Repeat.
- Beyond science: use previous data and/or theoretical models to predict future events.

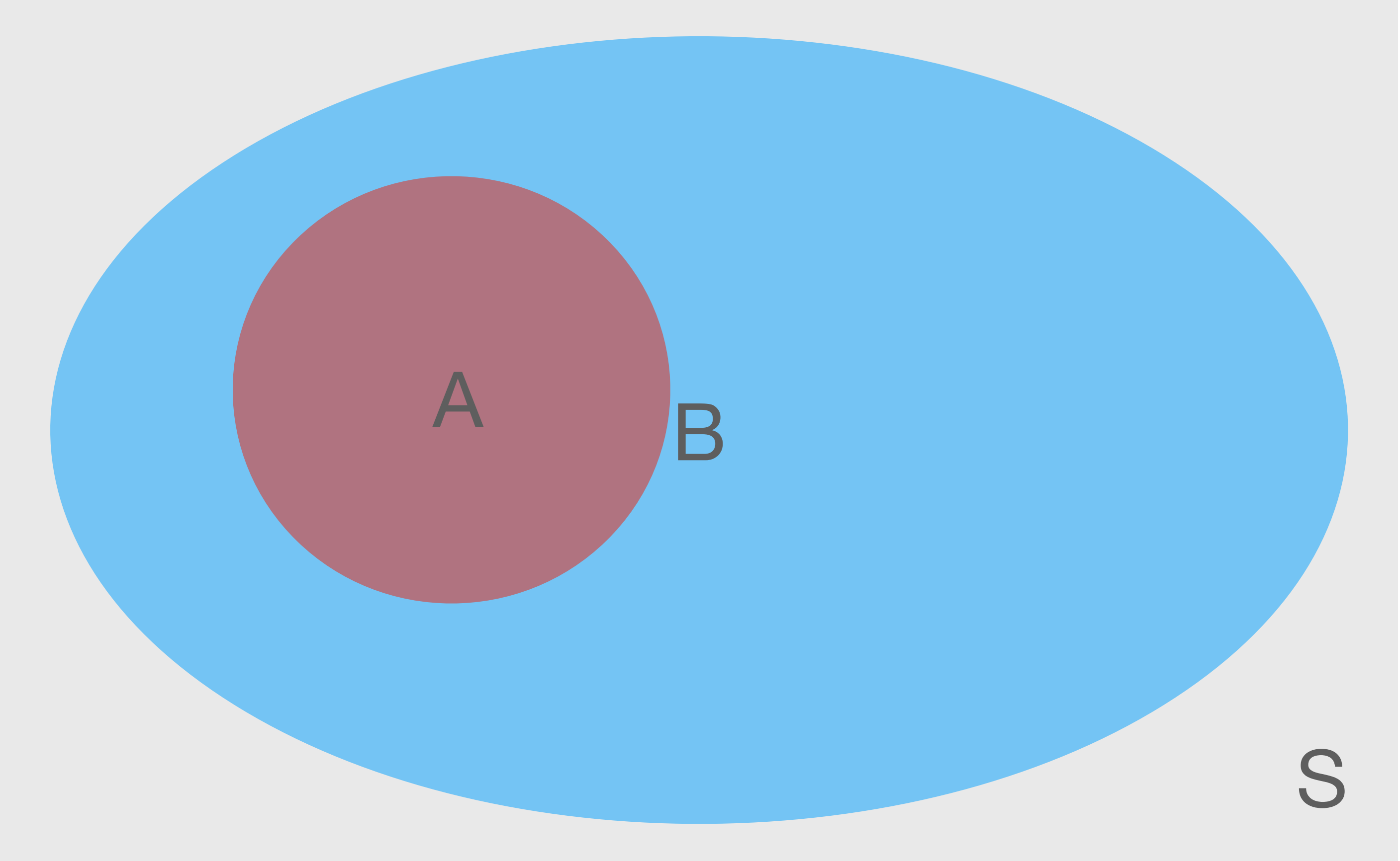
Probability Theory

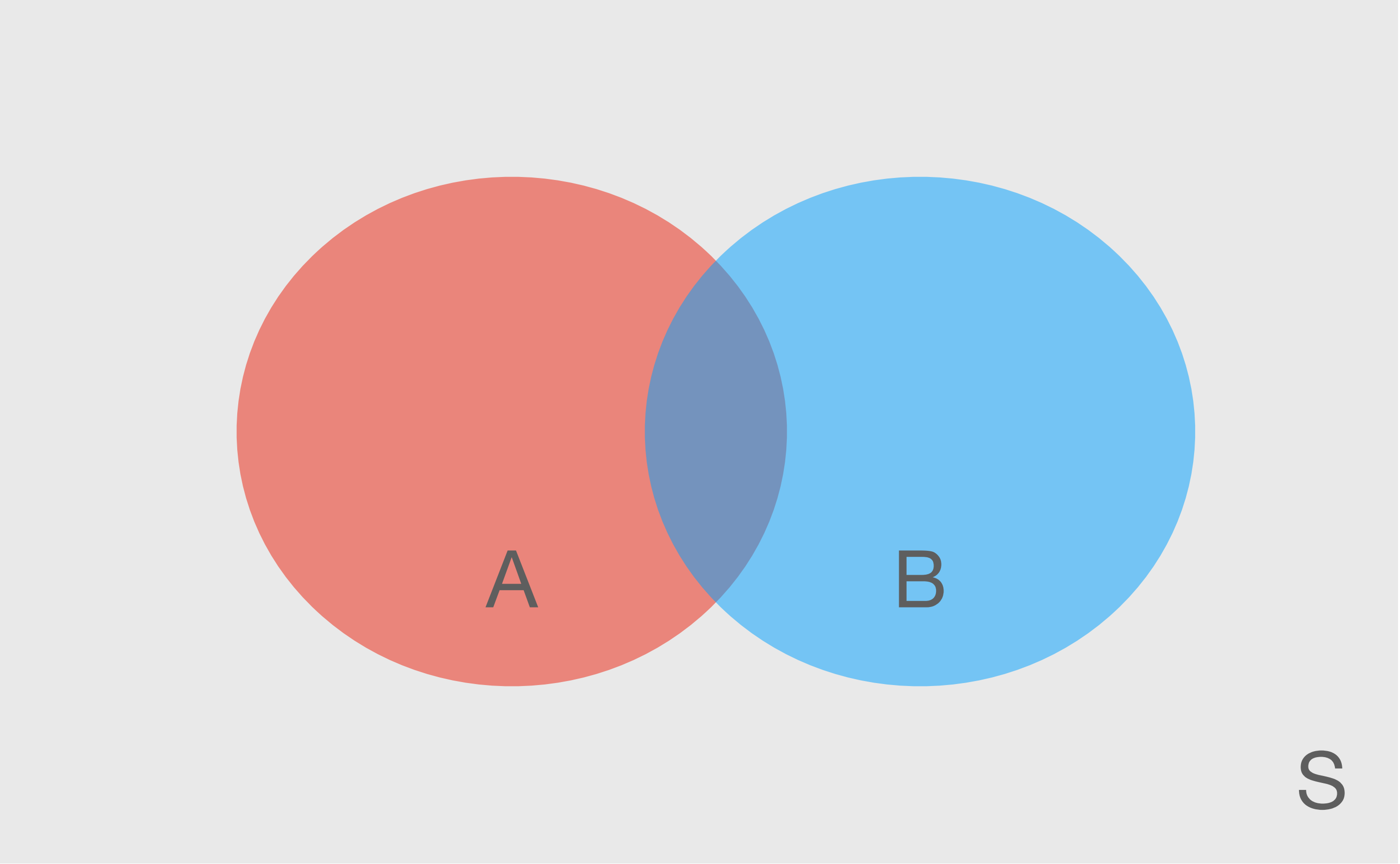
- Probability Theory → Analyze frequency of “events”
- Probability p of event x happening
  - → Repeatable observations of event
    - → p ~ fraction that x would be the outcome.
  - Example: Frequency of a disease in a population: 1 in 1000 → “Frequentist”
- What if you got a test → How do you interpret? You can’t repeat.
  - Accuracy of test
  - Frequency of disease
  - → Degree of belief → “Bayesian”

Basic Definitions

- Set S w subsets A and B
  - $P(S) = 1$
  - For all $\forall A : A \subset S \Rightarrow P(A) > 0$
  - $P(\bar{A}) = 1 - P(A)$
    - Bar means not in A.
    - $P(A \cup \bar{A}) = 1$
  - $P(\varnothing) = 0$
  - If there is no overlap between sets A and B
    - $A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$
    - $\cap$ is overlap
    - $\cup$ is union
  - $A \subset B \Rightarrow P(A) \leq P(B)$
    - $\subset$ means subset
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

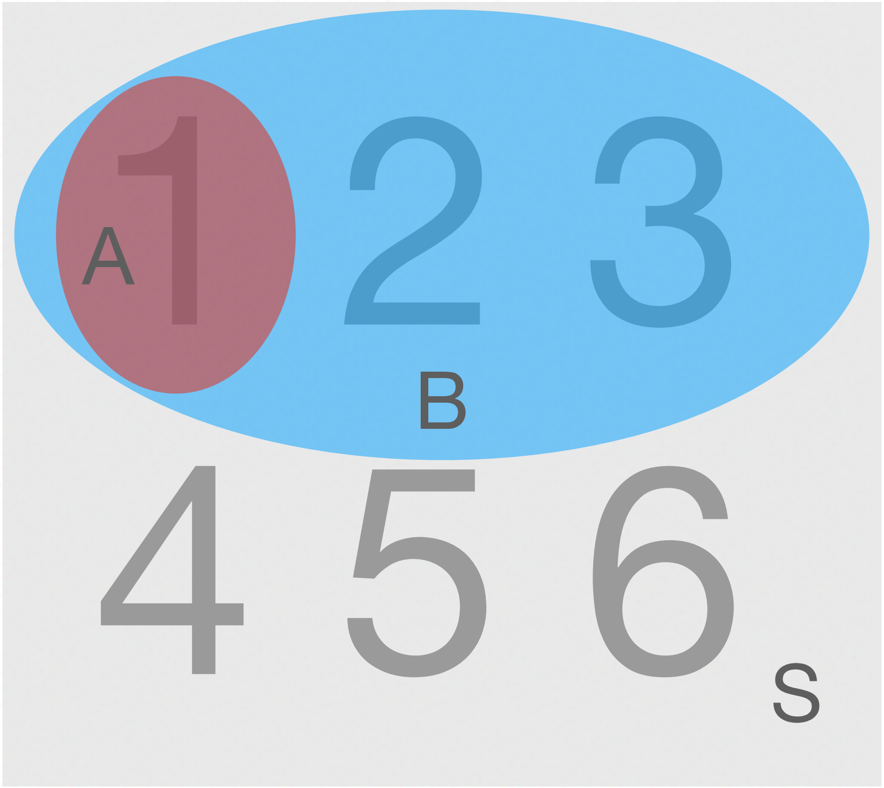






Conditional Probablilty

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Example:
  - Dice Roll (6-sided)
  - You are told rolled 3 or less
  - What is the Probability that you roled a 1?
  - $P(1) = \frac{P(<1 \text{ in } 6 \text{ rolls})}{P(<3 \text{ in } 6 \text{ rolls})} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$
- If $A \cup B = \varnothing \Rightarrow$
  - $P(A, B) = P(A)P(B)$
  - $P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$



Interpretation

- Relative Frequency:
  - A, B are outcomes of a repeatable experiment
  - $P(A) = \lim_{n \to \infty} \frac{\text{Times outcome is } A}{n}$
- Subjective Probabilty:
  - A, B are hypotheses (True/False) Statements.
  - $P(A)$ is degree of belief that A is true.

Bayes Theorem

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(B|A) = \frac{P(B \cap A)}{P(A)}$
- Since $P(A \cap B) = P(B \cap A)$
    - $\Rightarrow P(A|B)P(B) = P(B|A)P(A)$
    - $\Rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Bayes Example

Recall ⇒ $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

- Example: Disease Testing
  - Prior Knowledge about Population
    - $A := \{\text{sick}, \text{not sick}\}$
    - P(sick) = 0.001
    - P(not sick) = 0.999
  - Test
    - $B := \{+, -\}$
    - True Positive: P(+ | sick) = 0.98
    - False Negative: P(- | sick) = 0.02
    - False Positive: P(+ | not sick) = 0.03
    - True Negative: P(- | not sick) = 0.97
  - You get a postitive result -> what is the probablity that you are indeed sick?
    - P(sick | +) = $\frac{P(+|\text{sick})P(\text{sick})}{P(+|\text{sick})P(\text{sick})+P(+|\text{not sick})P(\text{not sick})}$
    - P(sick | +) = $\frac{(0.98)(0.001)}{(0.98)(0.001)+(0.03)(0.999)} = 0.032$
- Why: because of the prior.
- So why should I ever believe a test?
  - Because P(sick | symptoms) is high.

Data

- Use students as example
- Data can be viewed a table
  - Rows are students (data points)
  - Columns are features
- The features are Random Variables
- Make a distribution

Instance	Name	Age	Major	GPA	...
1	XXX	XXX	XXX	XXX	
2	XXX	XXX	XXX	XXX	
3	XXX	XXX	XXX	XXX	
...	XXX	XXX	XXX	XXX	