Probability and Statistics

Data Intelligence and Learning (DIAL) Lab
Prof. Jongwuk Lee

Probability Theory Basics

Why Study Probability Theory?

- **The world is full of uncertainty.**
  - Is the weather sunny tomorrow?
  - Is there a person in this image?
  - Is a user likely to prefer this movie?

- **We need to build a system that understands and interacts with uncertain real world.**

- **We often ask for the most likely explanation.**
  - Probability theory is nothing, but common sense reduced to calculation (Pierre Laplace, 1812).

Probability Space

- A probability space is a random process (or an experiment) with three components

$$(\Omega, \mathcal{F}, P)$$

- A sample space is the set of all possible outcomes.
- A set of all possible events, containing zero or more outcomes
- The assignment of probabilities to the event; P is a function from events to probabilities.

Sample Space

- The sample space Ω is **the set of all possible outcomes of an experiment.**

- Experiment: You rolled one die.

- What is Ω?
- $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Sample Space

- The sample space Ω is **the set of all possible outcomes of an experiment.**

- Experiment: You tossed two coins twice.
- What is Ω?
- $\Omega = \{HH, HT, TH, TT\}$
- The different elements of a sample space must be **mutually exclusive** and **collectively exhaustive**.

Events

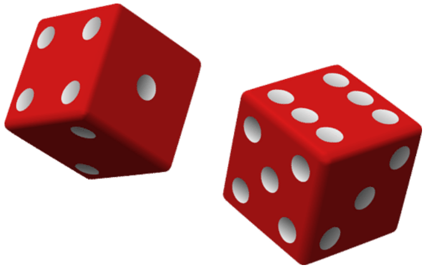
- An event is a set of outcomes of an experiment to which a probability is assigned.

- Experiment: You tossed two coins twice.
- Sample space
  - $\Omega = \{HH, HT, TH, TT\}$
- Event space
  - $\emptyset, \{HH\}, \{TT\}, \{HT\}, \{TH\}$
  - $\{HH, TT\}, \{HH, HT\}, \{HH, TH\}, \{TT, HT\}, \{TT, TH\}, \{HT, TH\}$
  - $\{HH, TT, HT\}, \{HH, TT, TH\}, \{HH, HT, TH\}, \{TT, HT, TH\}$
  - $\{HH, TT, HT, TH\}$

Experiments and Events

- Experiment: Tossing a coin twice.
- Event: You get two heads.

- Experiment: You throw two dice.



- Event: The sum of the rolls is six.
  - You got (1, 5), (2, 4), (3, 3), (4, 2) or (5, 1).
- Event: You get two odd faces.
  - You got (1, 1), (1, 3), (1, 5), (3, 1), (3, 3), (3, 5), …

Probability Measure Function

- With each event, we associate a **number** that measures the **probability**, i.e., $P: \mathcal{F} \rightarrow [0, 1]$.

Impossible | Even chance | Certain

Summary: Sample Space and Events

- **Sample space Ω**
  - The set of all possible outcomes of the experiment
    - If you toss a coin twice, Ω = {HH, HT, TH, TT}.
  - The number of possible outcomes $|\Omega| = N$

- **Event space $\mathcal{F}$**
  - The space of potential results of the experiment
  - $\mathcal{F}$ is often the powerset of Ω, i.e., $|2^N|$.

- **Let $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, \boldsymbol{P})$ be a probability space with sample space $\boldsymbol{\Omega}$, event space $\boldsymbol{\mathcal{F}}$ and probability measure function $\boldsymbol{P}$.**

Probability Axioms

- The probability law assigns to an event E a **non-negative number** $P(E)$ which encodes our **belief/knowledge** about the **likelihood** of the event E.

- **Nonnegativity**: $P(E_i) \geq 0$, for every event E_i.
- **Normalization**: The probability of the sample space Ω is equal to 1, i.e., $P(\Omega) = 1$.
- **Additivity**: If E_1 and E_2 are two disjoint events, the probability of their union satisfies $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.
  - It extends to the union of infinitely many disjoint events,

$$\boldsymbol{P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots}$$

Probability Axioms

- The probability of an event is a **non-negative real number.**

$$\boldsymbol{P(E_i) \in \mathbb{R}, P(E_i) \geq 0, \forall E_i \in \mathcal{F}}$$

- **Total probability** over all outcomes must be 1.

$$\boldsymbol{P(\Omega) = 1}$$

- **Additivity of disjoint** (or mutual exclusive) **events**:

$$\boldsymbol{P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)}$$

Simple Consequences of the Axioms

- $\boldsymbol{P(A) + P(A^c) = 1}$

- $\boldsymbol{A, B, C}$ **are disjoint:** $\boldsymbol{P(A \cup B \cup C) = P(A) + P(B) + P(C)}$

$$\Downarrow$$

- **For** k **disjoint events,**
$P(\{\boldsymbol{E_1}, \boldsymbol{E_2}, \ldots, \boldsymbol{E_k}\}) = \boldsymbol{P(\{E_1\}) + P(\{E_2\}) + \cdots + P(\{E_k\})}$

More Consequences of the Axioms

- If $A \subset B$, then $P(A) \leq P(B)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- $P(A \cup B) \leq P(A) + P(B)$

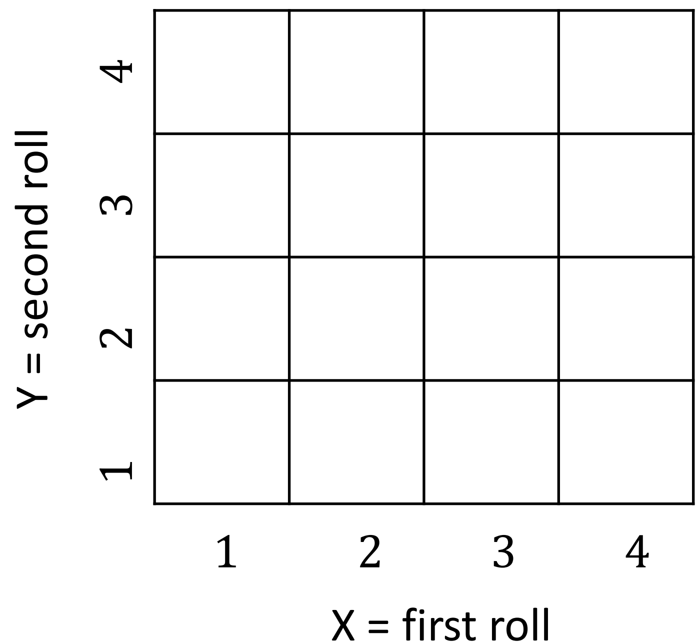
- $P(A \cup B \cup C) = P(A) + P\left(A^C \cap B\right) + P\left(A^C \cap B^C \cap C\right)$

Example: Discrete Case

- Rolling two 4-side dice

- Let every possible outcome have probability $1/16$.
  - $P(X = 1) =$

- Let $Z = \min(X, Y)$.
  - $P(Z = 4) =$
  - $P(Z = 2) =$

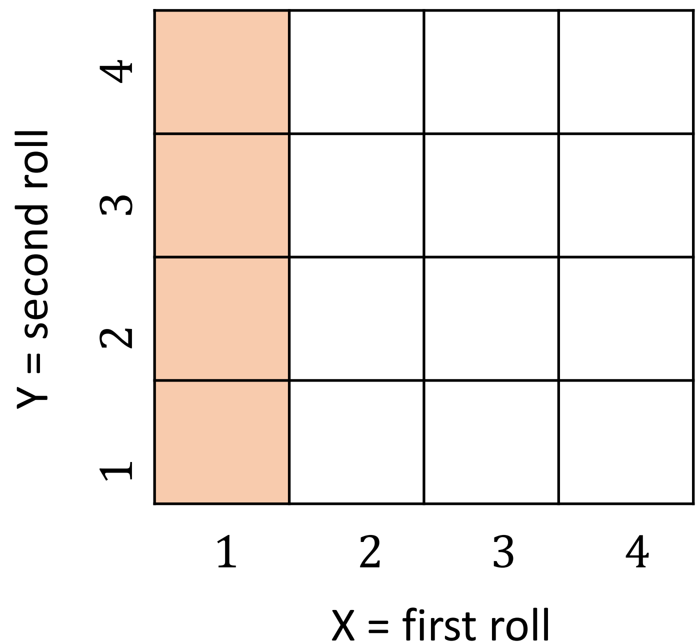


Example: Discrete Case

- **Rolling two 4-side dice**

- **Let every possible outcome have probability $1/16$.**
  - $P(X = 1) = 4/16$

- **Let $Z = \min(X, Y)$.**
  - $P(Z = 4) =$
  - $P(Z = 2) =$

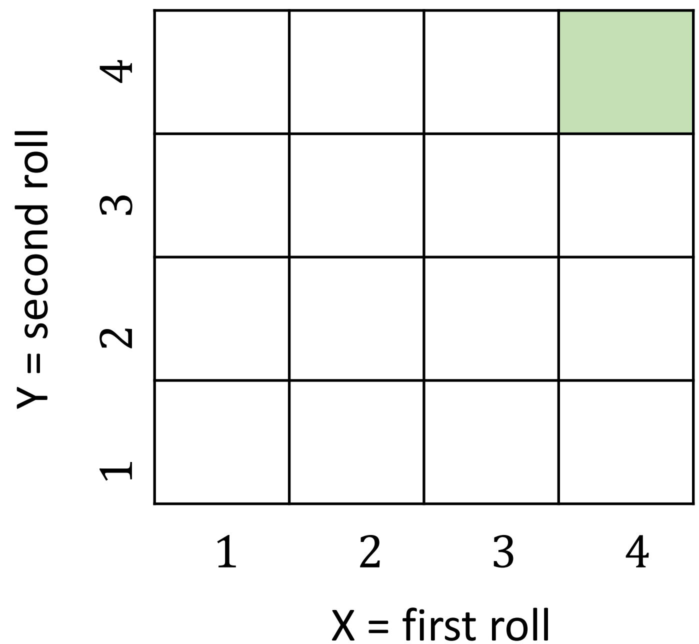


Example: Discrete Case

- **Rolling two 4-side dice**

- **Let every possible outcome have probability $1/16$.**
  - $P(X = 1) = 4/16$

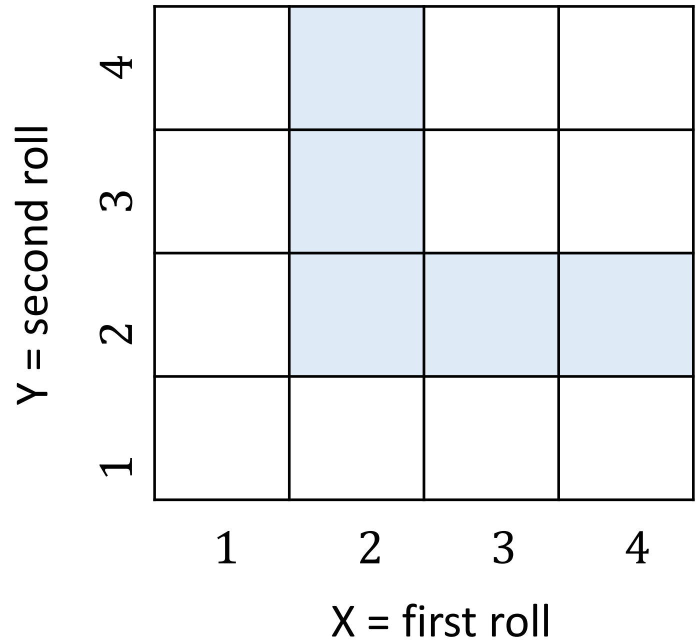
- **Let $Z = \min(X, Y)$.**
  - $P(Z = 4) = 1/16$
  - $P(Z = 2) =$



Example: Discrete Case

- **Rolling two 4-side dice**

- **Let every possible outcome have probability $1/16$.**
  - $P(X = 1) = 4/16$

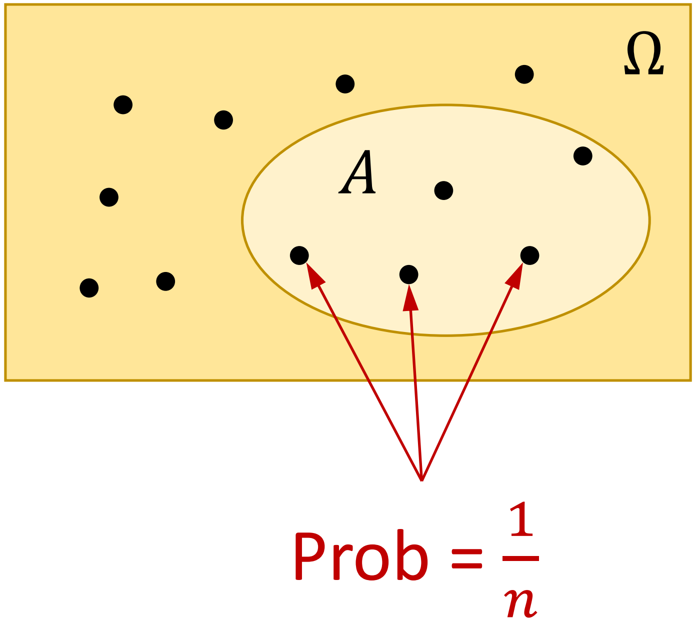


- **Let $Z = \min(X, Y)$.**
  - $P(Z = 4) = 1/16$
  - $P(Z = 2) = 5/16$

Discrete Uniform Law

- Assume Ω consists of n **equally** likely elements.
- Assume A consists of k elements.

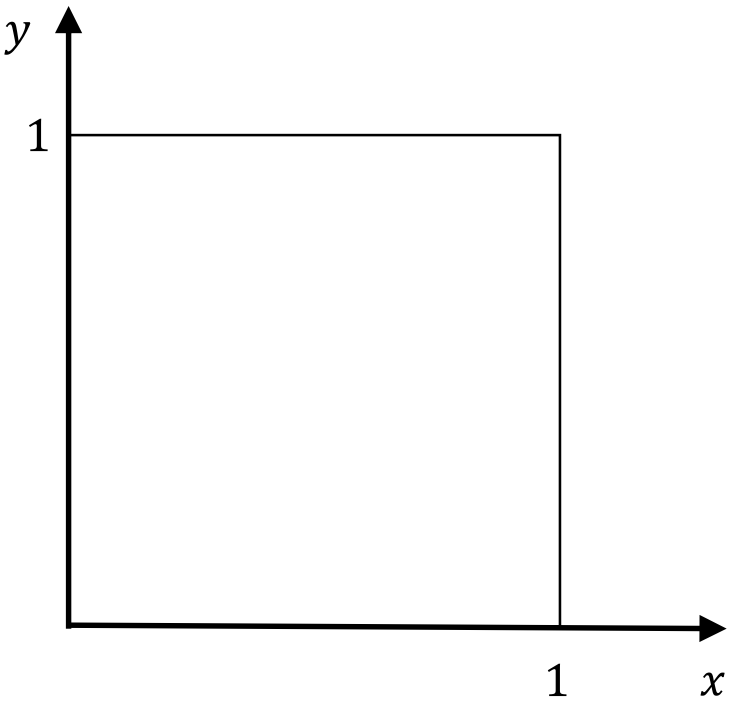
$$P(A) = k \cdot \frac{1}{n} = \frac{k}{n}$$



Example: Continuous Case

- **Uniform probability law:** **Probability = Area**

- (x, y) **such that** $\mathbf{0 \leq x, y \leq 1}$



$P(\{(x, y) \mid x + y \leq 1/2\}) =$

$P(\{(0.5, 0.3)\}) =$

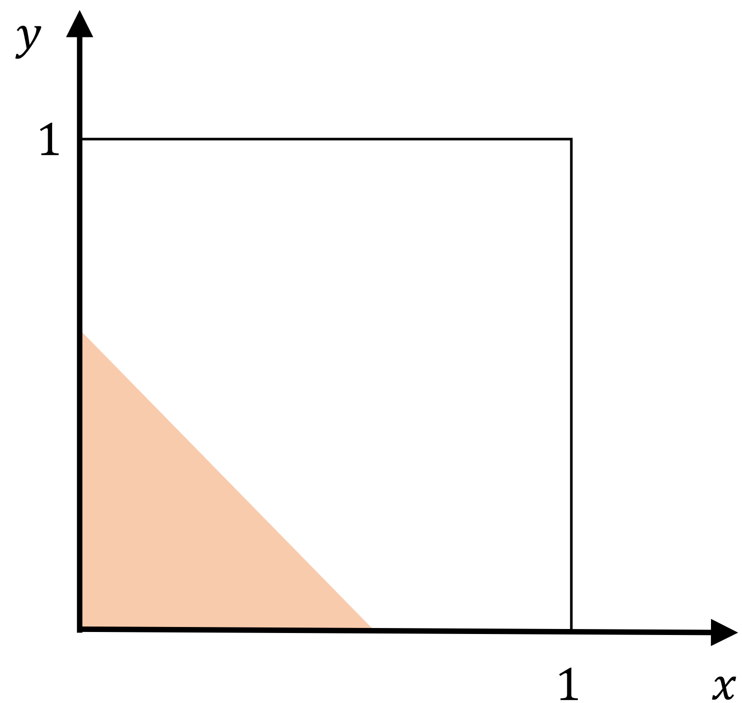
Example: Continuous Case

- Uniform probability law: **Probability = Area**

- (x, y) such that $\mathbf{0 \leq x, y \leq 1}$

$$P(\{(x,y) \mid x + y \leq 1/2\})$$
$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

$$P(\{(0.5,0.3)\}) =$$

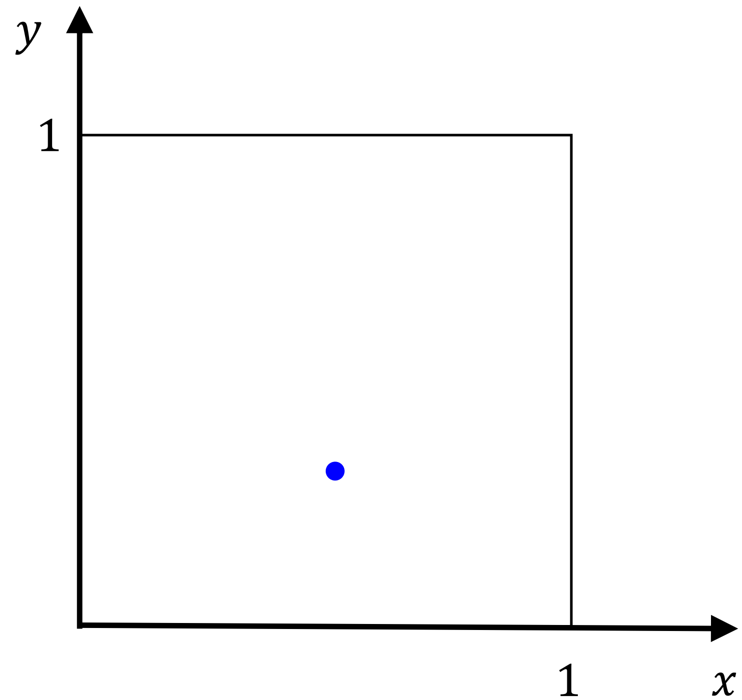


Example: Continuous Case

- Uniform probability law: **Probability = Area**
- (x, y) such that $\mathbf{0 \leq x, y \leq 1}$

$$P(\{(x, y) \mid x + y \leq 1/2\})$$
$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

$$P(\{(0.5, 0.3)\}) = 0$$



Probability Calculation Steps

- Specify the sample space.
- Specify a probability law.
- Identify an event of interest.
- Calculate ...

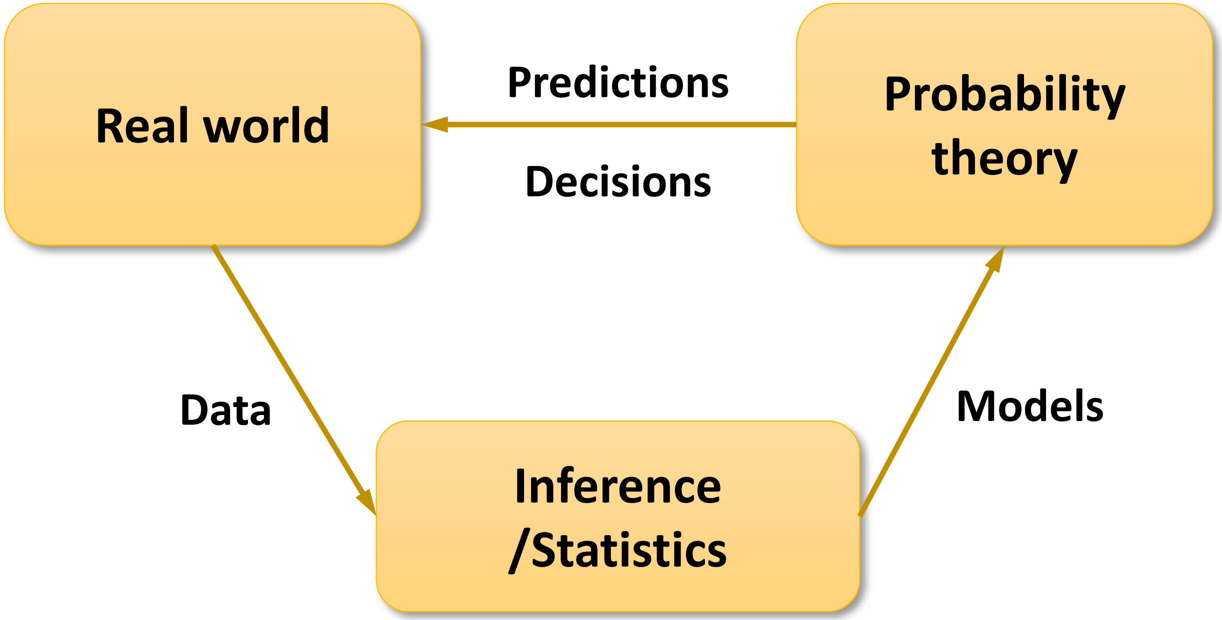
Interpretation of Probability Theory

- **A probability of 1 means it is certain.**
- **A probability of 0 means it is impossible.**

- **It is the study of uncertainty.**
  - **Relative frequency**: The faction of times an event occurs
    - If I were to toss the coin 10 times, roughly 5 times I would see a head.
  - **Belief**: A degree of belief about an event
    - The sun rises in the **east**. *vs.* The sun rises in the **west**.

The Role of Probability Theory

- **A framework for analyzing phenomena with uncertain outcomes**
  - Rules or consistent reasoning
  - Used for predictions and decisions



Conditional Probability

Motivation: Partial Information

- **We have assumed we know nothing about the outcome of our experiment.**

- **Sometimes, we have partial information that may affect the likelihood of a given event.**
  - Experiment: you **roll a die**.
  - Partial information: you are told that **the number is odd**.

  - Experiment: we predict the **weather tomorrow**.
  - Partial information: we know that the **weather today** is **rainy**.

Incorporating Partial Information

- **Knowing about event B (e.g., "it is raining today") changes our beliefs about event A (e.g., "will it rain tomorrow?").**

- **How to update our probability law to incorporate this new knowledge?**

- **Introduce a conditional probability.**

What is Conditional Probability?

➢ Original problem

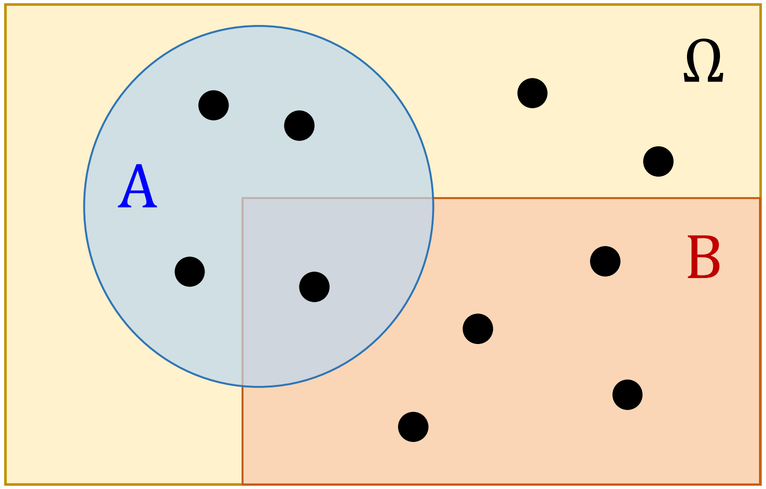
- What is the probability of some event A?
  - What is the probability that we roll a number less than 4?
- This is given by our probability law.

➢ New problem

- **Given event $\boldsymbol{B}$**, what is the probability of event A?
  - Given that the number rolled is an odd number, what is the probability that it is less than 4?
- We call this the **conditional distribution of $\boldsymbol{A}$ given $\boldsymbol{B}$**.
- We write this as $\boldsymbol{P(A \mid B)}$.
  - Read $|$ as **given** or **conditioned on the fact that**.
- Our **conditional probability** is still describing "**the probability of something**", so we expect it to behave like a **probability distribution**.

Idea of Conditioning

- $P(A \mid B)$ = "Probability of A, given that B occurred"



Usually, Ω is ignored.

$$P(A \mid \Omega) = \frac{P(A \cap \Omega)}{P(\Omega)}$$

$$\Downarrow$$

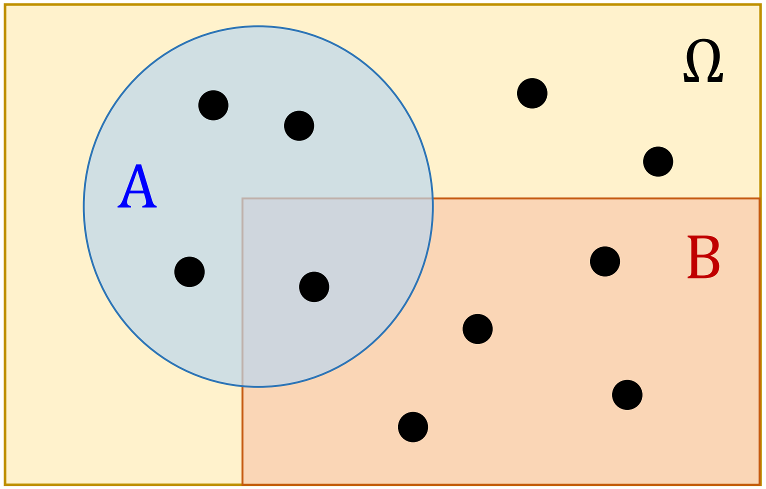
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

defined only if $P(B) > 0$

Idea of Conditioning

- Use **new information** to revise a model.

- If B occurred,

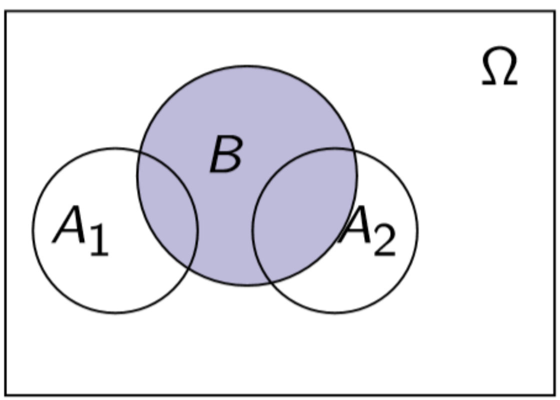


$$P(A \mid B) = \frac{1}{5}$$

$$P(B \mid B) = \frac{5}{5}$$

Conditional Probability Axioms

- **Suppose that our new universe is B instead of Ω.**
  - **Nonnegativity**: $P(A_i \mid B) \geq 0$ assuming $P(B) > 0$.
  - **Normalization**: We know $P(B \mid B) = 1$.
  - **Additivity**: $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B)$ for two disjoint sets A_1 and A_2.

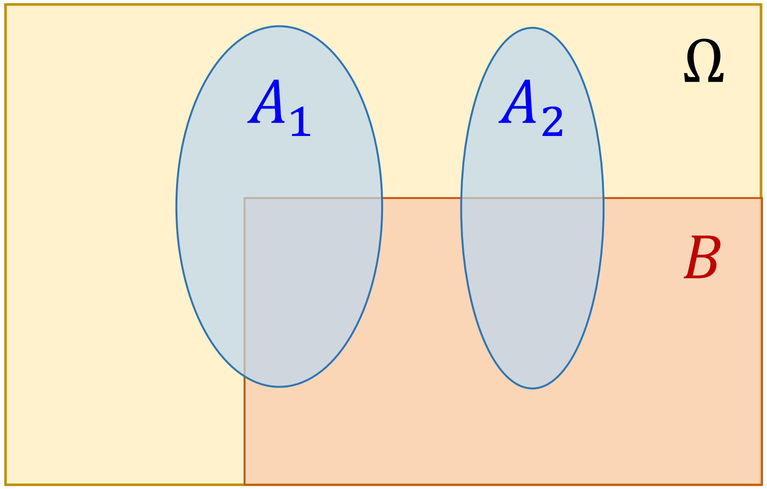


Conditioning on B $\longrightarrow$

Properties of Conditional Probability

- **If $P(B) > 0$,**
  - If $A_1 \subseteq A_2$, then $P(A_1 \mid B) \leq P(A_2 \mid B)$.
  - If A_i for $i \in \{1, \dots, n\}$ are all pairwise **disjoint**, then

$$P\left(\bigcup_{i=1}^{n} A_i \mid B\right) = \sum_{i=1}^{n} P(A_i \mid B)$$

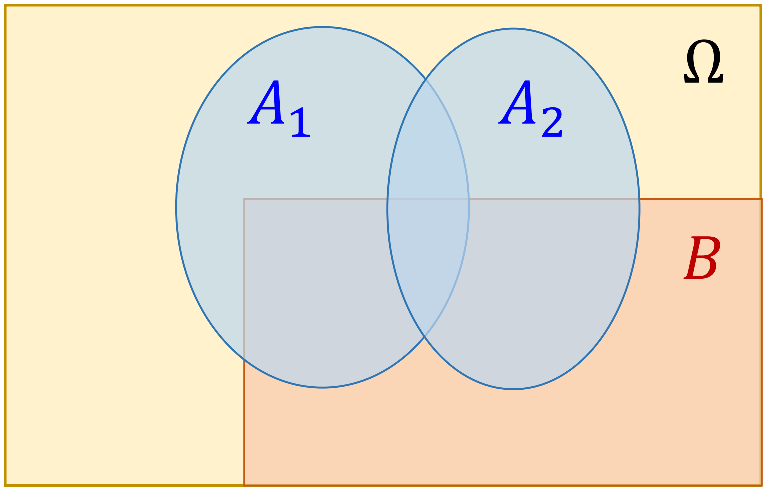


Properties of Conditional Probability

- **If $P(B) > 0$,**
  - $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) - P(A_1 \cap A_2 \mid B)$

- **Union bound:** $P(A_1 \cup A_2 \mid B) \leq P(A_1 \mid B) + P(A_2 \mid B)$

$$\boldsymbol{P\left(\bigcup_{i=1}^{n} A_i \mid B\right) \leq \sum_{i=1}^{n} P(A_i \mid B)}$$



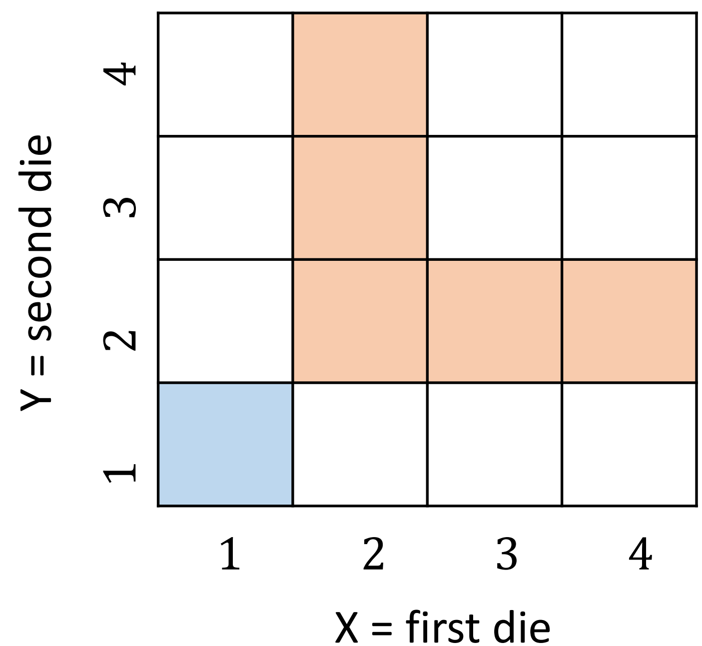
Example: Coin Tossing

- **Consider the experiment of tossing a fair coin three times. What is the probability of getting alternating heads and tails conditioned on the event that the first toss gives a head?**
- **Notation**
  - A = {Tosses yield alternating tails and heads.}
  - B = {The first toss is a head.}

- **How to compute $P(A \mid B)$?**
  - Sample space: {HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}.
  - A = {HTH,THT}, B = {HHH,HHT,HTH,HTT} and $A \cap B$ = {HTH}.
  - Our new sample space is B = {HHT,HTH,HTT,HHH}.
  - Each of these are equally likely. Out of these 1 event satisfies alternating heads and tails. So, $P(A|B) = 1/4$

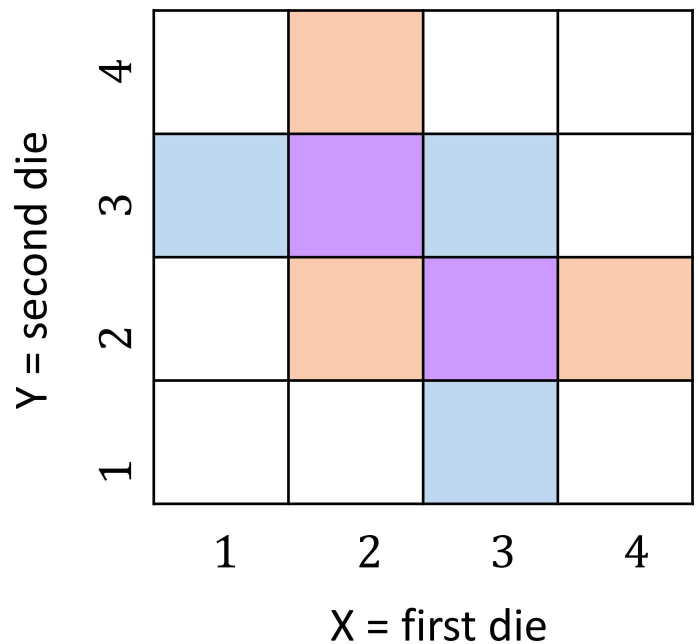
Example: Rolling a Die Twice

- **Rolling a 4-side die twice**
- **Let** B **be the event:** $\min(X, Y) = 2$
- **Let** A **be the event:** $max(X, Y)$
- $\boldsymbol{P(A = 1 \mid B) = 0\ /\ 5}$



Example: Rolling a Die Twice

- **Rolling a 4-side die twice**
- **Let B be the event:** $\min(X, Y) = 2$
- **Let A be the event:** $max(X, Y)$
- $\boldsymbol{P(A = 3 \mid B) = 2\ /\ 5}$



Example: Umbrella Sales

- Alice works for an umbrella company.
- If it is raining, the probability that she sells more than 10 umbrellas is 0.8.
- If it's not raining, the probability that she sells more than 10 umbrellas is 0.25.
- The probability that it rains tomorrow is 0.1.

- What is the probability that it doesn't rain tomorrow and she sells more than 10 umbrellas?

Example: Umbrella Sales

- Let S = {# of umbrella sold > 10} and R = { it is rainy.}.

- If it is raining, the probability that she sells more than 10 umbrellas is 0.8. $\Rightarrow$ $\boldsymbol{P(S \mid R) = 0.8}$
- If it's not raining, the probability that she sells more than 10 umbrellas is 0.25. $\Rightarrow$ $\boldsymbol{P(S \mid R^C) = 0.25}$
- The probability that it rains tomorrow is 0.1. $\Rightarrow$ $\boldsymbol{P(R) = 0.1}$

- What is the probability that it doesn't rain tomorrow and she sells more than 10 umbrellas? $\Rightarrow$ $\boldsymbol{P\left(S \cap R^C\right)} = ??$

Example: Umbrella Sales

- What is the probability that it doesn't rain tomorrow, and she sells more than 10 umbrellas? $\Rightarrow P(S \cap R^C) = ??$

- We can rearrange our formula for conditional probability.

$$P(S \mid R^C) = \frac{P(S \cap R^C)}{P(R^C)}$$

$$\Downarrow$$

$$P(S \cap R^C) = P(S \mid R^C)P(R^C)$$

Example: Umbrella Sales

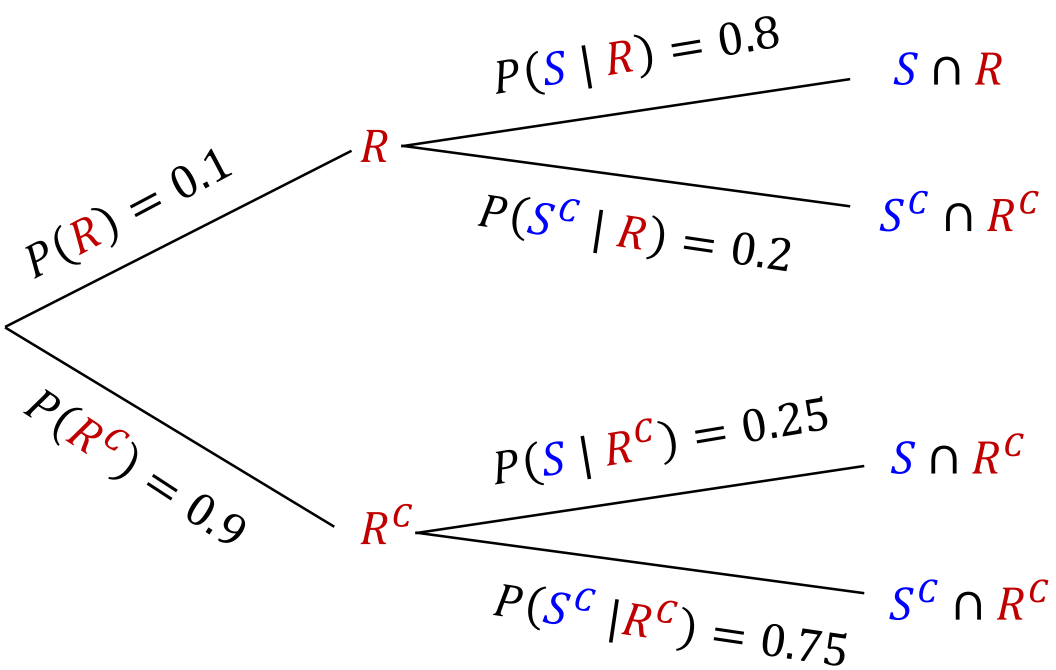
- What is the probability that it doesn't rain tomorrow and she sells more than 10 umbrellas? $\Rightarrow$ $\boldsymbol{P(S \cap R^C) = ??}$

$$\boldsymbol{P(S \cap R^C) = P(S \mid R^C)P(R^C)}$$

- We compute $\boldsymbol{P(S \mid R^C)P(R^C) = 0.25 \times 0.9 = 0.225}$.

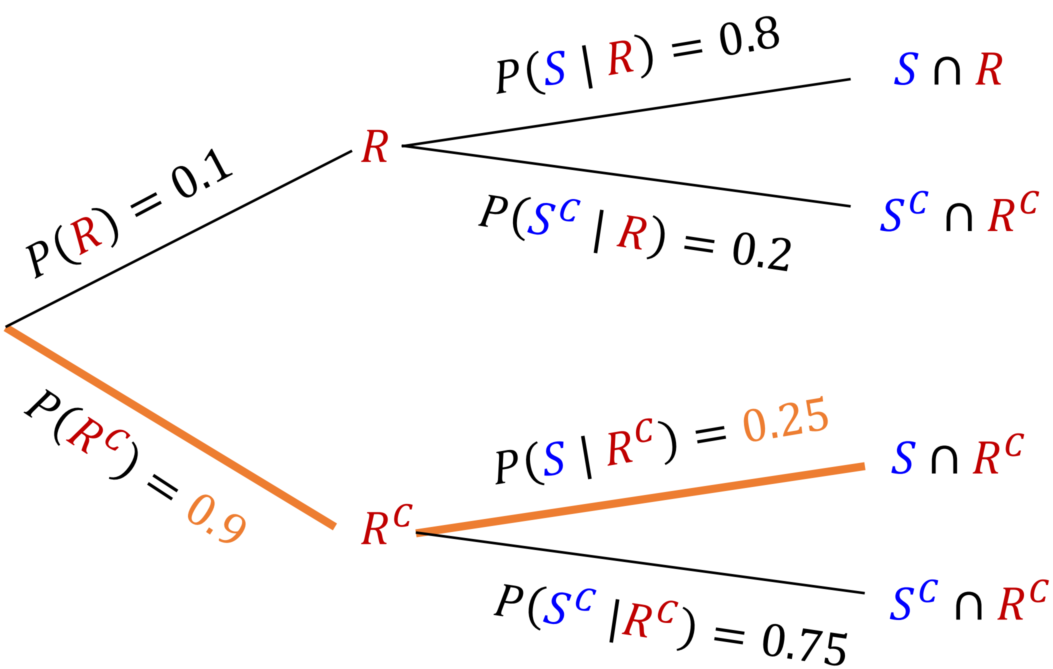
Representing Conditional Probabilities

- We represent conditional probabilities using a **tree structure**.
- The probability at a leaf node means the product of the probabilities along each path.



Representing Conditional Probabilities

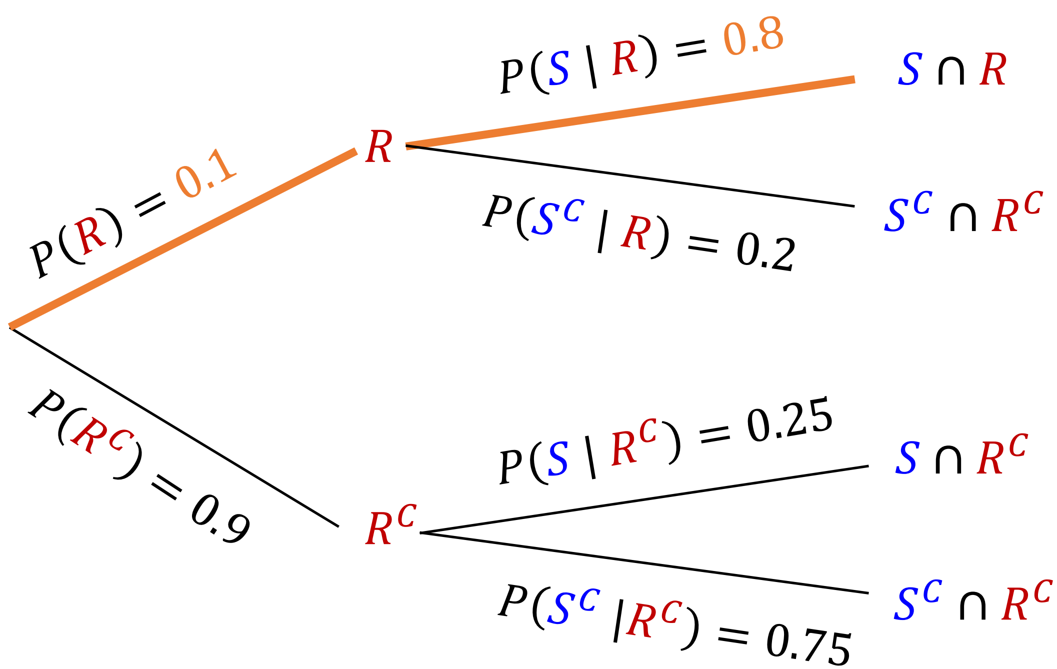
- We represent conditional probabilities using a **tree structure**.
- The probability at a leaf node means the product of the probabilities along each path.



$$P(S \cap R^C) = 0.9 \times 0.25 = 0.225$$

Representing Conditional Probabilities

- We represent conditional probabilities using a **tree structure**.
- The probability at a leaf node means the product of the probabilities along each path.

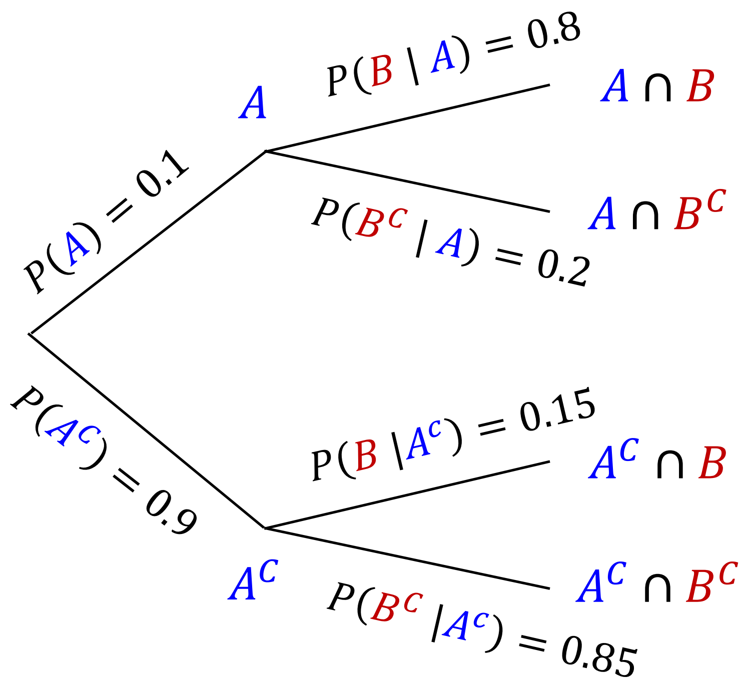


$P(S \cap R)$
$= 0.1 \times 0.8 = 0.08$

Representing Conditional Probabilities

- Event A: Airplane is flying above.
- Event B: Something registers on the radar screen.

- $P(A \cap B) =$
- $P(B) =$
- $P(A \mid B) =$



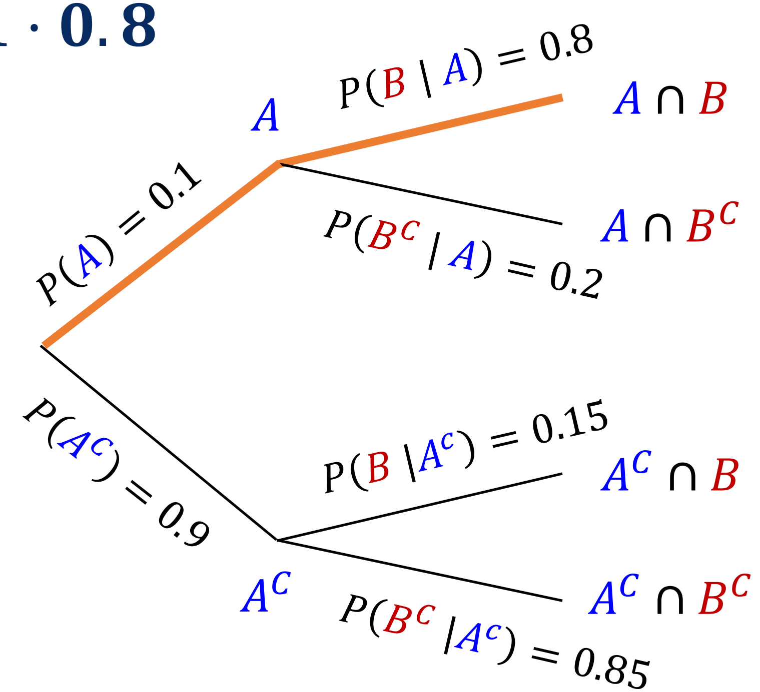
Representing Conditional Probabilities

- Event $\boldsymbol{A}$: Airplane is flying above.
- Event $\boldsymbol{B}$: Something registers on the radar screen.

- $\boldsymbol{P(A \cap B) = P(A)P(B \mid A) = 0.1 \cdot 0.8}$

- $\boldsymbol{P(B) =}$

- $\boldsymbol{P(A \mid B) =}$

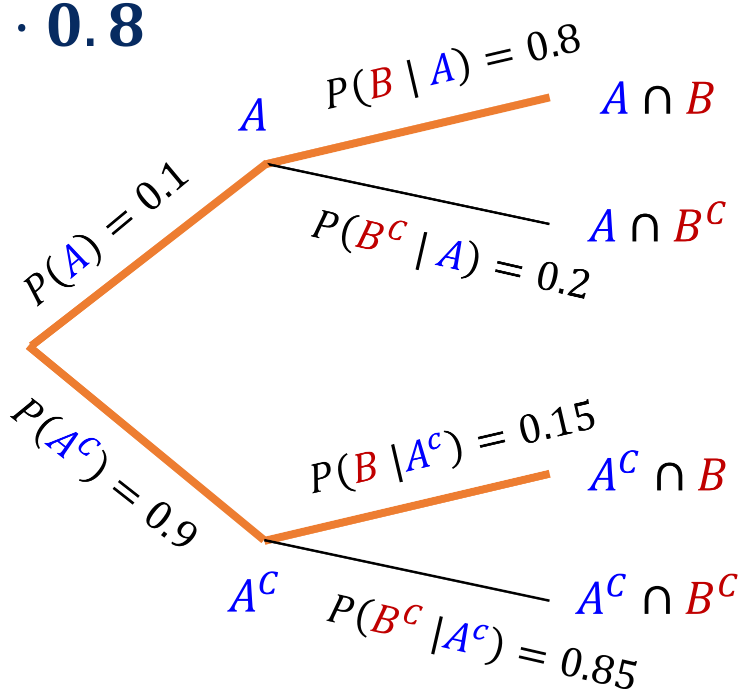


Representing Conditional Probabilities

- Event A: Airplane is flying above.
- Event B: Something registers on the radar screen.

- $P(A \cap B) = P(A)P(B \mid A) = 0.1 \cdot 0.8$

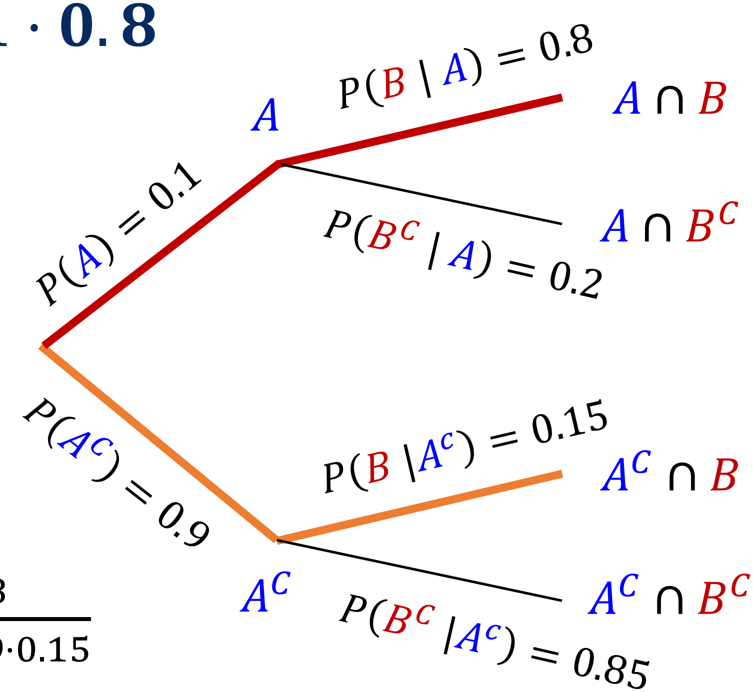
- $P(B) =$
  - $P(A)P(A \cap B) + P(A^C)P(A^C \cap B)$
  - $= \; 0.1 \cdot 0.8 + 0.9 \cdot 0.15$



Representing Conditional Probabilities

- Event A: Airplane is flying above.
- Event B: Something registers on the radar screen.

- $P(A \cap B) = P(A)P(B \mid A) = 0.1 \cdot 0.8$



- $P(B) =$
  - $P(A)P(A \cap B) + P(A^C)P(A^C \cap B)$
  - $= \; 0.1 \cdot 0.8 + 0.9 \cdot 0.15$

- $P(A \mid B) =$
  - $\frac{P(A \cap B)}{P(A)P(A \cap B) + P\left(A^C\right)P\left(A^C \cap B\right)} = \frac{0.1 \cdot 0.8}{0.1 \cdot 0.8 + 0.9 \cdot 0.15}$

Multiplication Rule

- **We know that $P(A \cap B) = P(A|B)P(B)$.**

- **What is $\boldsymbol{P(A \cap B \cap C)}$?**
  - Treat $(B \cap C)$ as an event. Call this R.

- $\boldsymbol{P(A \cap B \cap C) = P(A \cap R)}$
  - So, $P(A \cap R) = P(A|R)P(R) = P(A|B \cap C)P(B \cap C)$.
  - $P(R) = P(B \cap C) = P(B|C)P(C)$.

Multiplication Rule

➢ **Using induction, you can prove that:**

$$P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$$

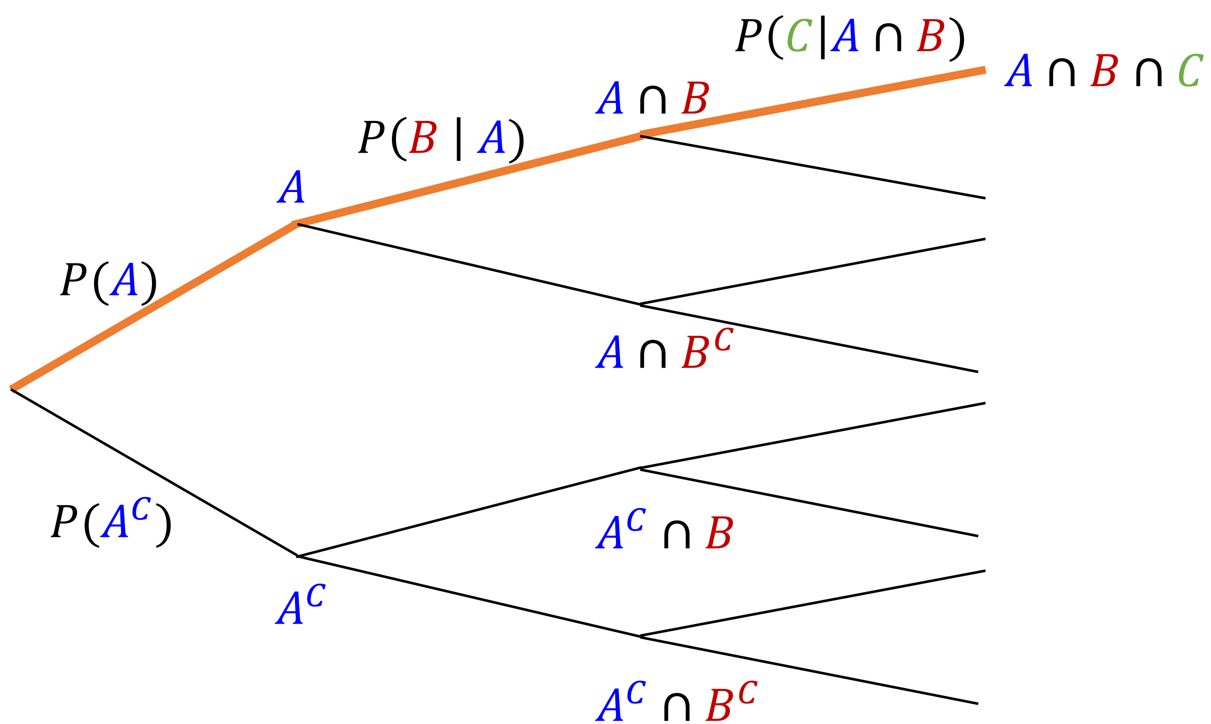
$$\Downarrow$$

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1 \mid A_2 \cap \cdots \cap A_n) \cdots P(A_{n-1} \mid A_n)P(A_n)$$

Multiplication Rule

$$P(C \mid A \cap B) = \frac{P(A \cap B \cap C)}{P(A \cap B)} \Rightarrow P(A \cap B \cap C) = P(C \mid A \cap B)P(A \cap B)$$

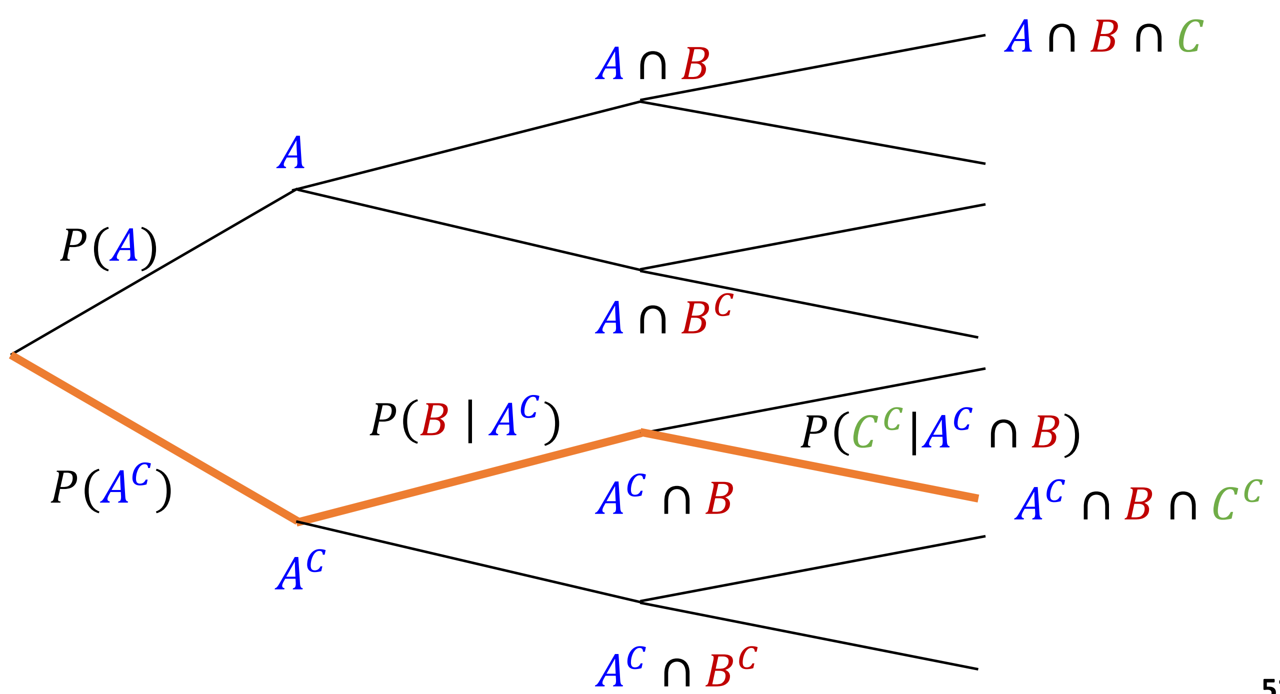
$$= P(C \mid A \cap B)P(B \mid A)P(A)$$



Multiplication Rule

$$P\left(A^C \cap B \cap C^C\right) = P\left(C^C \middle| A^C \cap B\right) P\left(A^C \cap B\right)$$

$$= P\left(C^C \middle| A^C \cap B\right) P\left(B \mid A^C\right) P\left(A^C\right)$$



Example: Umbrella Sales

- Alice works for an umbrella company.
- We have $P(R) = 0.1$, $P(S \mid R) = 0.8$ and $P\left(S \mid R^C\right) = 0.25$
- If we knew that Alice sells more than 10 umbrellas, then what is the probability it rained?

Example: Umbrella Sales

- Alice works for an umbrella company.
- We have $P(R) = 0.1$, $P(S \mid R) = 0.8$ and $P\left(S \mid R^C\right) = 0.25$

- If we knew that Alice sells more than 10 umbrellas, then what is the probability it rained?

- We are interested in $P(R \mid S)$. First, we need $P(R \cap S)$ and then we need $P(S)$.
- $P(R \cap S) = P(S|R)P(R) = 0.8 \times 0.1 = 0.08$.

Example: Umbrella Sales

- **Alice works for an umbrella company.**
- **We have $P(R) = 0.1$, $P(S \mid R) = 0.8$ and $P\left(S \mid R^C\right) = 0.25$**

- **Now, what about $P(S)$?**
- **Write S as a union of two disjoint events. Guesses?**
  - $S = S \cap \Omega = S \cap (R \cup R^C) = (S \cap R) \cup (S \cap R^C)$.

- $P(S) = P(S \cap R) + P(S \cap R^C)$. **Theorem of total probability**
- $P(S) = P(S \mid R)P(R) + P(S \mid R^C)P(R^C)$
- $P(S) = 0.8 \times 0.1 + 0.25 \times 0.9 = 0.305$

Example: Umbrella Sales

- Alice works for an umbrella company.
- We have $P(R) = 0.1$, $P(S \mid R) = 0.8$ and $P\left(S \mid R^C\right) = 0.25$
- If we knew that Alice sells more than 10 umbrellas, then what is the probability it rained?
- $P(R \mid S) = P(R \cap S)/P(S)$
- $P(R \mid S) = P(S|R)P(R)/(P(S \mid R)P(R) + P(S \mid R^C)P(R^C))$
- $P(R \mid S) = 0.08/0.305 \approx 0.262$
- This is known as **Bayes' rule**.

Example: Card Decks

- Three cards are drawn from an ordinary 52-card deck **without replacement**.
  - **Without replacement**: Drawn cards are not placed back into the deck.

- What is the probability that there is **no heart** among the three?

Example: Card Decks

- **Three cards are drawn from an ordinary 52-card deck without replacement.**
  - **Without replacement**: Drawn cards are not placed back into the deck.

- **What is the probability that there is no heart among the three?**
- **Notation: A_i = {i-th card is not a heart}**
- **We want: $P(A_1 \cap A_2 \cap A_3)$.**
  - Remember: There are **thirteen** cards with hearts.
- **Use multiplication rule:**
  - $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$.

Example: Card Decks

- **Three cards are drawn from an ordinary 52-card deck without replacement.**
  - **Without replacement**: Drawn cards are not placed back into the deck.

- **What is the probability that there is no heart among the three?**
- **Notation: A_i = {i-th card is not a heart}**
- **Use multiplication rule:**
  - $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2).$
  - $P(A_1) = \frac{39}{52}, P(A_2|A_1) = \frac{38}{51}, P(A_3|A_1 \cap A_2) = \frac{37}{50}.$

Example: Three Balls

- I have two blue balls, and one red ball. I pick two balls randomly without replacement.

- What is the probability that the first ball is blue?
- What is the probability that the second ball is blue?

Example: Three Balls

- I have two blue balls, and one red ball. I pick two balls randomly without replacement.

- What is the probability that the first ball is blue?
- What is the probability that the second ball is blue?

- Notation: X_i is color of the i-th ball.
- We want $P(X_1 = B) = \mathbf{2/3}$.
- We want $P(X_2 = B) = \mathbf{2/3}$.
  - $P(X_\mathbf{2} = B) = P(X_2 = B \cap X_1 = B) + P(X_2 = B \cap X_1 = R)$
  - $P(X_\mathbf{2} = B) = P(X_2 = B \mid X_1 = B)P(X_1 = B) + P(X_2 = B \mid X_1 = R)P(X_1 = R) = 1/2 \times 2/3 + 1 \times 1/3 = \mathbf{2/3}$

Bayes’ Theorem

Bayes' Theorem

- A simple rule to get conditional probability of A given B, from the conditional formula of B given A

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

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

- It is useful for inferring hidden causes from our observation.

Typical Bayes Rule Example

- Considering testing for **some latent (hidden/unobservable) disease**, it will not be symptomatic until a future time point.
- We can directly observe **the outcome of the test**.
- Assuming the test is not 100% accurate, we cannot directly observe whether we have the disease.

- **Two possible hidden causes** for a **positive test result.**
  - We have the disease, and the test is correct.
  - We don't have the disease, and the test is a false positive.

- **Inferring which hidden cause underlies our observation**

Example: Disease Testing

- **Assume that the disease affects 2% of the population.**
  - The false positive rate is **1%**.
  - The false negative rate is **5%**.
  - We take the test, and the result is **positive**.

- **Given that you tested positive, what is the probability you have the disease?**

Example: Disease Testing

- **Assume that the disease affects 2% of the population.**
  - The false positive rate is **1%**.
  - The false negative rate is **5%**.
  - We take the test, and the result is **positive**.

- **Given that you tested positive, what is the probability you have the disease?**

- **Let T be the event "tests positive" and D be the event "has disease."**
  - $P(D) = 0.02,\ P(T \mid D^C) = 0.01,\ P(T^C \mid D) = 0.05$

Example: Disease Testing

- Given that you tested positive, what is the probability you have the disease?

- What is $P(D \mid T)$? Bayes' rule gives us:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D^C)P(D^C)}$$

- We get from the **conditional probability of an observation given a hidden cause** (which we usually know) to **the conditional probability of a hidden cause given an observation** (which we usually care about!)

Example: Disease Testing

- What is $P(D \mid T)$? Bayes' rule gives us:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D^C)P(D^C)}$$

- So, let's plug in the numbers. Recall
  - $P(D) = 0.02, \; P(T \mid D^C) = 0.01, \; P\left(T^C \mid D\right) = 0.05$
  - So, $P(T \mid D) = 0.95, P\left(D^C\right) = 0.98$

$$P(D \mid T) = \frac{0.95 \times 0.02}{0.95 \times 0.02 + 0.01 \times 0.98} = \frac{0.019}{0.0288} = 0.66$$

Example: Coding Message

- Alice is sending a coded message to Bob using “dot” and “dash,” which are known to occur in the proportion of 3 : 4 for Morse codes.
- Because of interference on the transmission line, a dot can be mistakenly received as a dash with a probability 1/8 and vice-versa.
- If Bob receives a “dot,” what is the probability that Alice sent a “dot”?

Example: Coding Message

- If Bob receives a "dot," what is the probability that Alice sent a "dot"? $\Rightarrow P(dotS \mid dotR)$

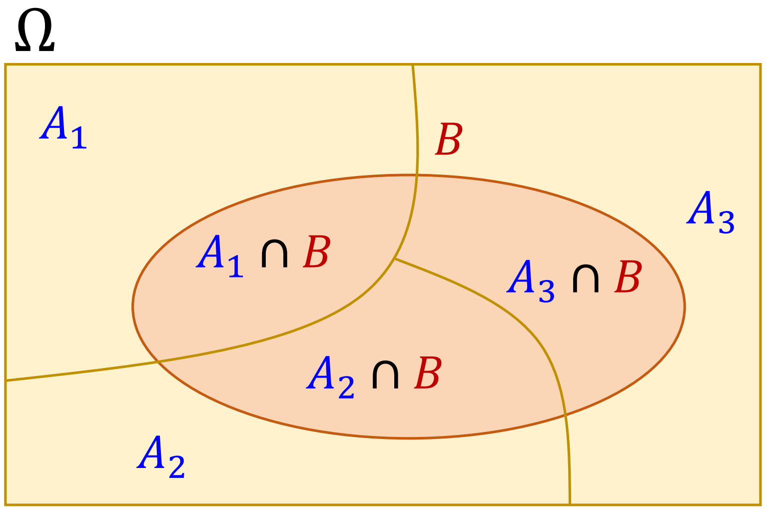
- $P(\mathrm{dot}S) = 3/7, P(\mathrm{dash}S) = 4/7$
- $P(\mathrm{dash}R \mid \mathrm{dot}S) = P(\mathrm{dot}R \mid \mathrm{dash}S) = 1/8$

$$\boldsymbol{P(dotS \mid dotR) = \frac{P(dotR \mid dotS)P(dotS)}{P(dotR)}}$$

$$\boldsymbol{= \frac{P(dotR \mid dotS)P(dotS)}{P(dotR \mid dotS)P(dotS) + P(dotR \mid dashS)P(dashS)} = \frac{\left(1-\frac{1}{8}\right)\times\frac{3}{7}}{\left(1-\frac{1}{8}\right)\times\frac{3}{7}+\frac{1}{8}\times\frac{4}{7}} = \frac{25}{56}}$$

Total Probability Theorem

- **Obtaining the probability of a subset, using conditional probabilities**
  - Let $A_1, \ldots, A_n$ be a partition of Ω, such that $P(A_i) > 0$ for all A_i.
- **Let B be an event. Note that $\boldsymbol{B} = \cup_i (\boldsymbol{A_i} \cap \boldsymbol{B})$.**
- $P(B) = P(A_1 \cap B) + P(A_1 \cap B) + \cdots + P(A_n \cap B).$



Bayes' Theorem

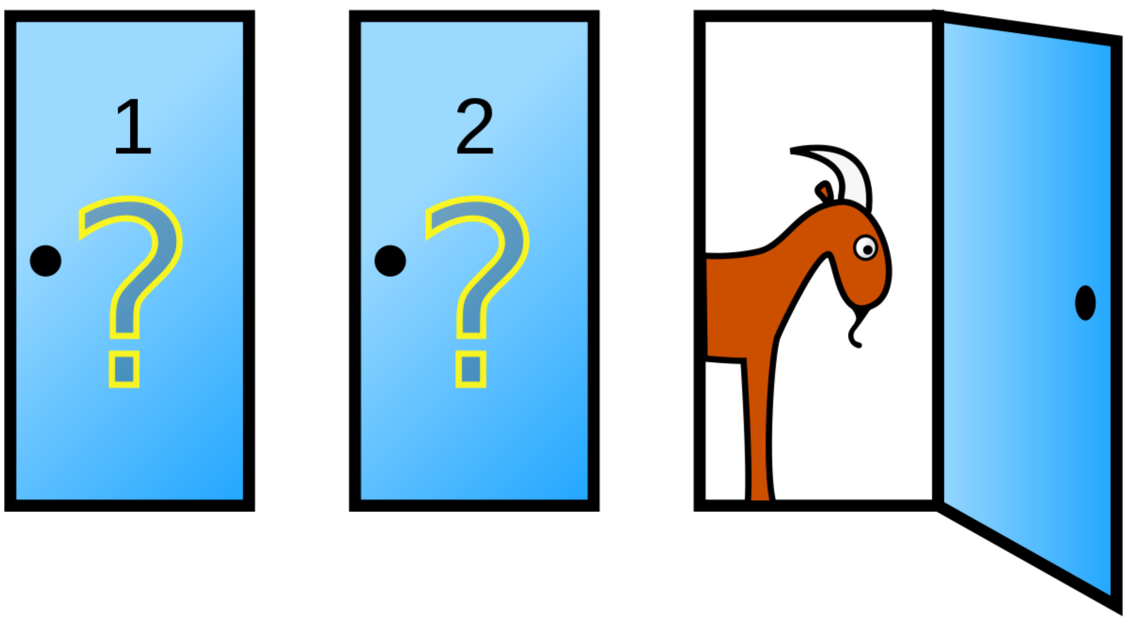
- **Let $A_1, A_2, \dots, A_n$ be a partition of the sample space.**
- **Let B be any set. Then, for each $i = 1, 2, \dots, n$**

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

Thomas Bayes (1701-1761).
English statistician, philosopher and Presbyterian minister

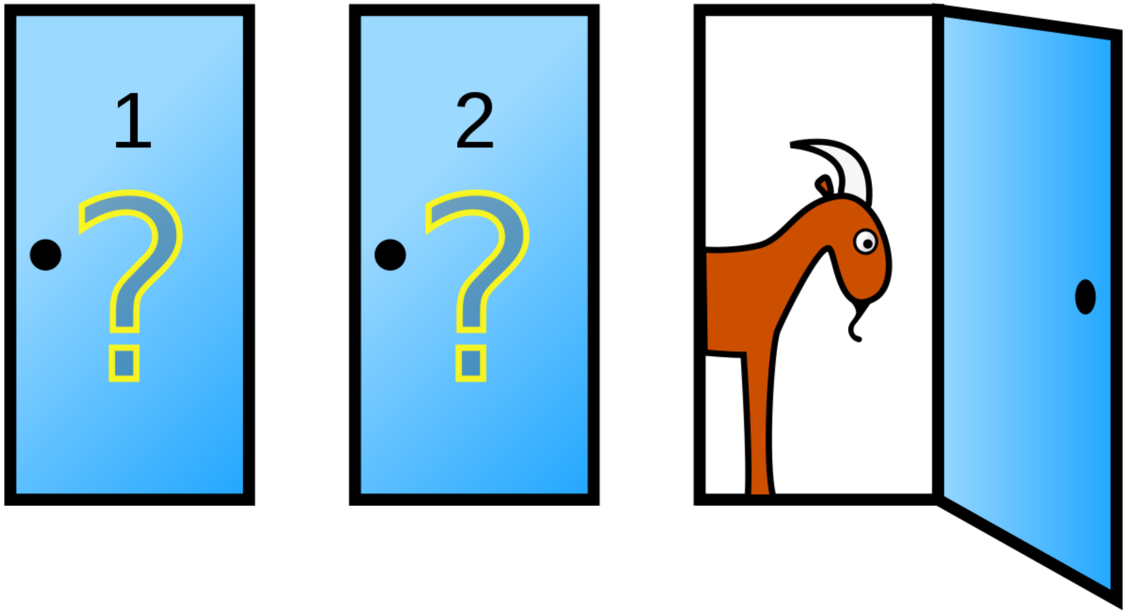
Example: The Monty Hall Problem

- You are a contestant on a game show, where you have to pick one of three doors (say A, B, and C) to open.
- One of the three doors contains a **goat**; the rest are **empty**.
- Assume the host knows which door contains the **goat**.



Example: The Monty Hall Problem

- You pick a door, say A. To build suspense, the host opens one of the other two doors (say B), revealing it is empty.
- The host asks, do you want to stick with your existing door or switch? What do you do? Does it make any difference?



Example: The Monty Hall Problem

- **Our outcome consists of two random variables: where the goat is, and which door the host opens.**
- **We will observe which door is opened; we want to infer where the goat is.**
  - G_A for the event "goat in A," G_B for "goat in B," G_C for "goat in C."
  - H_A for "host opens A," H_B for "host opens B," H_C for "host opens C."
- **What is $P(\boldsymbol{G_A})$?**
  - $\boldsymbol{P(G_A) = P(G_B) = P(G_C) = 1/3}$.

Example: The Monty Hall Problem

- **You picked door A (without loss of generality).**

- **For every possible location of the goat, we can calculate the conditional probability of the host opening a given door.**
  - We assume that the host opened a door she knew to be **empty**.
  - We know she is not going to open door that we picked.

- **Three possible cases for $P(H_B)$**
  - If the **goat** is in A, what is the probability that she opens door B?
  - If the **goat** is in B, what is the probability that she opens door B?
  - If the **goat** is in C, what is the probability that she opens door B?

Example: The Monty Hall Problem

- **If the goat is in A, what is the probability that she opens door B?**
  - $P(H_B \mid G_A) = 1/2.$

- **If the goat is in B, what is the probability that she opens door B?**
  - $P(H_B \mid G_B) = 0.$

- **If the goat is in C, what is the probability that she opens door B?**
  - $P(H_B \mid G_C) = 1.$

Example: The Monty Hall Problem

- **If the host opens door B, what's the probability that the goat is in door C?**

- **By Bayes' Rule,**

$$\boldsymbol{P(G_C \mid H_B) = \frac{P(H_B \mid G_C)P(G_C)}{P(H_B)}}$$

- **We know that $\boldsymbol{P(G_C) = 1/3}$ and $\boldsymbol{P(H_B \mid G_C) = 1}$.**
- **By the law of total probability, $\boldsymbol{P(H_B)}$ is**
  - $P(H_B) = P(H_B|G_A)P(G_A) + P(H_B|G_B)P(G_B) + P(H_B|G_C)P(G_C)$
  - $P(H_B) = \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{1}{2}$

Example: The Monty Hall Problem

- Therefore,

$$P(G_C \mid H_B) = \frac{P(H_B \mid G_C)P(G_C)}{P(H_B)}$$

$$P(G_C \mid H_B) = \frac{1/3 \times 1}{1/2} = \frac{2}{3}$$

- Given the partial information that the host has opened door B, the probability that the goat is in door C is 2/3.

- So, we should switch!

Bayes' Theorem and Inference

- **Systematic approach for incorporating new evidence**

- **Bayesian inference**
  - **Initial beliefs** $P(A_i)$ on possible causes of an **observed event** B
  - A **model** of the world under each A_i: $P(B \mid A_i)$

$$A_i \xrightarrow[P(B \mid A_i)]{\textbf{model}} B$$

  - Draw **conclusions** about **causes**.

$$B \xrightarrow[P(A_i \mid B)]{\textbf{Inference}} A_i$$

Bayes' Theorem in ML

- It is useful for **inferring hidden causes** from **our observation.**

Posterior probability   Likelihood   Prior probability

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(\theta)} \propto P(X \mid \theta)P(\theta)$$

θ: parameter, X: data

- It is also commonly used for parameter estimation methods.
  - Maximum likelihood estimation (MLE)
  - Maximum a posteriori estimation (MAP)

Bayes' Theorem in ML

Notations

- Posterior is the probability of the parameters $\boldsymbol{\theta}$ given $\boldsymbol{X}$.
- Prior encapsulates our subjective prior knowledge of the observed (latent) variable $\boldsymbol{\theta}$ before observing any data.
- Likelihood is the function of $\boldsymbol{\theta}$ given fixed $\boldsymbol{X}$.

$$\underbrace{P(\boldsymbol{\theta} \mid \boldsymbol{X})}_{\text{Posterior}} = \frac{\overbrace{P(\boldsymbol{X} \mid \boldsymbol{\theta})}^{\text{Likelihood}}\overbrace{P(\boldsymbol{\theta})}^{\text{Prior}}}{\underbrace{P(\boldsymbol{X})}_{\text{Evidence}}} \propto P(\boldsymbol{X} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})$$

It is also commonly used for parameter estimation methods.

- Maximum likelihood estimation (MLE)
- Maximum a posteriori estimation (MAP)

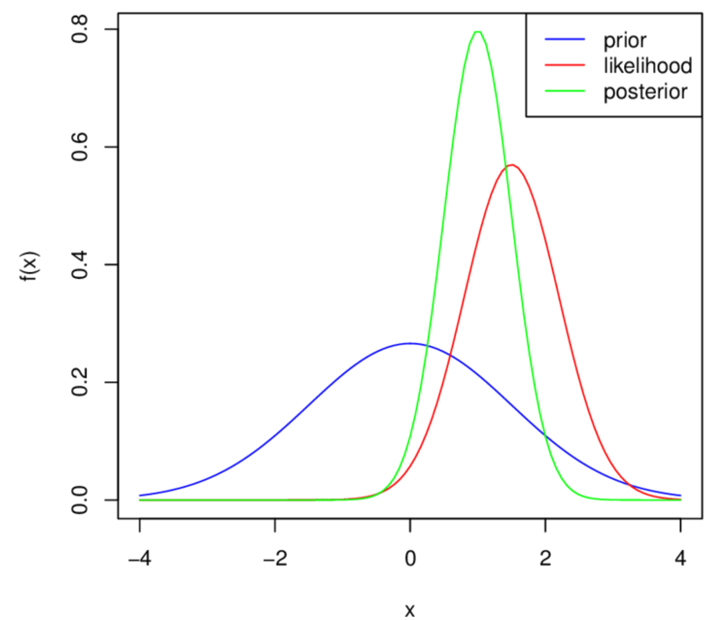
Bayes' Theorem in ML

➢ Intuition

- **Prior:** how plausible is the model a priori **before observing the data**?
  - The probability of being a head is 0.5.
- **Likelihood**: how well does **the model explain the data**?
  - For 4 out of 5 trials, the coin is head.
- **Posterior**: how plausible is the model **after observing the data**?
  - For this coin, the probability of a head is 0.7.

Posterior Likelihood Prior

$$P(\theta \mid X) \propto P(X \mid \theta) P(\theta)$$



Bayes' Theorem: Model Version

- **Let M be model, E be evidence.**
- $P(M|E)$ **proportional to** $P(E|M) \times P(M)$

$$P(M \mid E) \propto P(E \mid M)P(M)$$

Posterior Likelihood Prior

- **Intuition**
  - **Prior** = how plausible is the event (model, theory) a priori before seeing any evidence?
  - **Likelihood** = how well does the model explain the data?

Statistical Independence

Independence of Two Events

- Two events A and B are independent if the probability of A does not affect the probability of B.

$$P(A,B) = P(A)P(B) \quad \Leftrightarrow \quad P(B \mid A) = P(B)$$

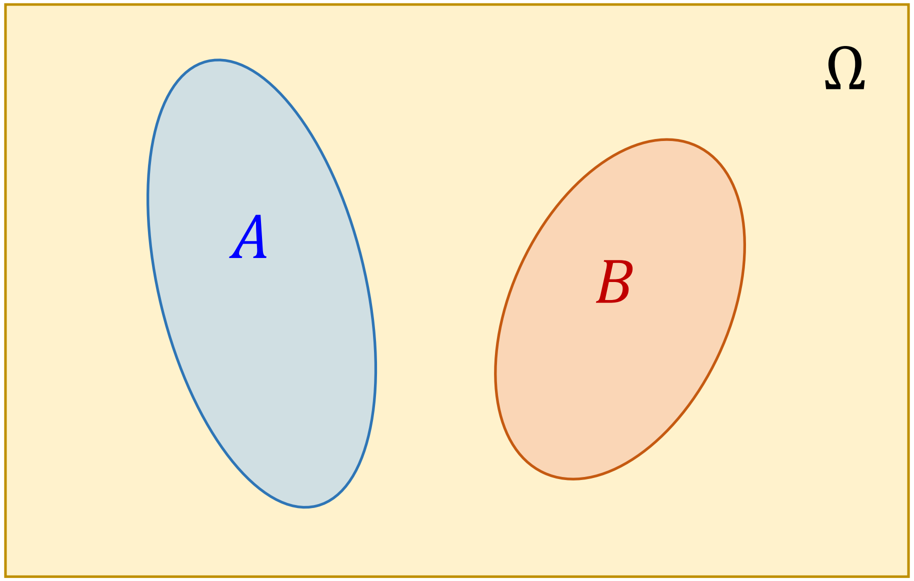
- **Intuitive definition:** $P(B \mid A) = P(B)$
  - The occurrence of A provides **no new information** about B.

- **Symmetric with respect to A and B**
  - Implies that $P(A \mid B) = P(A)$
  - It applies even if $P(A) = 0$

Example: Independence of Two Events

- Are they **independent**?

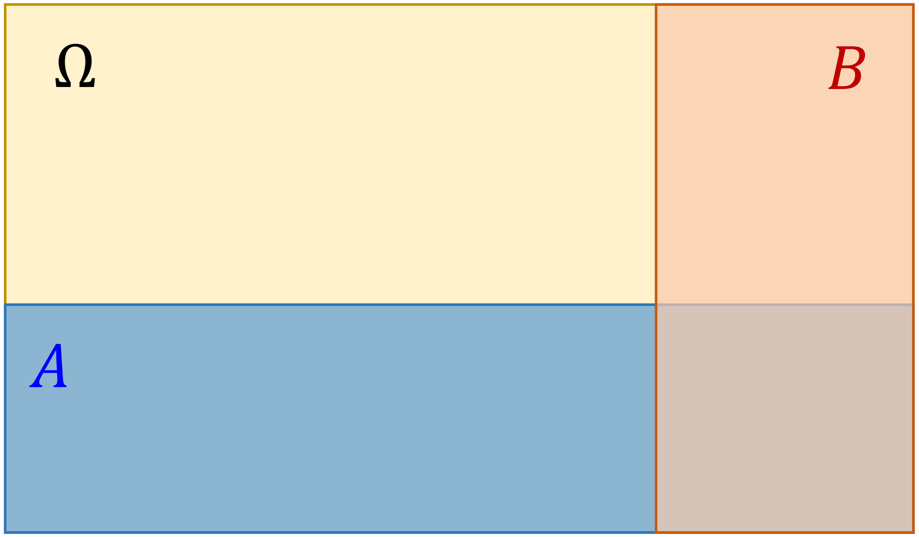
$$P(A, B) = P(A)P(B)$$



Example: Independence of Two Events

- Are they **independent**?

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



Example: Gambler

- A gambler is rolling 4 fair dice. What is the probability that there is at least one 6 in 4 rolls?

Example: Gambler

- A gambler is rolling 4 fair dice. What is the probability that there is at least one 6 in 4 rolls?

- Each roll is independent.
- Let X_i denote the event that there is no six in the i-th roll.

- P(at least 1 six in 4 rolls) $= 1 -$ P(no sixes in 4 rolls)

Example: Gambler

- A gambler is rolling 4 fair dice. What is the probability that there is at least one 6 in 4 rolls?

- Each roll is independent.
- Let X_i denote the event that there is no six in the i-th roll.

- P(at least 1 six in 4 rolls) $= 1 - P$(no sixes in 4 rolls)

$$= 1 - P(\boldsymbol{X_1} \cap \boldsymbol{X_2} \cap \boldsymbol{X_3} \cap \boldsymbol{X_4})$$

$$= 1 - P(\boldsymbol{X_1})P(\boldsymbol{X_2})P(\boldsymbol{X_3})P(\boldsymbol{X_4})$$

$$= 1 - \left(\frac{\mathbf{5}}{\mathbf{6}}\right)^{\mathbf{4}} = \mathbf{0.518}$$

Independence of Two Events

- If A and B are independent, then A and B^C are independent.

- Is it true or false?

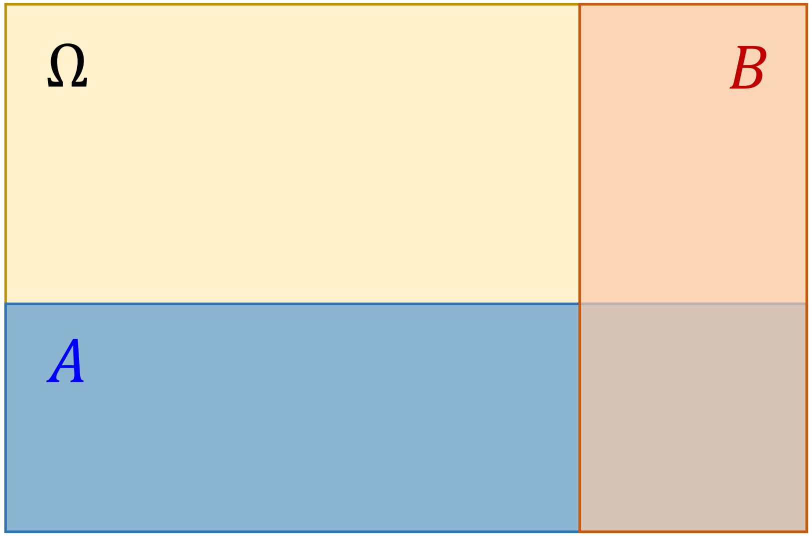
- $P(A) = P(A \cap B) + P\left(A \cap B^C\right)$
- $P(A) = P(A)P(B) + P\left(A \cap B^C\right)$
- $P\left(A \cap B^C\right) = P(A) - P(A)P(B) = P(A)\left(1 - P(B)\right)$
- $P\left(A \cap B^C\right) = P(A)P\left(B^C\right)$

Some Ground Rules

- **Theorem. If A and B are independent ($A \perp B$), then so are $\boldsymbol{A}$ and $\boldsymbol{B^C}$ are independent.**
  - $P(A \cap B^C) = P(A) - P(A \cap B) = P(A) - P(A)P(B)$
    $= P(A)\big(1 - P(B)\big) = P(A)P(B^C)$

- **$\boldsymbol{A^C}$ and $\boldsymbol{B}$ are independent.**

- **$\boldsymbol{A^C}$ and $\boldsymbol{B^C}$ are independent.**



Conditional Independence

- **Bob and Alice mostly go to their 9 am probability class when the weather is sunny. Are the events {Bob goes to class} and {Alice goes to class} independent events?**

- **No. If I know Bob went to class. Then it is likely that it is sunny. This makes it likely that Alice goes too.**

Conditional Independence

- **Bob and Alice mostly go to their 9am probability class when the weather is sunny. Are the events {Bob goes to class} and {Alice goes to class} independent events?**

- **Given the event {its sunny}, {Bob went to class} does not give us any information about {Alice went to class}.**
- **{Bob goes to class} and {Alice goes to class} are conditionally independent given {its sunny}.**

- **Two events A and B are conditionally independent given another event C if $P(A \cap B|C) = P(A|C)P(B|C)$**
  - We write this as $\boldsymbol{A} \perp \boldsymbol{B}|\boldsymbol{C}$.

Conditional Independence

- **Recall, we said two events A and B were independent if**

$$\boldsymbol{P(A \cap B) = P(A)P(B)}$$

- **If $P(B) > 0$, this means that $P(A|B) = P(A)$**
  - Knowing $\boldsymbol{B}$ tells us **nothing** about the probability of $\boldsymbol{A}$

- **We can extend this definition to conditional probabilities.**
- **We say two events A and B are conditionally independent given some event C if $\boldsymbol{P(A \cap B|C) = P(A|C)P(B|C)}$.**
  - We write this as $A \perp B|C$.

Conditional Independence

- **Conditional independence:** $P(A \cap B|C) = P(A|C)P(B|C)$
- **Intuitively, what we are thinking is,** $P(A|B \cap C) = P(A|C)$**.**
- **Is this true? Assume that** $P(B \cap C) > 0$**.**

$$P(A \mid B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)}$$

$$= \frac{P(A \cap B \mid C)P(C)}{P(B \mid C)P(C)}$$

$$= \frac{P(A \mid C)P(B \mid C)P(C)}{P(B \mid C)P(C)} = P(A \mid C)$$

Conditional Independence

- **So, provided $P(B \cap C) > 0$, we can write**

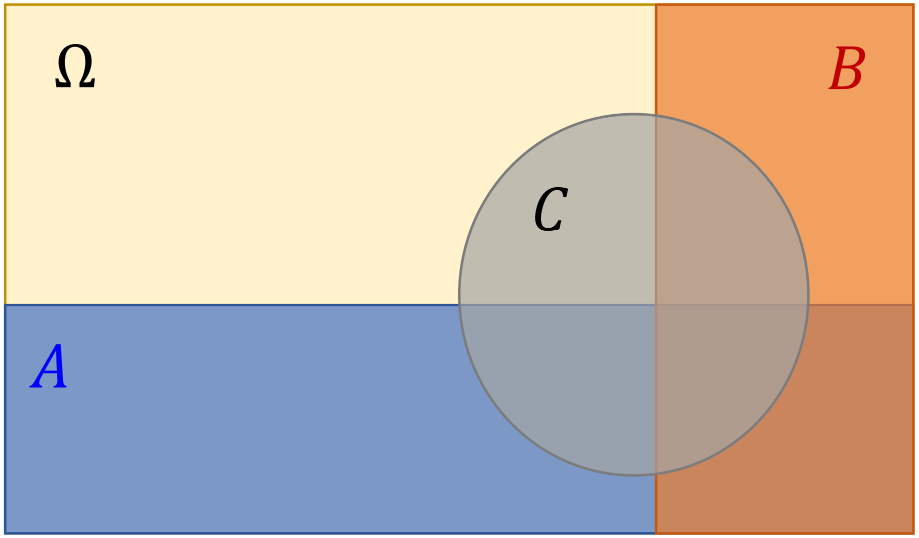
$$\boldsymbol{P(A \mid B \cap C) = P(A \mid C)}$$

- **Given we know C, also knowing B tells us nothing about A.**

Example: Conditional Independence

- Assume A and B are independent.

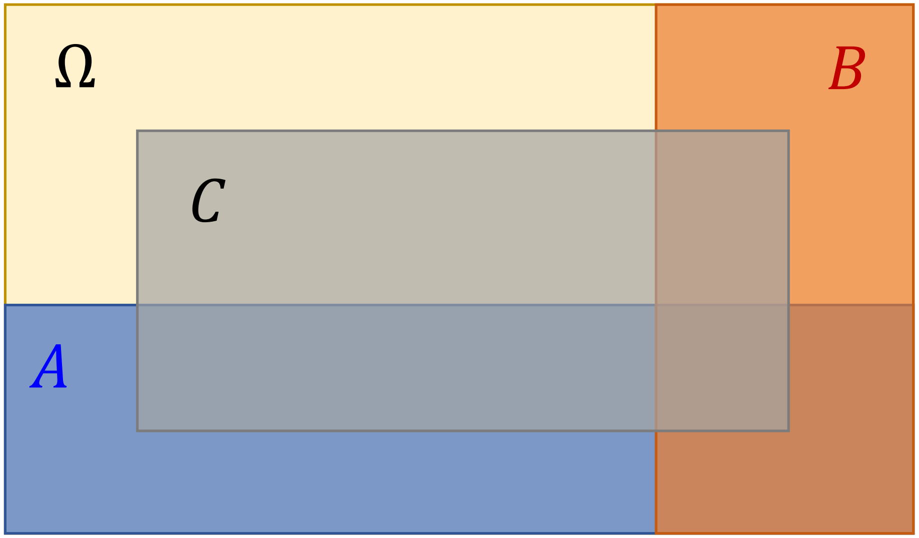
- If C occurred, are A and B independent?



Example: Conditional Independence

- Assume A and B are independent.

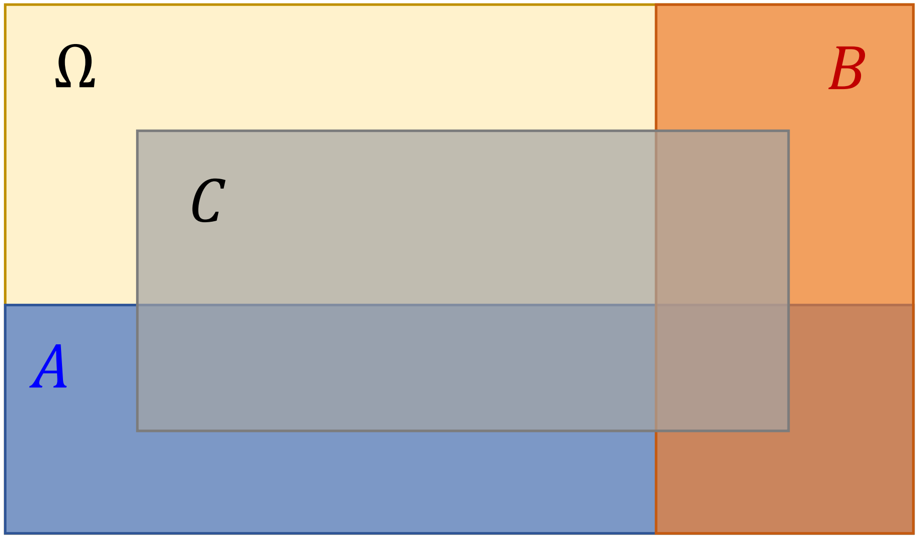
- If C occurred, are A and B independent?



Conditional Independence

- Conditional independence is defined as independence under the probability law $P(\cdot | C)$.

$$P(A\,,B \mid C) = P(A \mid C)P(B \mid C) \iff P(A \mid B, C) = P(A \mid C)$$



Discrete and Continuous Probability Distribution

Motivation: Random Variable

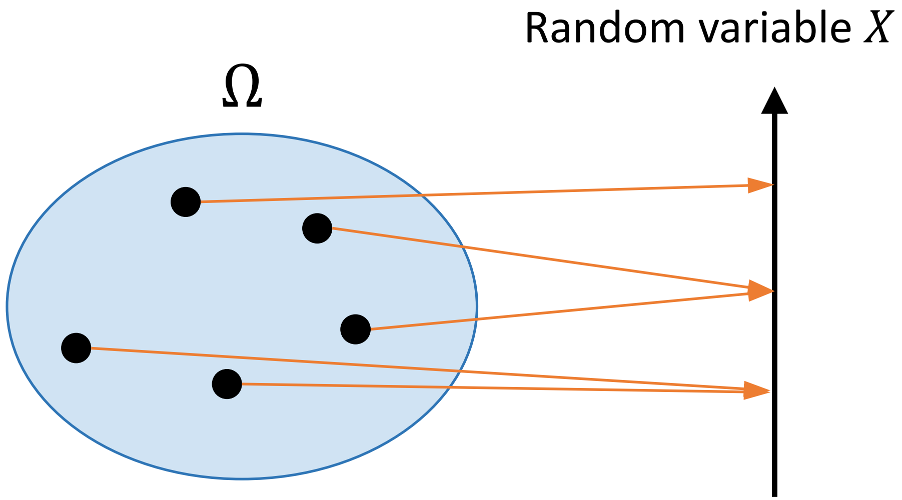
- **Example**
  - We are taking an opinion poll among 100 students about how understandable the lectures are.
  - If "1" is used for understandable and "0" is used for not, then there are 2^{100} **possible outcomes**!!
  - The thing that matters most is the number of students who think the class is understandable (or equivalently not). If we define a variable X to be that number, then the range of X is $\{0,1,\ldots,100\}$.
  - Much **easier** to handle that!

- **For many experiments, it is easier to use a new variable that summarizes all possible outcomes.**

Random Variable as a Mapping

- **A random variable X is a function that takes an outcome O and returns a particular quantity of interest x.**

$$X(O) = x, \textit{or just } X = x$$



- **Basically, a way to redefine a probability space to a new probability space**
  - X must obey axioms of probability.

Example of Random Variables

- You toss a coin: is it head or tail?
  - $f: \{H, T\} \rightarrow \{0, 1\}$

- You roll a die: what number do you get?
  - $f: \{1, 2, \dots, 6\} \rightarrow \{1, 2, \dots, 6\}$

- Number of heads in three coin tosses
  - $f: \{HHH, HHT, \dots, TTT\} \rightarrow \{0, 1, 2, 3\}$

- The sum of two rolls of dice
  - $f: \{(1, 1), (1, 2), \dots, (6, 6)\} \rightarrow \{2, 3, \dots, 12\}$

Example of Random Variables

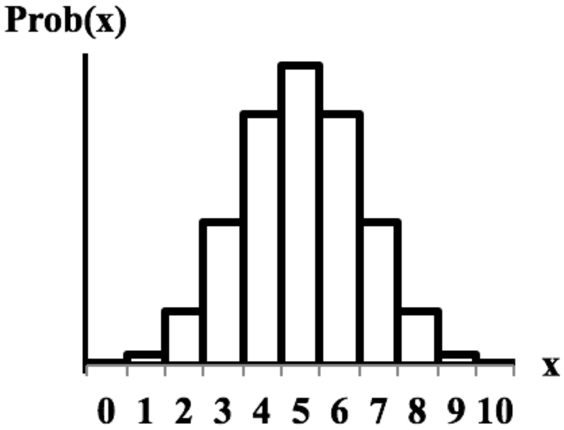
- X = **“The number of heads” is the random variable.**
  - There can be 0 heads, 1 head, 2 heads, 3 heads.
- **Target space =** $\{0, 1, 2, 3\}$

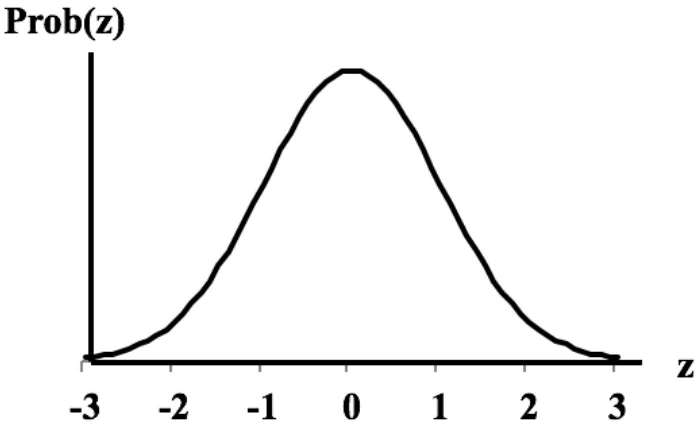
Types of Probability Space

- **Define $|\Omega|$ = the number of possible outcomes O**

- **Discrete space: $|\Omega|$ is finite.**
  - Analysis includes **summations** ($\sum$).

- **Continuous space: $|\Omega|$ is infinite.**
  - Analysis includes **integrals** ($\int$).





Examples of Discrete Probability Space

- **Consider two consecutive flips of a coin**
  - 4 possible outcomes: $\Omega = \{HH, HT, TH, TT\}$
  - $2^4 = 16$ possible events
    - $E = \{TH, HT\}$, i.e., one of the coins is head.
- **If the coin is fair, then the probabilities of outcomes are equal.**
  - $P(HH) = P(HT) = P(TH) = P(TT) = 1/4$
- **Probability of the event with one head is**
  - $P(HT) + P(TH) = 1/2$

Examples of Discrete Probability Space

- **Consider single roll of a six-sided die.**
  - 6 possible outcomes: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - $2^6 = 64$ possible events

- **If the die is fair, then the probabilities of outcomes are equal.**
  - $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$

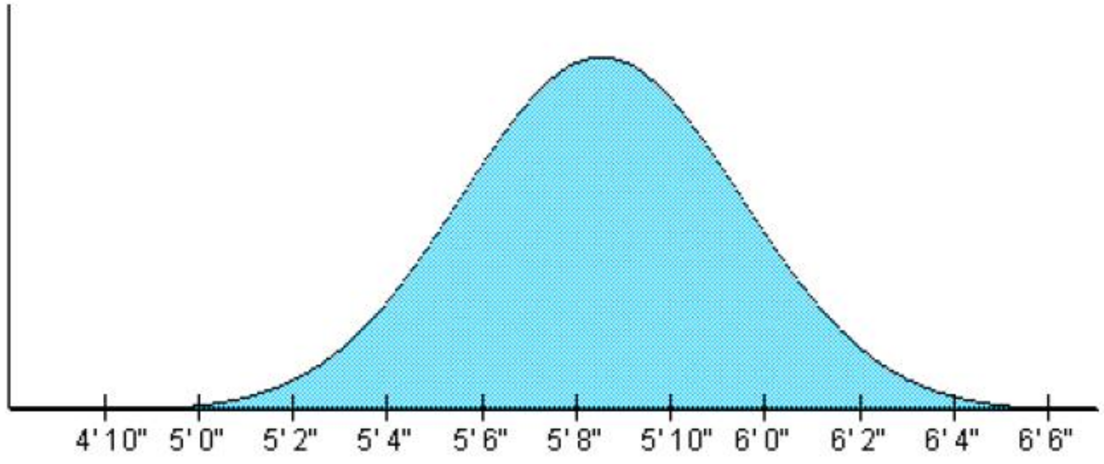
- **Probability of the event with odd numbers is**
  - $P(1) + P(3) + P(5) = 1/2$

Example of Continuous Probability Space

➢ Height of a randomly chosen male

- Infinite number of possible outcomes: O has some single value in range 2 feet to 8 feet
- Infinite number of possible events
  - $E = (O \mid O < 5.5\, feet)$, i.e., individual chosen is less than 5.5 feet

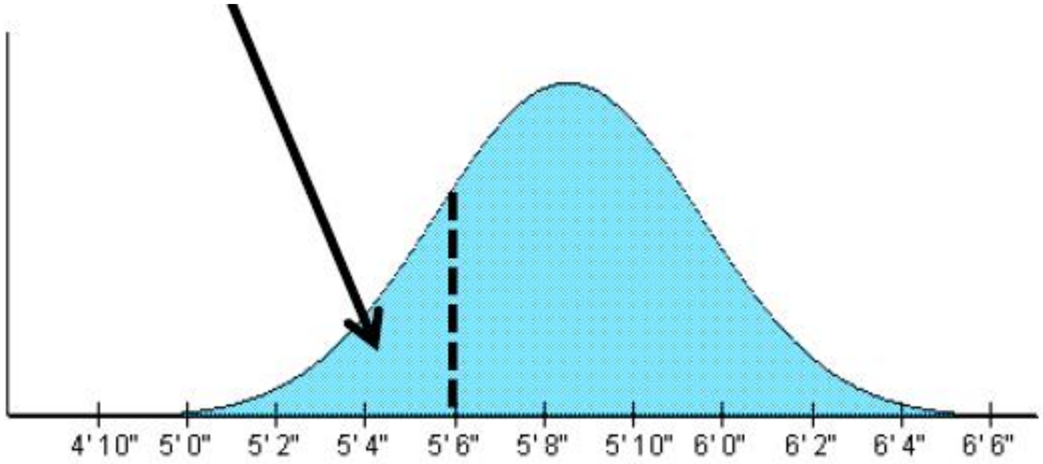
- Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$



Example of Continuous Probability Space

- **Height of a randomly chosen male**
  - Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$

- **Examples**
  - $P(O = 5.8) > P(P = 6.2)$
  - $P(O < 5.6) = \int P(O)$ from $O = -\infty$ to $5.6 \approx 0.25$

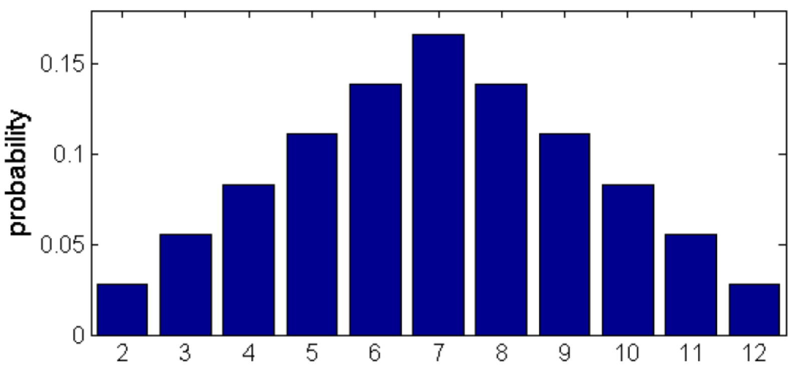


Probability Distributions

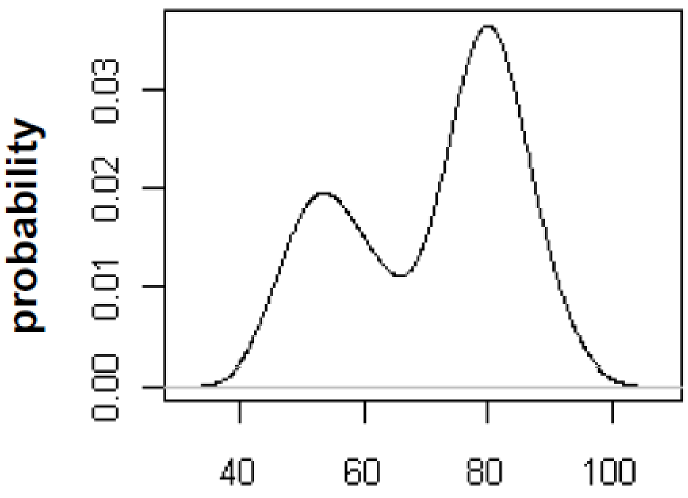
- Discrete probability distribution

- Continuous probability distribution

Probability mass function (PMF)



Probability density function (PDF)



Multivariate Probability Distributions

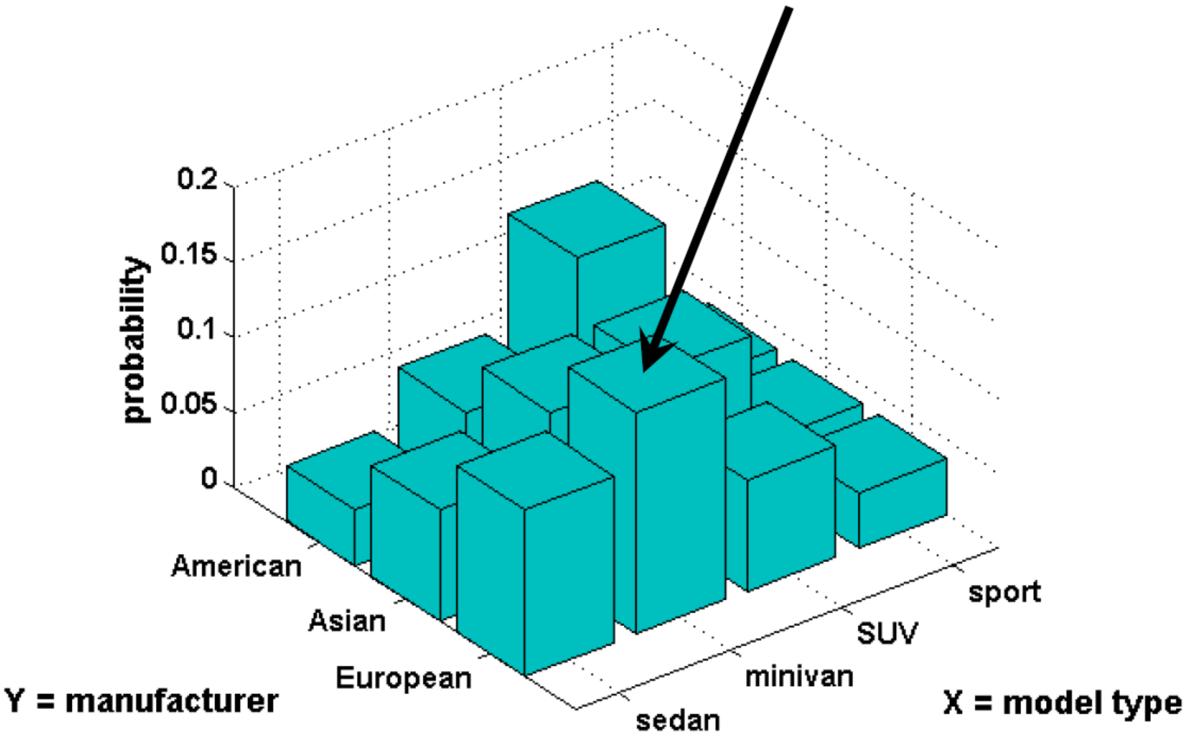
- **Several random processes occur (doesn't matter whether in parallel or in sequence)**
  - Want to know probabilities for each possible combination of outcomes

- **Joint probability**
  - Two processes whose outcomes are represented by random variables X and Y.
  - Probability that process X has outcome x and process Y has outcome y is denoted as:

$$\boldsymbol{P(X = x, Y = y)}$$

Example of Joint Probability

$$P(X = minvan, Y = European) = 0.14$$



Multivariate Probability Distributions

Marginal probability

- Probability distribution of a single variable in a joint distribution

$$P(X = x) = \sum_{b=all\ values\ of\ Y} P(X = x, Y = b)$$

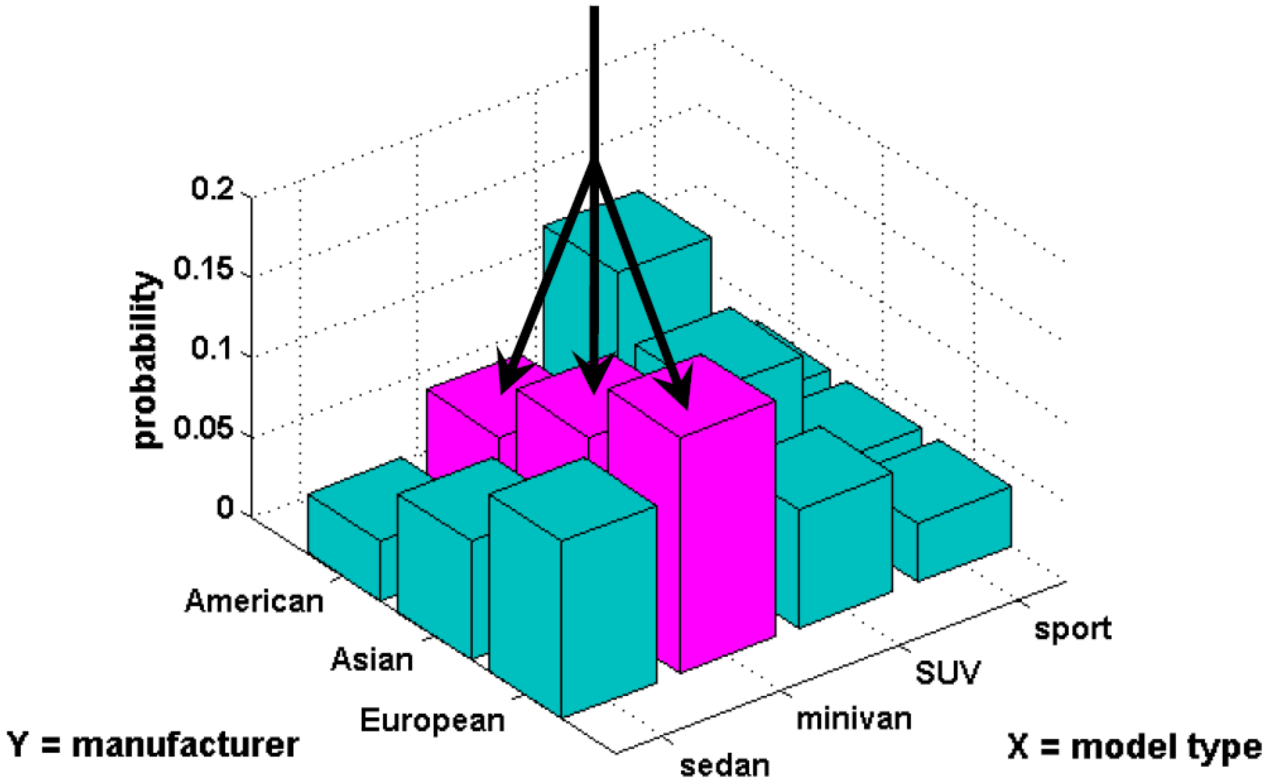
Conditional probability

- Probability distribution of one variable given that another variable takes a certain value

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

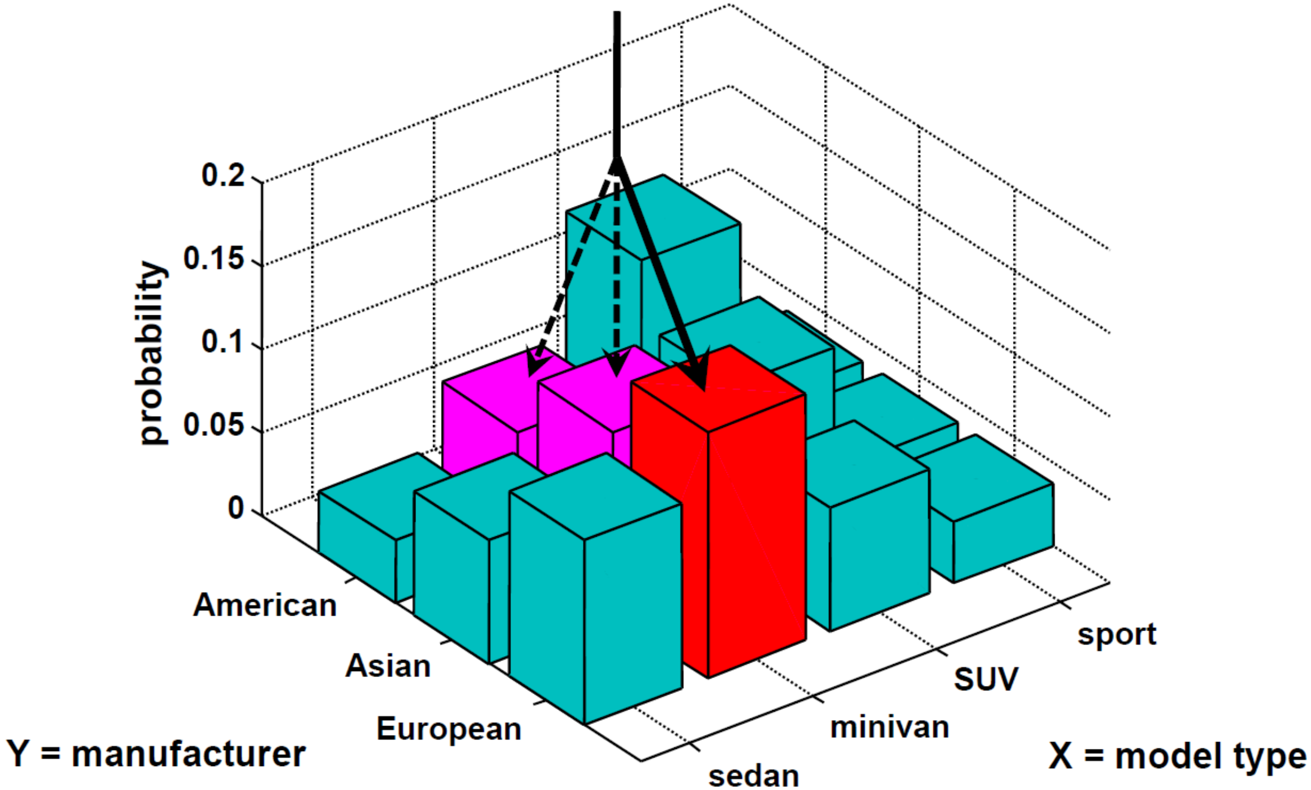
Example of Marginal Probability

$$P(X = minivan) = 0.07 + 0.11 + 0.14 = 0.32$$



Example of Conditional Probability

$$P(Y = European \mid X = minivan)$$
$$= 0.14/(0.07 + 0.11 + 0.14) = 0.4375$$



Q&A