Introduction to Probability

Why Randomness

- **Randomization.** We allow a fair coin flip in unit time.
- Why randomize?
  - Deterministic algorithms offer little flexibility
  - Randomness often enables to surprisingly simple & fast algorithms
- Very important in computer science:
  - Symmetry-breaking protocols, memory management, learning algorithms, contention resolution, hashing, load balancing, cryptography, AI, game theory
- Gives insight in “real world” issues
  - Polling, risk assessment, scientific testing, gambling, etc.

Probability Review

- Before we design/analyze randomized algorithms, we need a foundation in probability
- Plan: we'll start with some things you've likely seen before
    - Will be a "review" of probability from Discrete Math
    - Since each Math 200 differs, ensure everyone has same background
- Will move on to randomized algorithms and data structures:
    - Randomized Quicksort
    - Hashing
    - Skip lists
    - Fingerprinting
    - etc.

“Deathbed” Formulas

- You should remember these even on your deathbed [MAB]
- *Extremely* useful in probability
- $\left(1+\frac{1}{n}\right)^n \approx e \qquad \left(1-\frac{1}{n}\right)^n \approx \frac{1}{e}$ for large enough n (gets close quite quickly)
- More precisely: $\left(1+\frac{1}{n}\right)^n \leq e \qquad \left(1-\frac{1}{n}\right)^n \leq \frac{1}{e}$
- $\left(\frac{x}{y}\right)^y \leq \binom{x}{y} \leq \left(\frac{ex}{y}\right)^y$

$$\binom{x}{y} = \frac{x!}{y!(x-y)!}$$ is the number of

y-sized subsets of x items

Discrete Probability Review

Sample Space

- A discrete probability space consists of a non-empty, countable set Ω, called the *sample space*, and a probability mass function $\Pr : \Omega \to \mathbb{R}$ s.t.
  - $\Pr[\omega] \geq 0 \quad \forall \omega \in \Omega$, and $\sum_{\omega \in \Omega} \Pr[\omega] = 1$
- **Idea**: the sample space consists of *all possible outcomes*
- When flipping a coin, the sample space is $\Omega = \{\text{heads, tails}\}$
- When rolling a six-sided die, $\Omega = \{1,2,3,4,5,6\}$
- If you're stuck on a probability question, sometimes it may help to list all possible outcomes!

Discrete Probability Review

- An **event** is a set of outcomes
  - E.g. Seeing a heads when we toss a fair coin
  - E.g. Seeing a six when we roll a fair die
- Probability of an event is the weight of all outcomes satisfying that event
  - A fair coin: $\Pr[\text{heads}] = \Pr[\text{tails}] = 1/2$
  - A fair six-sided die: $\Pr[\omega] = 1/6 \quad \forall \omega \in \Omega$

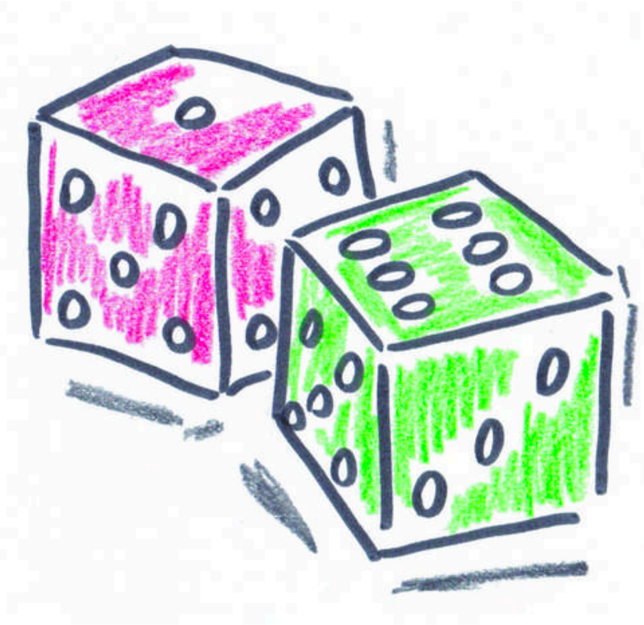
Four Step Method

- Step 1. Find the sample space
- Step 2. Define events of interest
- Step 3. Determine outcome probabilities
- Step 4. Determine event probabilities

When it comes to probability:

Intuition: Bad

Formalism: Good



Example: Baby Sex Likelihood

- Let's say every baby born is a girl or a boy with probability $1/2$ each
- If someone has four children, is it more likely that they have two girls and two boys? Or three of one, and one of the other?
- **First**: what is the sample space/how many outcomes do we have

Example: Baby Sex Likelihood

- Let's say every baby born is a girl or a boy with probability $1/2$ each
- If someone has four children, is it more likely that they have two girls and two boys? Or three of one, and one of the other?
- **First**: what is the sample space/how many outcomes do we have

BBBG	GGGG	GGGG	GBBG
BBGB	GGGB	BBBB	BGGB
BGBB	GGBG	GGBB	BGBG
GBBB	GBGG	GBGB	BBGG

Example

- If someone has four children, is it more likely that they have two girls and two boys? Or three of one, and one of the other?
- Pr[three same sex out of four] $= 8/16 = 1/2$

BBBG	BGGG	GGGG	GBBG
BBGB	GGGB	BBBB	BGGB
BGBB	GGBG	GGBB	BGBG
GBBB	GBGG	GBGB	BBGG

Example

- If someone has four children, is it more likely that they have two girls and two boys? Or three of one, and one of the other?
- Pr[two boys and two girls] $= 6/16 = 3/8 < 1/2$

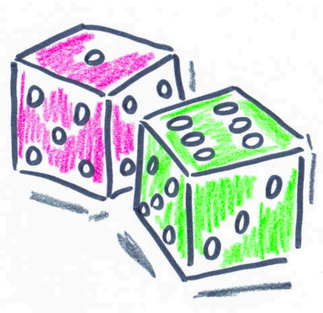
BBBG	BGGG	GGGG	GBBG
BBGB	GGGB	BBBB	BGGB
BGBB	GGBG	GGBB	BGBG
GBBB	GBGG	GBGB	BBGG

Same Example: Let's Do the Math

- Let’s say every baby is a girl or a boy with probability $1/2$ each
- If someone has four children, is it more likely that they have two girls and two boys? Or three of one, and one of the other?
- 2^4 outcomes, each outcome occurs with equal probability: $1/2^4 = 1/16$
- $\binom{4}{1} = 4$. 4 ways to have one girl; 4 ways to have one boy; total = $8/16$
- $\binom{4}{2} = 6$ ways to have two girls and two boys; total = $6/16$

Independence

- **Intuition**: two events are independent if they do not affect each other
- Example: let's say I flip two coins: the event that the first is a head, and the event that the second is a head, are independent.
- Not-independent example: Say I flip a coin 10 times, then let:
    - Event 1: Flips 1, 2, and 3 are all heads
    - Event 2: Flips 2, 3, and 4 are all heads
- These are not independent. If Event 1 is true, Event 2 is more likely. If Event 1 is false, Event 2 is less likely.



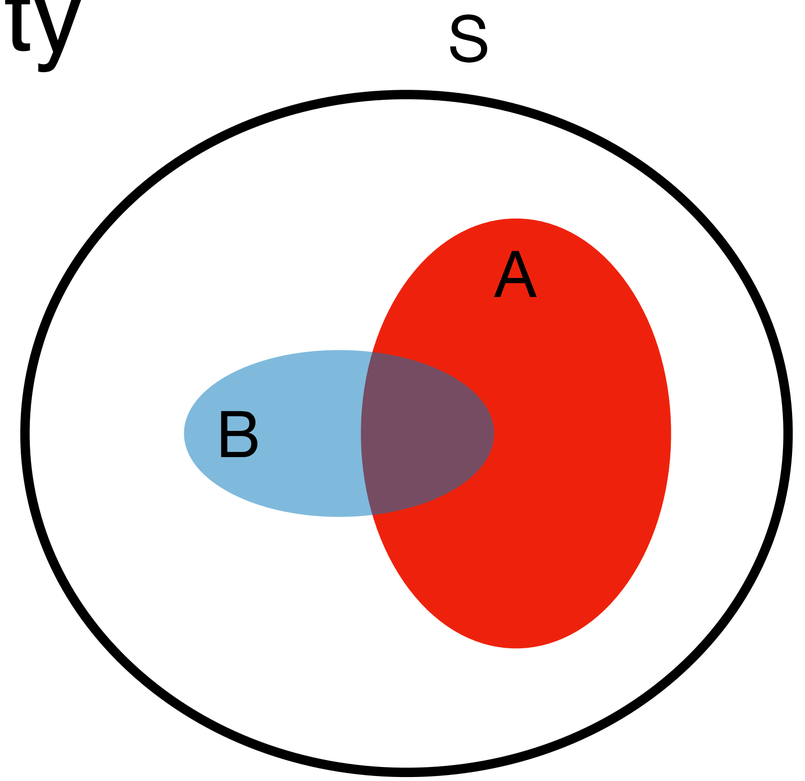
Independent Probabilities

- Definition of independence:
  - A and B are independent events if and only if:

    $$\Pr[A \text{ and } B] = \Pr[A] \cdot \Pr[B]$$

- Here A and $B = A \cap B$ (events are just subsets of the outcome space)
- Probability of flipping 10 heads in a row is $1/2^{10}$
- Probability of flipping a heads, and then rolling a 1 on a die, is $1/12$

Conditional Probability



- *What is the probability that it will rain this afternoon, given that it is cloudy this morning?*
- Conditional probability is the probability that one event happens, given that some other event definitely happens or has already happened
- **Notation**. $\Pr(A \mid B)$ denotes the probability that event A happens given that event B happens
- $\Pr[A]$ is the fraction of S that is red
- $\Pr[A \mid B]$ captures weight of A that is purple (overlaps with B) normalized over B

Conditional Probability (Def):

$$\Pr[A \mid B] = \frac{\Pr[A \text{ and } B]}{\Pr[B]} = \frac{\Pr[A \cap B]}{\Pr[B]}$$

Conditional Probability

- **Definition of conditional probability:**

$$\Pr[A \mid B] = \frac{\Pr[A \text{ and } B]}{\Pr[B]}$$

- **(Product rule).** This means that

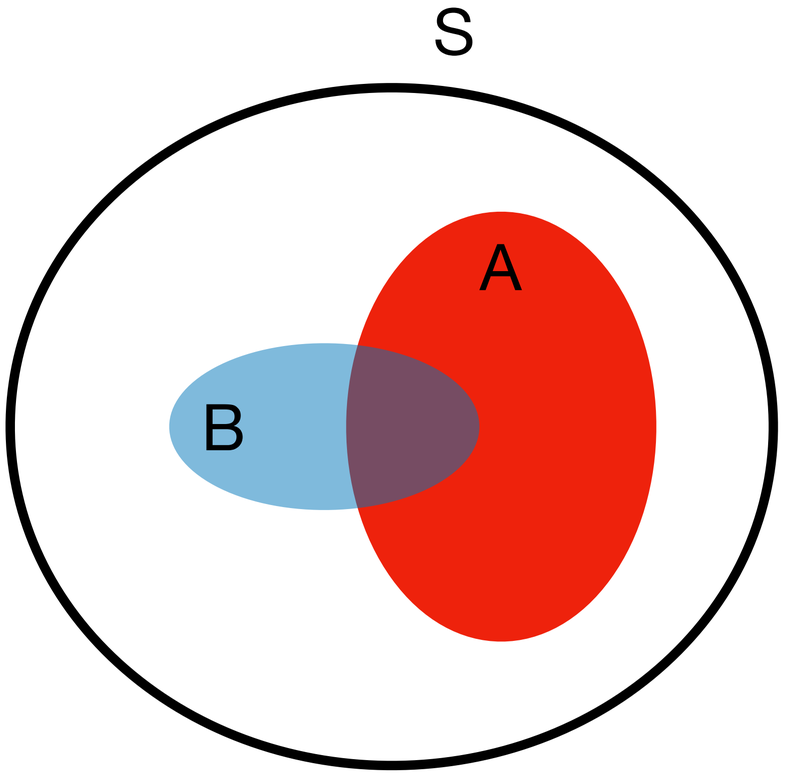
$$\Pr[A \text{ and } B] = \Pr[A \mid B] \cdot \Pr[B]$$

- We know for independent events A and B that

$$\Pr[A \text{ and } B] = \Pr[A] \cdot \Pr[B]$$

  - Means that A and B are independent if and only if

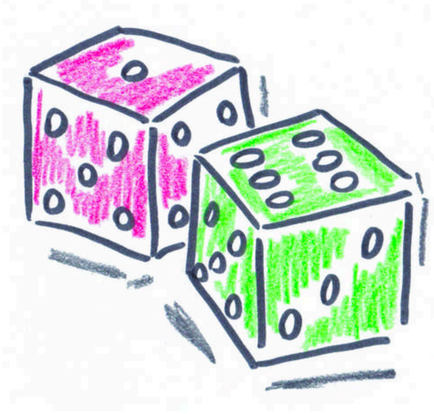
$$\Pr[A \mid B] = \Pr[A]$$

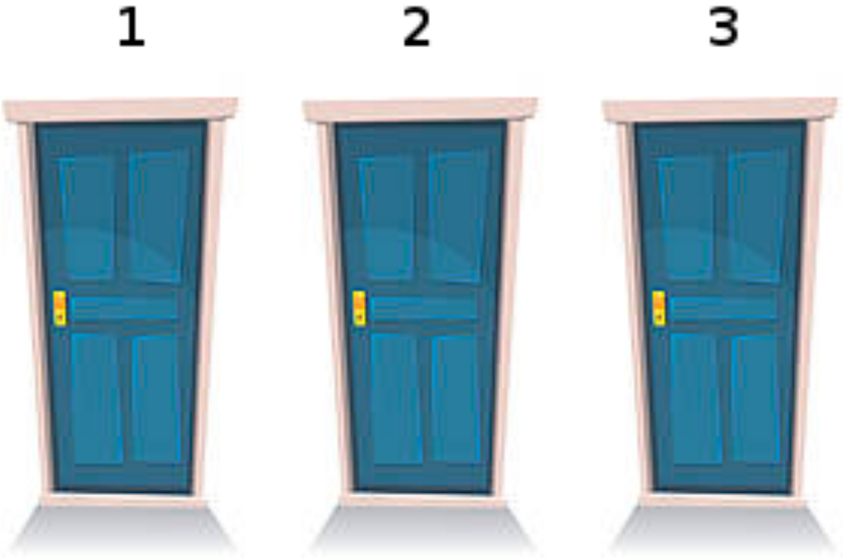


Monty Hall Problem

- "*Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say number 1, and the host, who knows what's behind the doors, opens another door, say number 3, which has a goat. He says to you, "Do you want to pick door number 2?" **Is it to your advantage to switch your choice of doors?***"

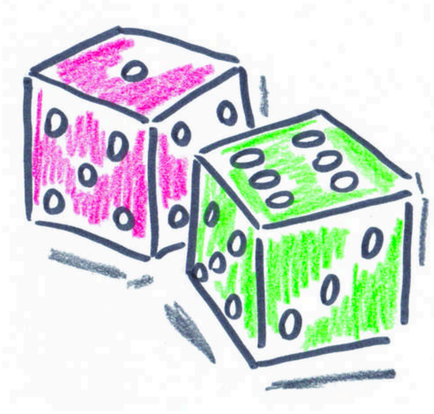
--- Craig. F. Whitaker Columbia, MD

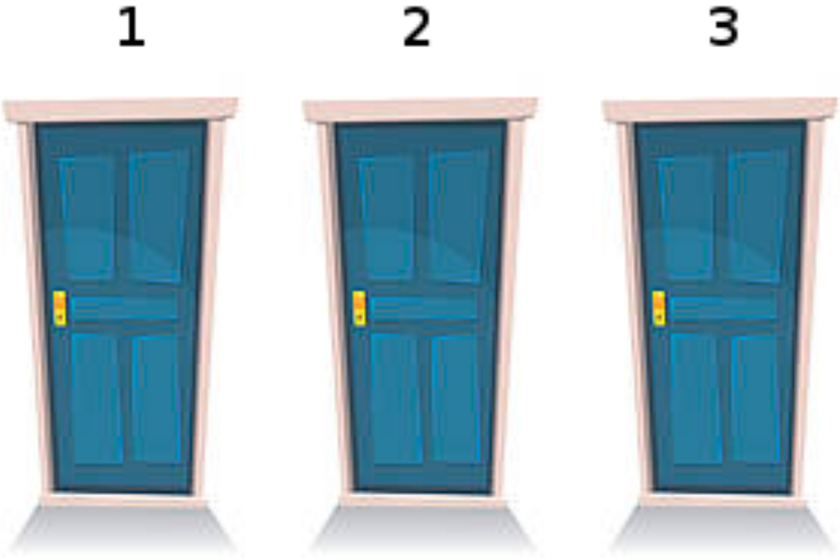




Clarifying the Problem

- The car is equally likely to be hidden behind any of the 3 doors
- The player is equally to pick any of the 3 doors, regardless of the car's location
- After the player picks a door, the host *must* open a *different* door with a goat behind it and offer the choice to switch
- If the host has a choice of which door to open, he is equally likely to select each of them

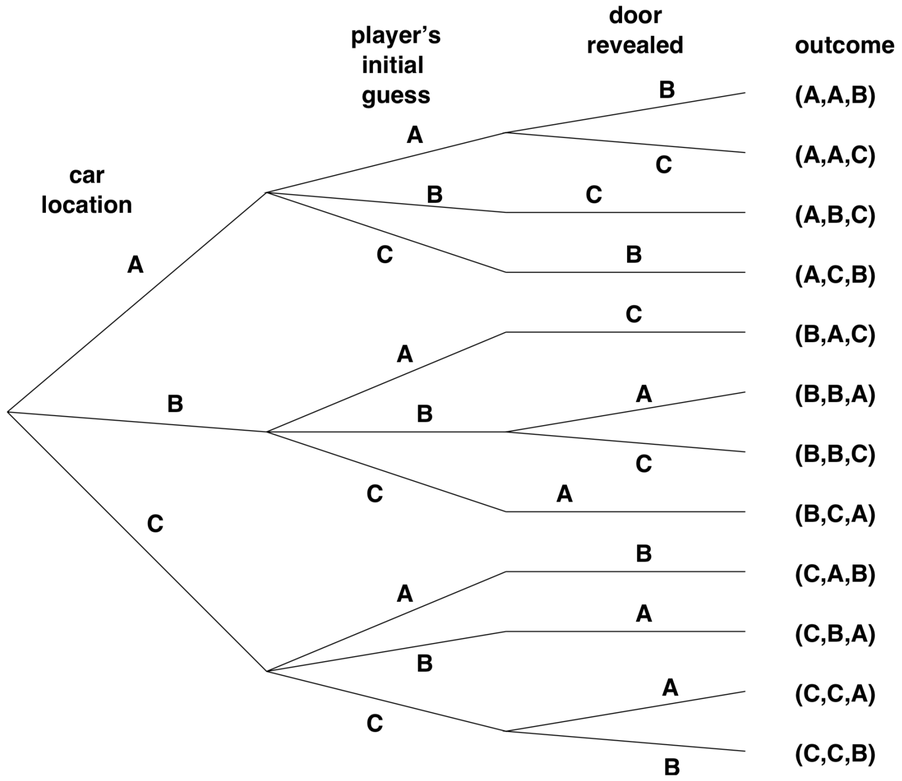




Find the Sample Space

- Sample space: set of all possible outcomes
- Here, an outcome involves 3 things:
  - door concealing the car
  - door initially chosen by the player
  - door that host opens to reveal a goat
- Every possible combination of this is an *outcome*
- We can visualize these as a *tree diagram*
- Sample space S is then:

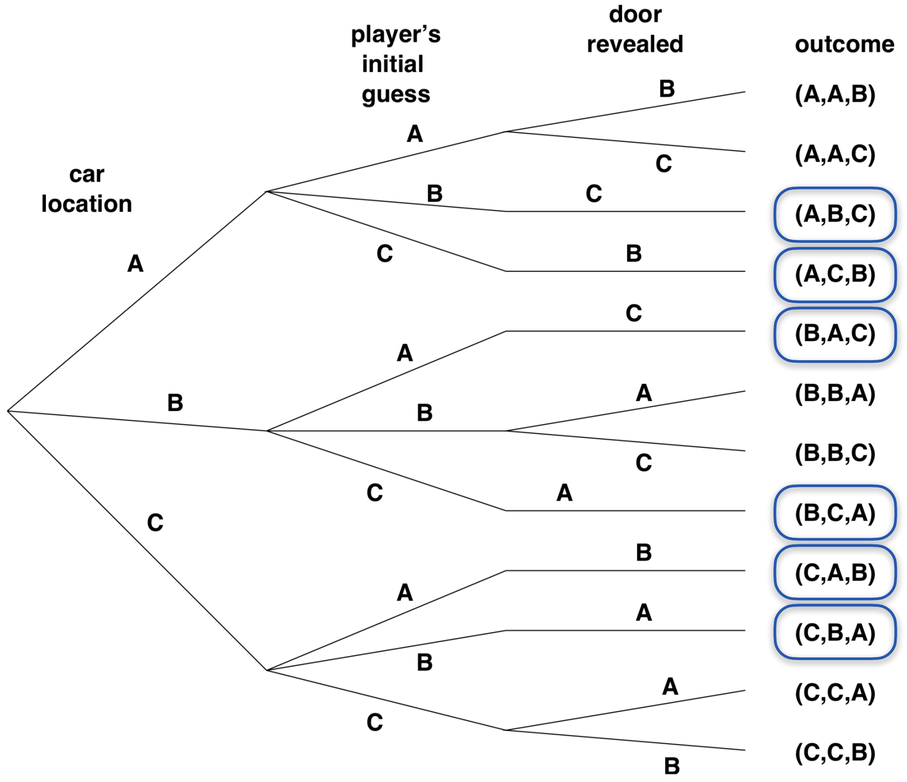
$$S = \left\{ \begin{array}{cccccc} (A,A,B), & (A,A,C), & (A,B,C), & (A,C,B), & (B,A,C), & (B,B,A), \\ (B,B,C), & (B,C,A), & (C,A,B), & (C,B,A), & (C,C,A), & (C,C,B) \end{array} \right\}$$



Define Events of Interest

- **Question**. *What is the probability that _____ ?*
- Model as an ***event*** (subset of the sample space)
- Event that player wins by switching:
  - $\{(A,B,C),(A,C,B),(B,A,C),(B,C,A),(C,A,B),(C,B,A)\}$
  - Exactly half of the outcomes
- Does switching lead to wins with probability half?
  - No!

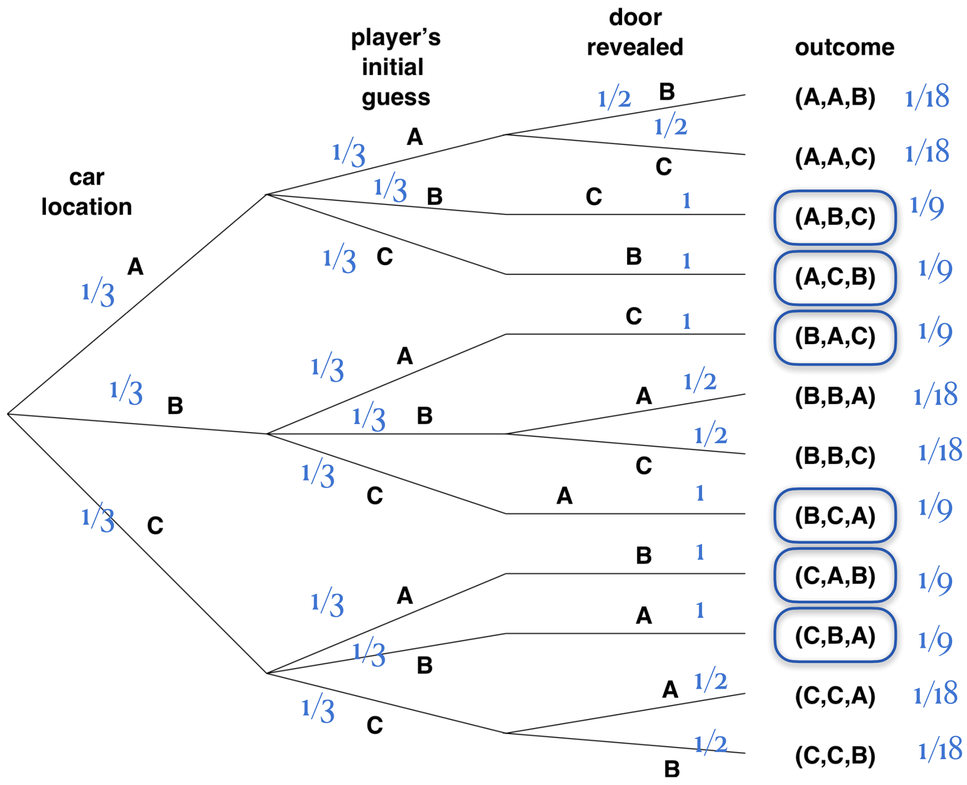
$$S = \left\{ \begin{array}{cccccc} (A,A,B), & (A,A,C), & (A,B,C), & (A,C,B), & (B,A,C), & (B,B,A), \\ (B,B,C), & (B,C,A), & (C,A,B), & (C,B,A), & (C,C,A), & (C,C,B) \end{array} \right\}$$



Determine Outcome Probabilities

- Each outcome is not equally likely!
- To determine probability, assign edge probabilities
    - Edge probabilities are conditional on previous parts of tree!

- $\Pr(A, B, C) = \frac{1}{18}$
- $\Pr(A, A, C) = \frac{1}{18}$
- $\Pr(A, B, C) = \frac{1}{9}$, etc.



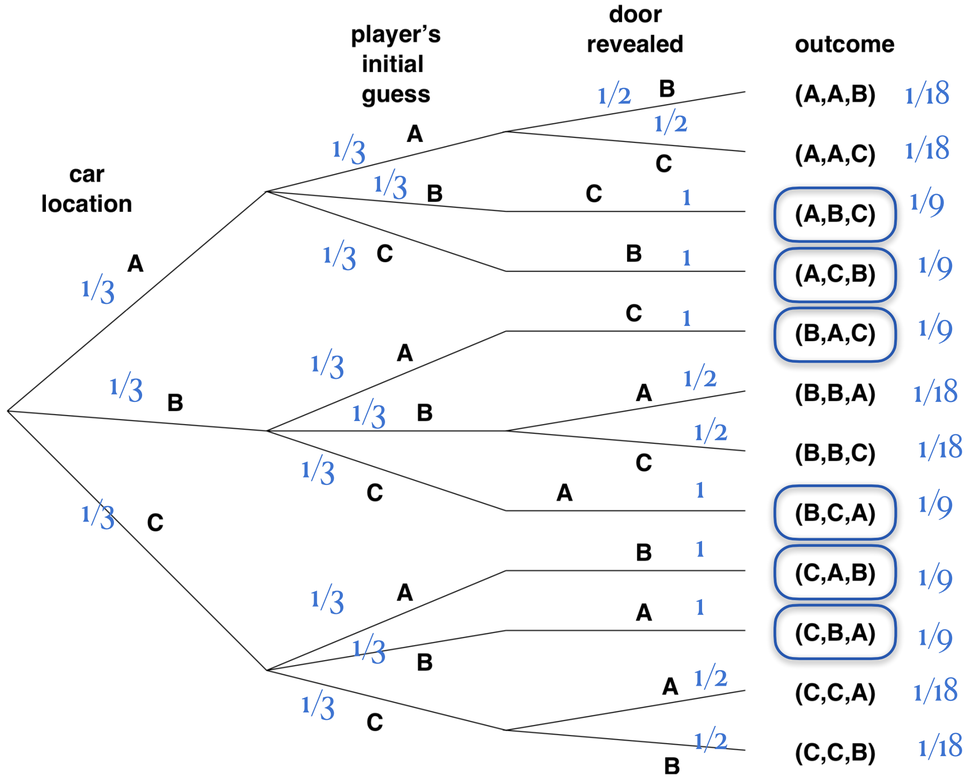
Compute Event Probabilities

- We now have a probability of each outcome
- Probability of an event is the sum of the probabilities of the outcomes it contains, i.e., $\Pr(E) = \sum_{x \in E} \Pr(x)$
- Pr(switching wins) = $\frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{2}{3}$
- It is better to switch!
- Takeaway: resist the intuitively appealing answer

$$S = \left\{ \begin{array}{llllll} (A,A,B), & (A,A,C), & (A,B,C), & (A,C,B), & (B,A,C), & (B,B,A), \\ (B,B,C), & (B,C,A), & (C,A,B), & (C,B,A), & (C,C,A), & (C,C,B) \end{array} \right\}$$

Event (Switching Wins) =

$$\{(A,B,C),(A,C,B),(B,A,C),(B,C,A),(C,A,B),(C,B,A)\}$$



The Birthday Paradox

- Suppose that there are m students in a lecture hall
- Assume for each student, any of the $n = 365$ possible days are equally likely as their birthday
- Assume birthday are mutually independent
- **Question**. What is the likelihood that no two students have the same birthday?
- Let A_i be the event that the i^{th} persons birthday is different from the previous $i - 1$ people
- Pr (all m different birthdays)

$$= \Pr(A_1 \cap A_2 \cap \ldots \cap A_m)$$

$$= \Pr(A_1) \cdot \Pr(A_2 \mid A_1) \cdot \Pr(A_3 \mid A_1 \cap A_2) \ldots \Pr(A_n \mid A_1 \cap \ldots \cap A_{n-1})$$

The Birthday Paradox

- Pr (all m different birthdays)

$$= 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdot \left(1 - \frac{3}{n}\right) \ldots \left(1 - \frac{m-1}{n}\right)$$

$$= \prod_{j=1}^{m-1} \left(1 - \frac{j}{n}\right) \leq \prod_{j=1}^{m-1} e^{-j/n} \approx e^{-m^2/2n}$$

- $m \approx \sqrt{2n \ln 2}$ for probability to be $1/2$
- For $n = 365$, we get $m = 22.49$
- Thus, with around 23 people in this class, we have a 50% chance of two people having the same birthday

Death-bed Inequality:

$$(1 - x) \leq \left(\frac{1}{e}\right)^x \text{ for } x \geq 1$$

Birthday problem

From Wikipedia, the free encyclopedia

For yearly variation in mortality rates, see birthday effect. For the mathematical brain teaser that was asked in the Math Olympiad, see Cheryl's Birthday.

In probability theory, the **birthday problem** or **birthday paradox** concerns the probability that, in a set of n randomly chosen people, some pair of them will have the same birthday. By the pigeonhole principle, the probability reaches 100% when the number of people reaches 367 (since there are only 366 possible birthdays, including February 29). However, 99.9% probability is reached with just 70 people, and 50% probability with 23 people. These conclusions are based on the assumption that each day of the year (excluding February 29) is equally probable for a birthday.

Actual birth records show that different numbers of people are born on different days. In this case, it can be shown that the number of people required to reach the 50% threshold is 23 *or fewer*.[1] For example, if half the people were born on one day and the other half on another day, then any *two* people would have a 50% chance of sharing a birthday.

It may well seem surprising that a group of just 23 individuals is required to reach a probability of 50% that at least two individuals in the group have the same birthday: this result is perhaps made more plausible by considering that the comparisons of birthday will actually be made between every possible pair of individuals = 23 × 22/2 = 253 comparisons, which is well over half the number of days in a year (183 at most), as opposed to fixing on one individual and comparing his or her birthday to everyone else's. The birthday problem is not a "paradox" in the literal logical sense of being self-contradictory, but is merely unintuitive at first glance.

Real-world applications for the birthday problem include a cryptographic attack called the birthday attack, which uses this probabilistic model to reduce the complexity of finding a collision for a hash function, as well as calculating the approximate risk of a hash collision existing within the hashes of a given size of population.

Acknowledgments

- Some of the material in these slides are taken from
  - Kleinberg Tardos Slides by Kevin Wayne (https://www.cs.princeton.edu/~wayne/kleinberg-tardos/pdf/04GreedyAlgorithmsI.pdf)
  - Jeff Erickson's Algorithms Book (http://jeffe.cs.illinois.edu/teaching/algorithms/book/Algorithms-JeffE.pdf)
  - Hamiltonian cycle reduction images from Michael Sipser's Theory of Computation Book