# Mathematical concepts for computer science

# Events and Probability Spaces

- 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**?
- The letter describes a situation like one faced by contestants in the 1970's game show Let's Make a Deal, hosted by Monty Hall and Carol Merrill.
- Marilyn replied that the contestant should indeed switch.

# Events and Probability Spaces

- The problem became known as the **Monty Hall Problem** and it generated thousands of hours of heated debate.

# Clarifying the Problem

- Craig's original letter to Marilyn vos Savant is a bit vague, so we must make some assumptions in order to have any hope of modeling the game formally.
  - The car is equally likely to be hidden behind each of the three doors.
  - The player is equally likely to pick each of the three 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 player the choice of staying with the original door or switching.
  - If the host has a choice of which door to open, then he is equally likely to select each of them.

"**What is the probability that a player who switches wins the car?**"

# The Four Step Method

- Every probability problem involves some sort of randomized experiment, process, or game. And each such problem involves two distinct challenges:
  1. How do we model the situation mathematically?
  2. How do we solve the resulting mathematical problem?

# The Four Step Method

Step 1: Find the Sample Space

Step 2: Define Events of Interest

Step 3: Determine Outcome Probabilities

Step 4: Compute Event Probabilities

# Step 1: Find the Sample Space

- Our first objective is to identify all the possible outcomes of the experiment.
- A typical experiment involves several randomly-determined quantities.
- For example, the Monty Hall game involves three such quantities:

  1. The door concealing the car.
  2. The door initially chosen by the player.
  3. The door that the host opens to reveal a goat.

# Step 1: Find the Sample Space

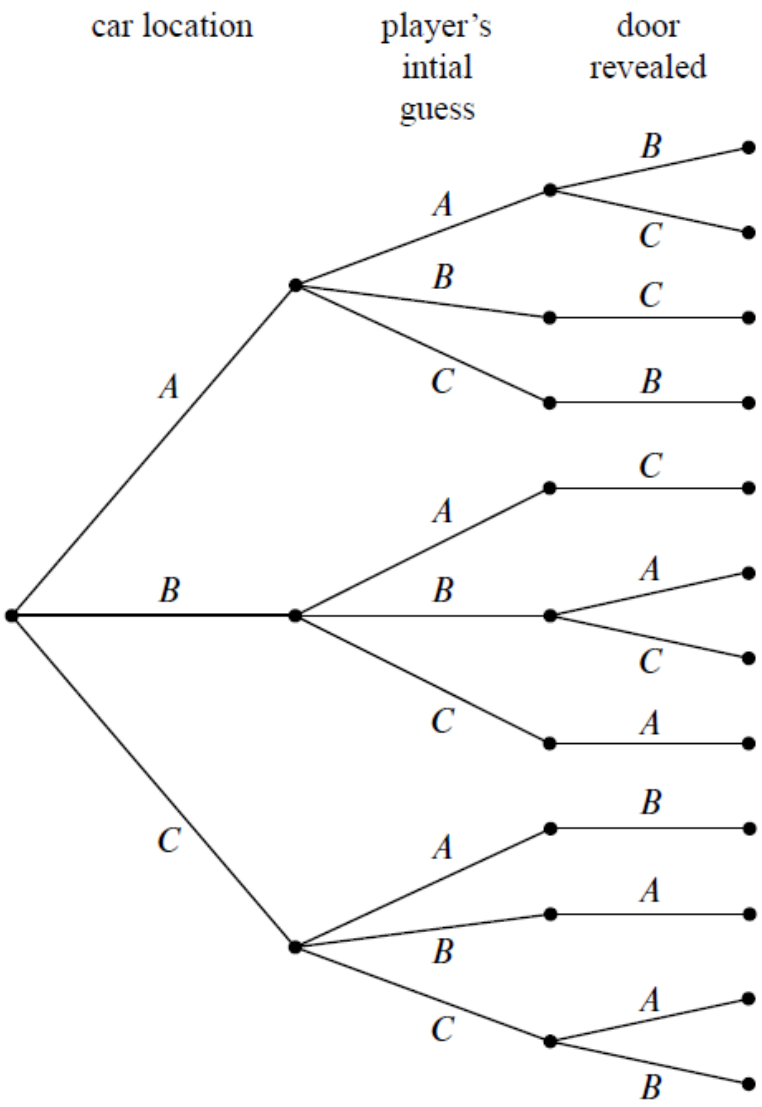
1. The door concealing the car.
2. The door initially chosen by the player.
3. The door that the host opens to reveal a goat.

- Every possible combination of these randomly-determined quantities is called an outcome.
- The set of all possible outcomes is called the **sample space** for the experiment

# Step 1: Find the Sample Space

- A tree diagram is a graphical tool that can help us work through the four step approach when the number of outcomes is not too large or the problem is nicely structured.
- In particular, we can use a tree diagram to help understand the sample space of an experiment.
- The first randomly-determined quantity in our experiment is the door concealing the prize.

# Step 1: Find the Sample Space




The full tree diagram for the Monty Hall Problem.

The second level indicates the door initially chosen by the player.

The third level indicates the door revealed by Monty Hall.

The leaves of the tree represent outcomes of the experiment, and the set of all leaves represents the sample space.

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

# Step 2: Define Events of Interest

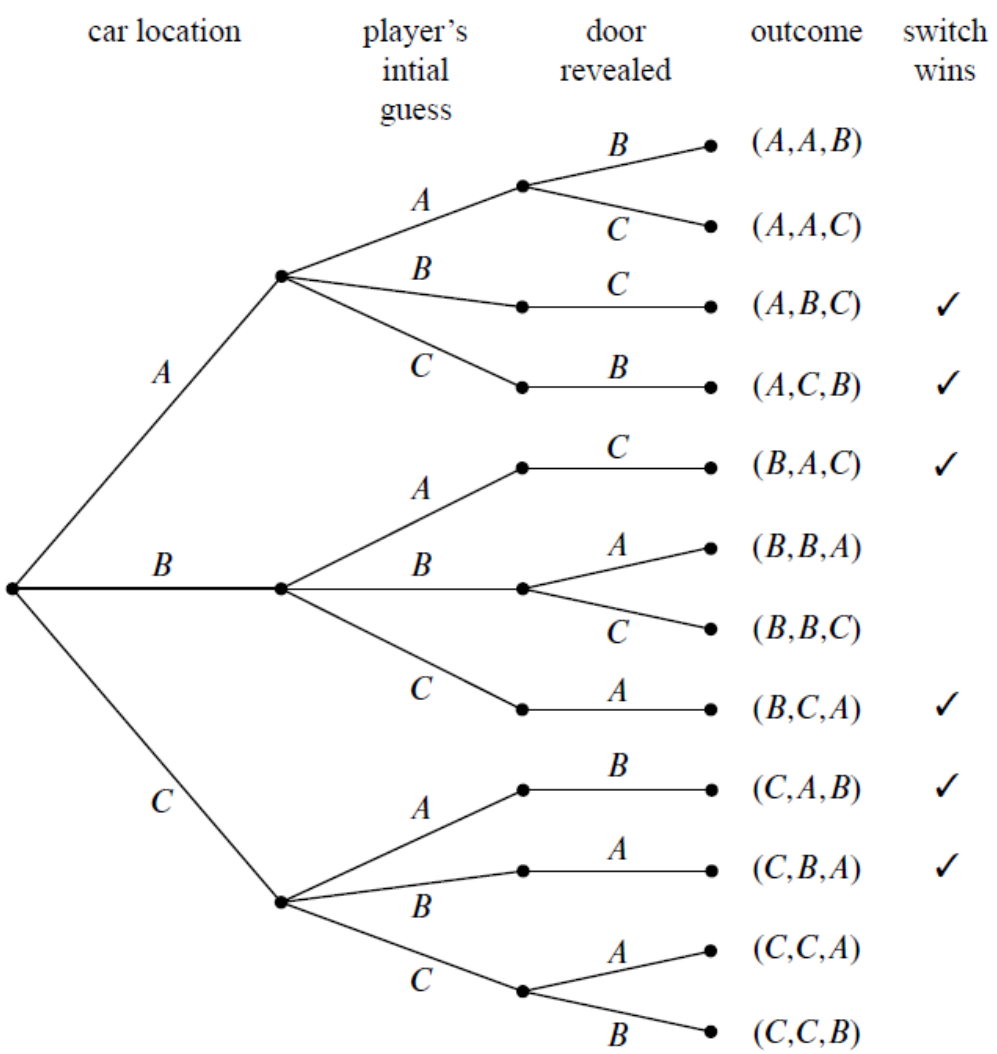
- The event - prize is behind door C refers to the set: $\{(C, A, B), (C, B, A), (C, C, A), (C, C, B)\}$
- The event - prize is behind the door first picked by the player is:

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

- The event - player wins by switching

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

# Step 2: Define Events of Interest




The tree diagram for the Monty Hall Problem, where the outcomes where the player wins by switching are denoted with a check mark.

Notice that exactly half of the outcomes are checked, meaning that the player wins by switching in half of all outcomes. You might be tempted to conclude that a player who switches wins with probability 1/2. This is wrong.

# Step 3: Determine Outcome Probabilities

- The goal of this step is to assign each outcome a probability, indicating the fraction of the time this outcome is expected to occur.
- The sum of all the outcome probabilities must equal one, reflecting the fact that there always must be an outcome.

# Step 3: Determine Outcome Probabilities

- The goal of this step is to **assign each outcome a probability**, indicating the fraction of the time this outcome is expected to occur.
- The sum of all the outcome probabilities must equal one, reflecting the fact that there always must be an outcome.
- Outcome probabilities are determined by the phenomenon we're modeling and thus are not quantities that we can derive mathematically.

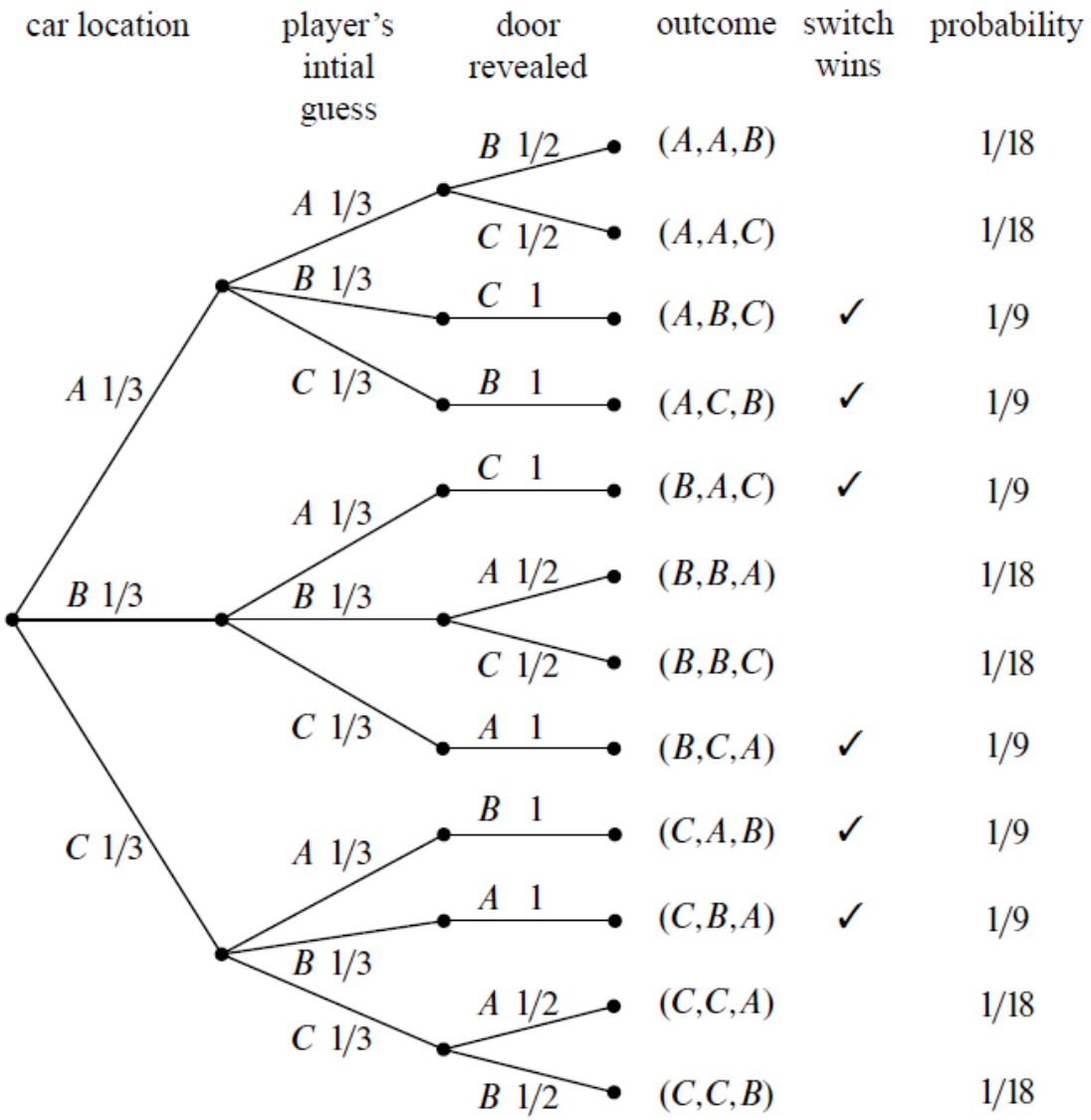
# Step 3: Determine Outcome Probabilities

- We can break the task of determining outcome probabilities into two stages.
  - Assign Edge Probabilities
  - Compute Outcome Probabilities

# Assign Edge Probabilities

- First, we record a probability on each edge of the tree diagram.
- These edge probabilities are determined by the assumptions we made at the outset: that the **prize is equally likely to be behind each door**, **that the player is equally likely to pick each door**, and that **the host is equally likely to reveal each goat**, if he has a choice.
- Notice that when the host has no choice regarding which door to open, the single branch is assigned probability 1

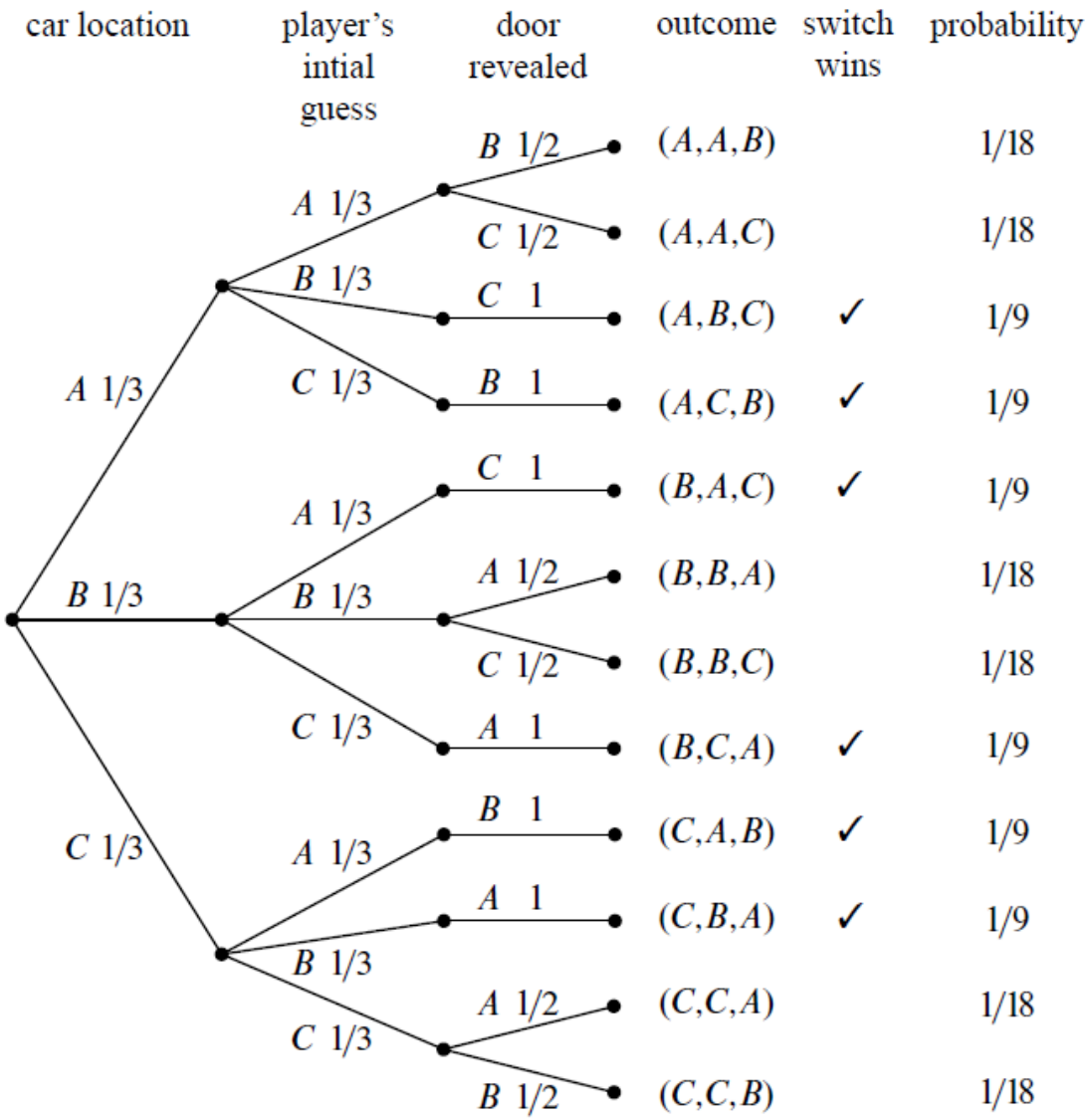
# Assign Edge Probabilities



# Compute Outcome Probabilities

- Convert edge probabilities into outcome probabilities.
- Calculate the probability of an outcome by multiplying the edge-probabilities on the path from the root to that outcome.
- For example, the probability of the topmost outcome (A, A, B), is $\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{18}$.

# Compute Outcome Probabilities

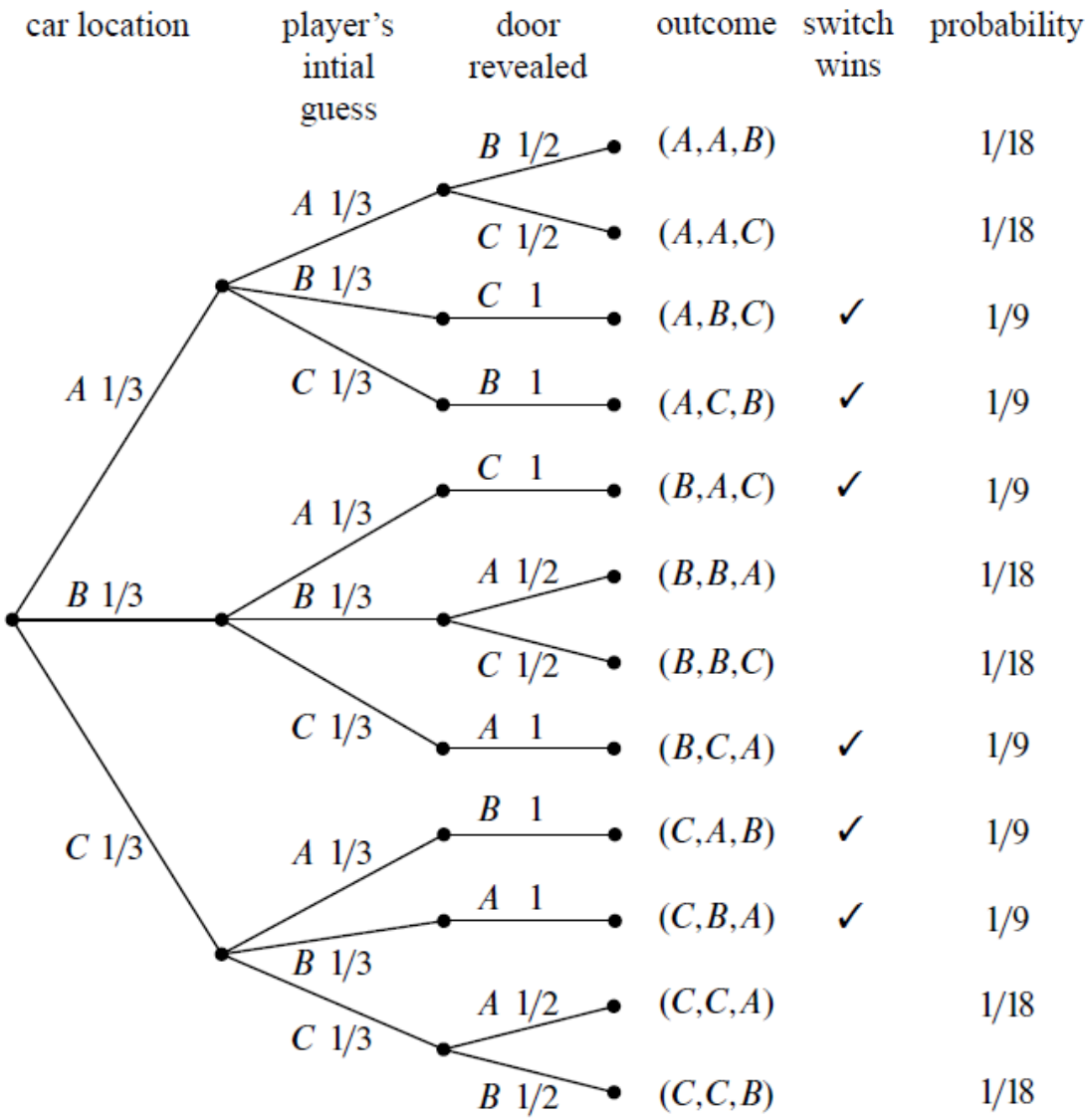


# Step 4: Compute Event Probabilities

- we want to determine the probability of an **event**.
- The probability of an event E is denoted by Pr[E], and it is the sum of the probabilities of the outcomes in E.
- The probability of the [switching wins] event

$$
\begin{aligned}
&\Pr[\text{switching wins}] \\
&= \Pr[(A,B,C)] + \Pr[(A,C,B)] + \Pr[(B,A,C)] + \\
&\quad \Pr[(B,C,A)] + \Pr[(C,A,B)] + \Pr[(C,B,A)] \\
&= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} \\
&= \frac{2}{3}.
\end{aligned}
$$

# Step 4: Compute Event Probabilities



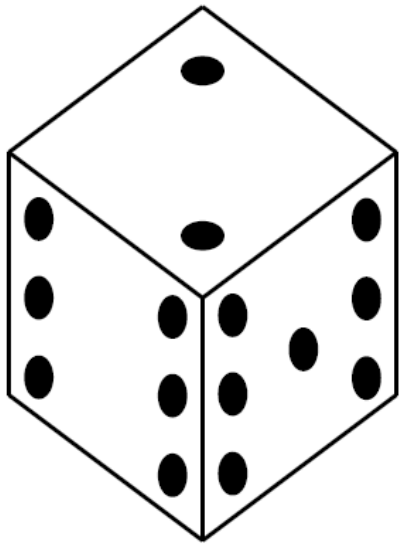
# Monty Hall Problem

- It seems Marilyn's answer is correct!
- A player who switches doors wins the car with probability **2/3**.
- In contrast, a player who stays with his or her original door wins with probability 1/3, since staying wins if and only if switching loses.

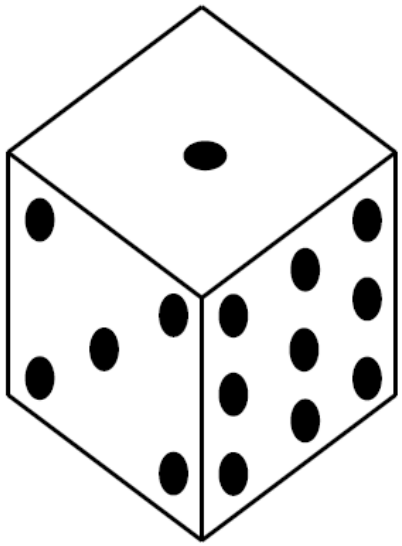
# An Alternative Interpretation of the Monty Hall Problem

- Notice that Craig Whitaker's original letter **does not say** that the **host is required to reveal a goat and offer the player the option to switch**, merely that he did these things.
- In fact, on the Let's Make a Deal show, Monty Hall sometimes simply opened the door that **the contestant picked initially**.

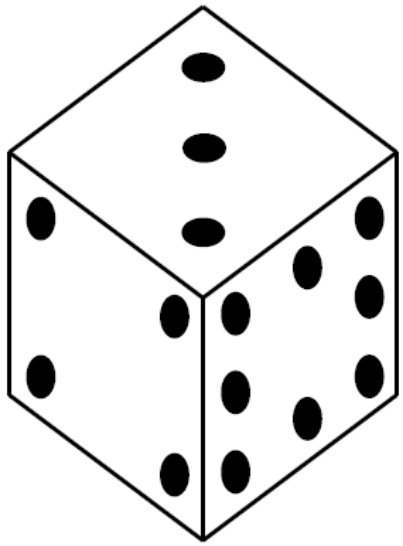
# Strange Dice



*A*



*B*

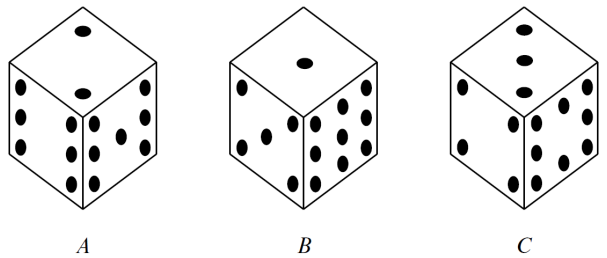


*C*

The number of pips on each concealed face is the same as the number on the opposite face. For example, when you roll die A, the probabilities of getting a 2, 6, or 7 are each 1/3.

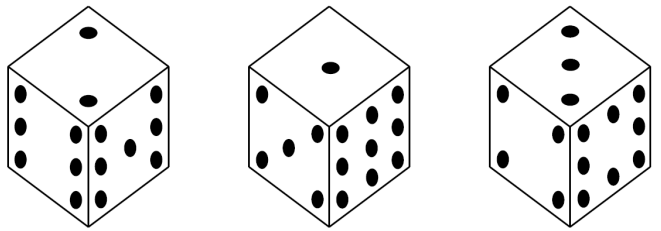
# Strange Dice

- You choose die B because it has a 9, and then other guy selects die A.
- Let's see what the probability is that you will win.



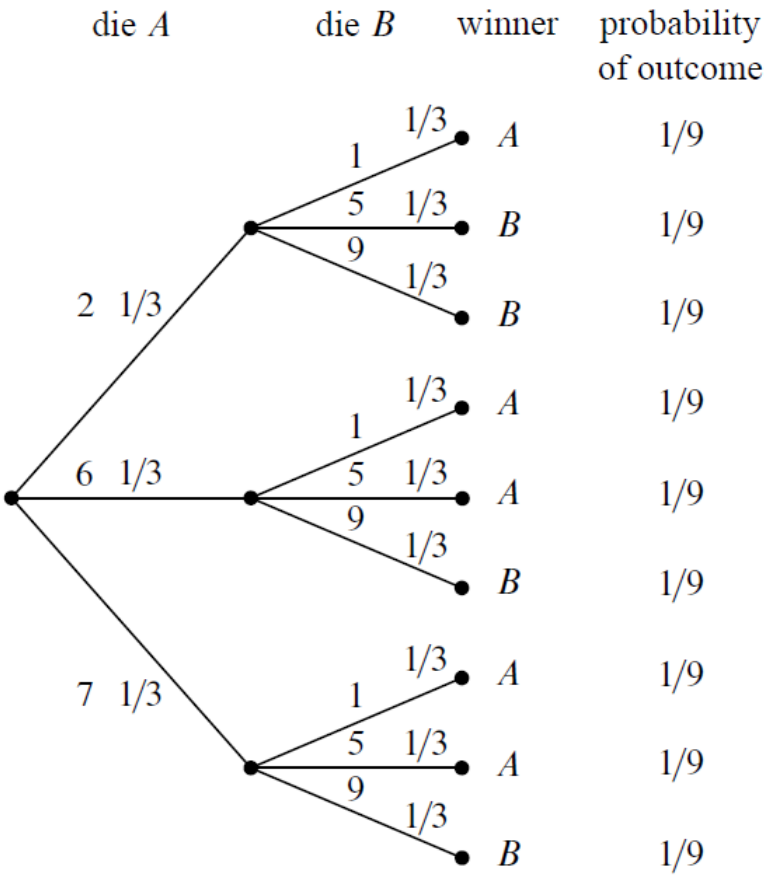
# Die A versus Die B

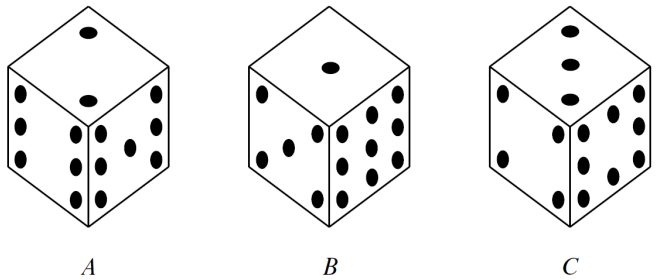
- Step 1: Find the sample space.
  - S = {(2, 1), (2, 5), (2, 9), (6, 1), (6, 5), (6, 9), (7, 1), (7, 5), (7, 9)}
- Step 2: Define events of interest.
  - We are interested in the event that the number on die A is greater than the number on die B.
  - {(2, 1), (6, 1), (6, 5), (7, 1), (7, 5)}



# Die A versus Die B

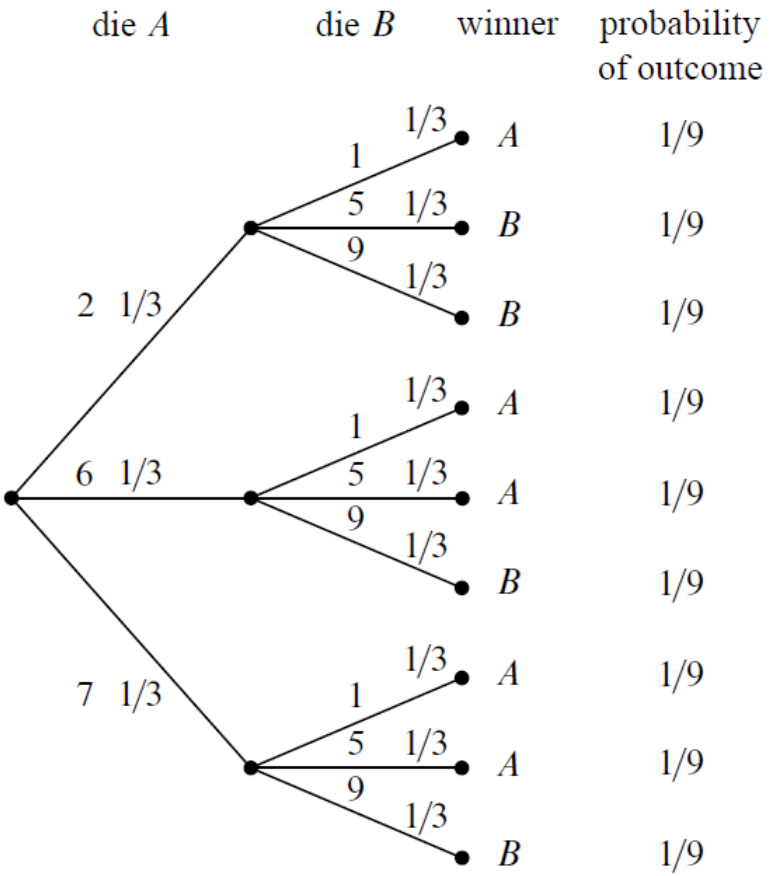
- Step 3: Determine outcome probabilities.



# Die A versus Die B

- Step 4: Compute event probabilities.



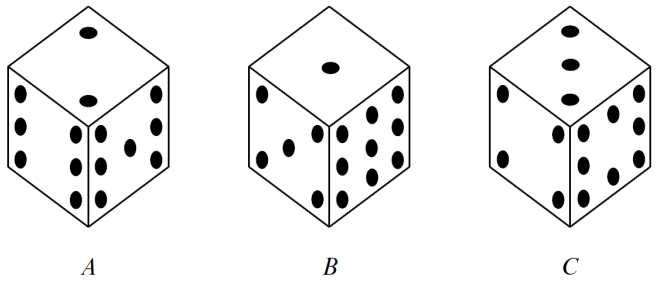

Any event E in a uniform sample space S

$$\Pr[E] = \frac{|E|}{|\mathcal{S}|}.$$

In this case, E is the event that die A beats die B, so |E| = 5, |S| = 9,

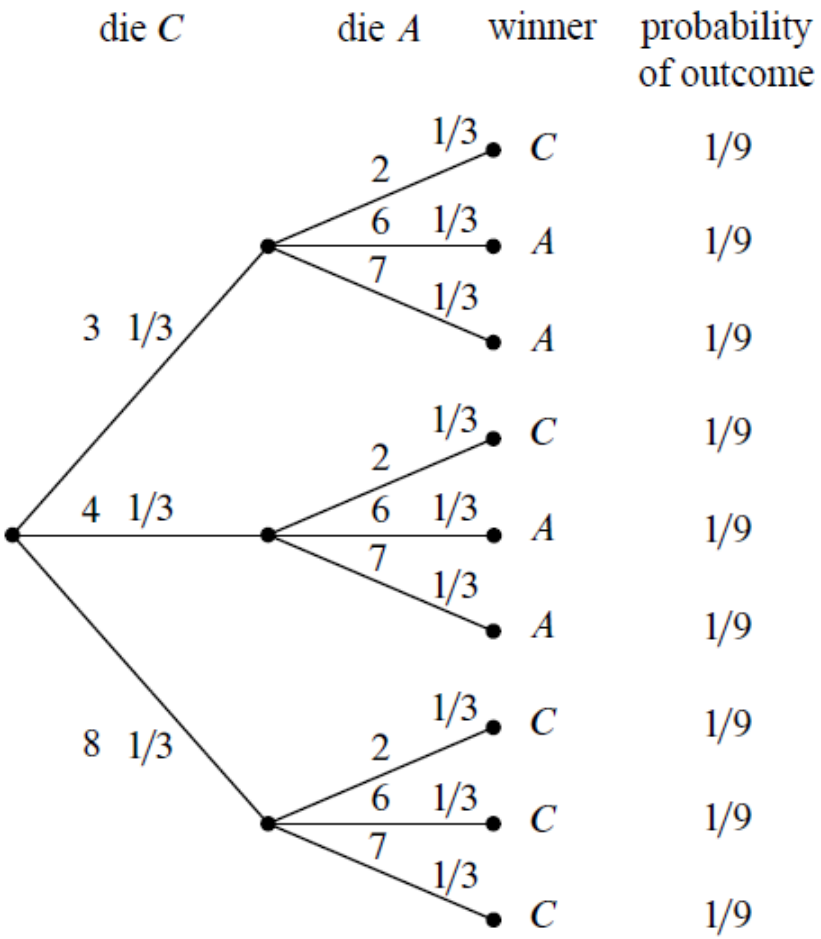
$$\Pr[E] = 5/9.$$

Die A beats die B more than half the time



# Die A versus Die C

- You choose A, and then other guy chooses C



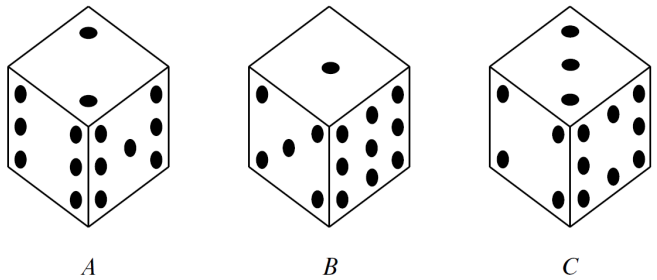

Any event E in a uniform sample space S

$$\Pr[E] = \frac{|E|}{|\mathcal{S}|}.$$

In this case, E is the event that die C beats die A, so |E| = 5, |S| = 9,

$$\Pr[E] = 5/9.$$

Die C beats die A more than half the time



# Die B versus Die C

- You choose C, and then other guy chooses B

$$A \succ B \succ C \succ A$$

whatever die you pick, the other guy can pick one of the others and be likely to win.
So picking first is actually a **disadvantage.**

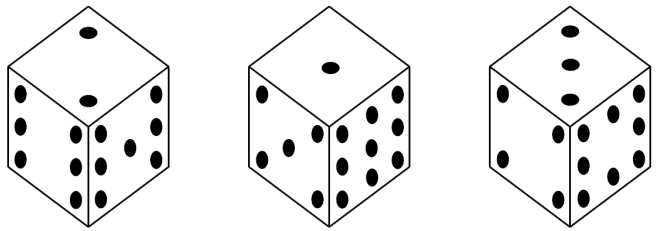
Any event E in a uniform sample space S

$$\Pr[E] = \frac{|E|}{|\mathcal{S}|}.$$

In this case, E is the event that die B beats die C, so |E| = 5, |S| = 9,

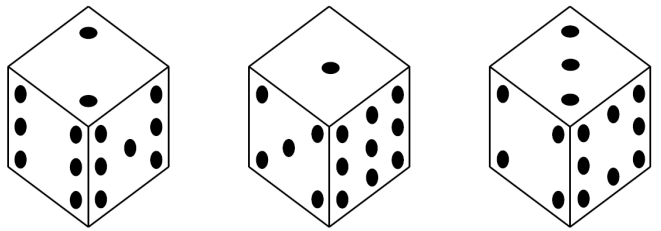
$$\Pr[E] = 5/9.$$

Die B beats die C more than half the time



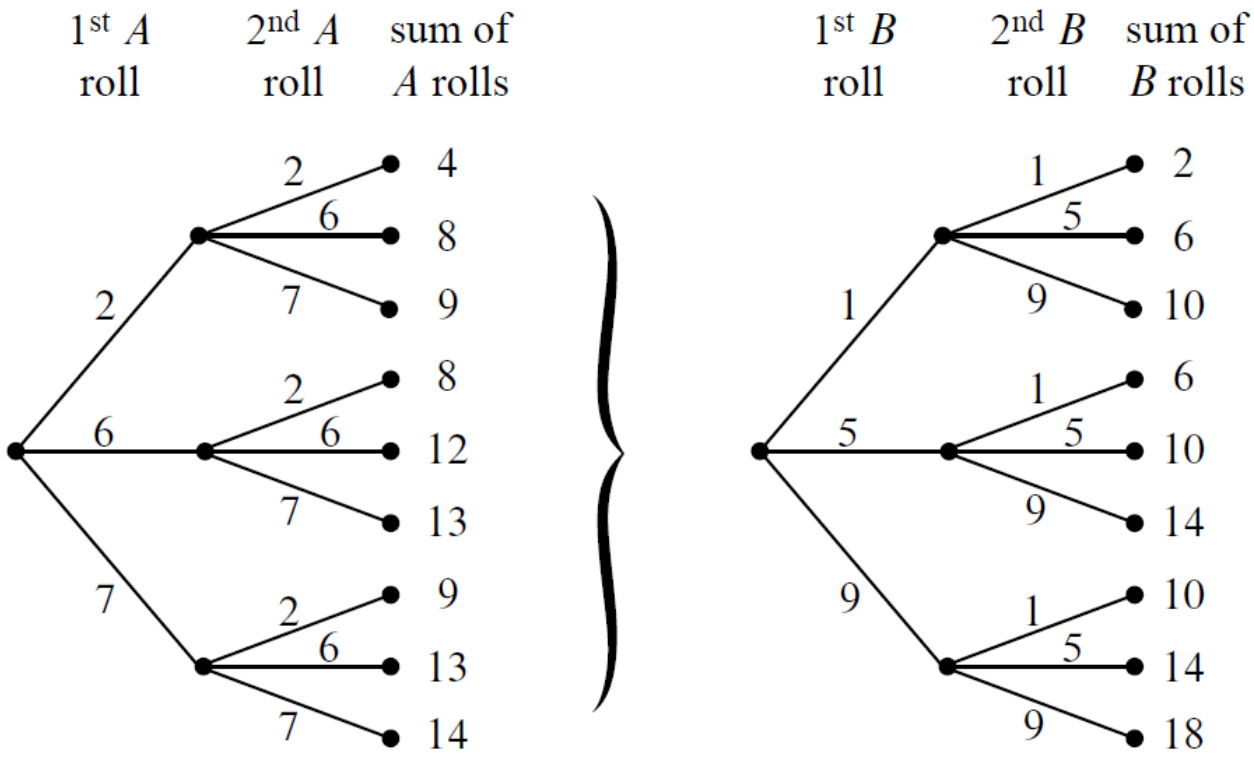
# Rolling Twice

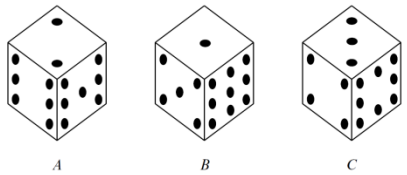
- If each player rolls twice, the tree diagram will have four levels and $3^4 = 81$ outcomes.
- The probability of each outcome is $(1/3)^4 = 1/81$
- The probability that A wins is the number of outcomes where A beats B divided by 81.



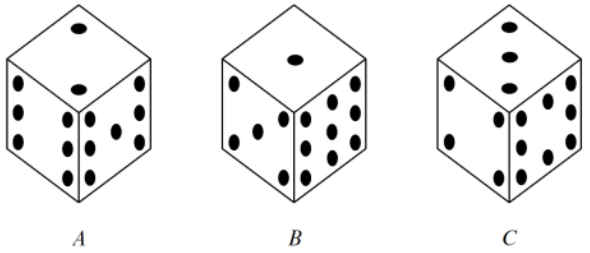
# Rolling Twice

- The other guy chose B, and you chooses A

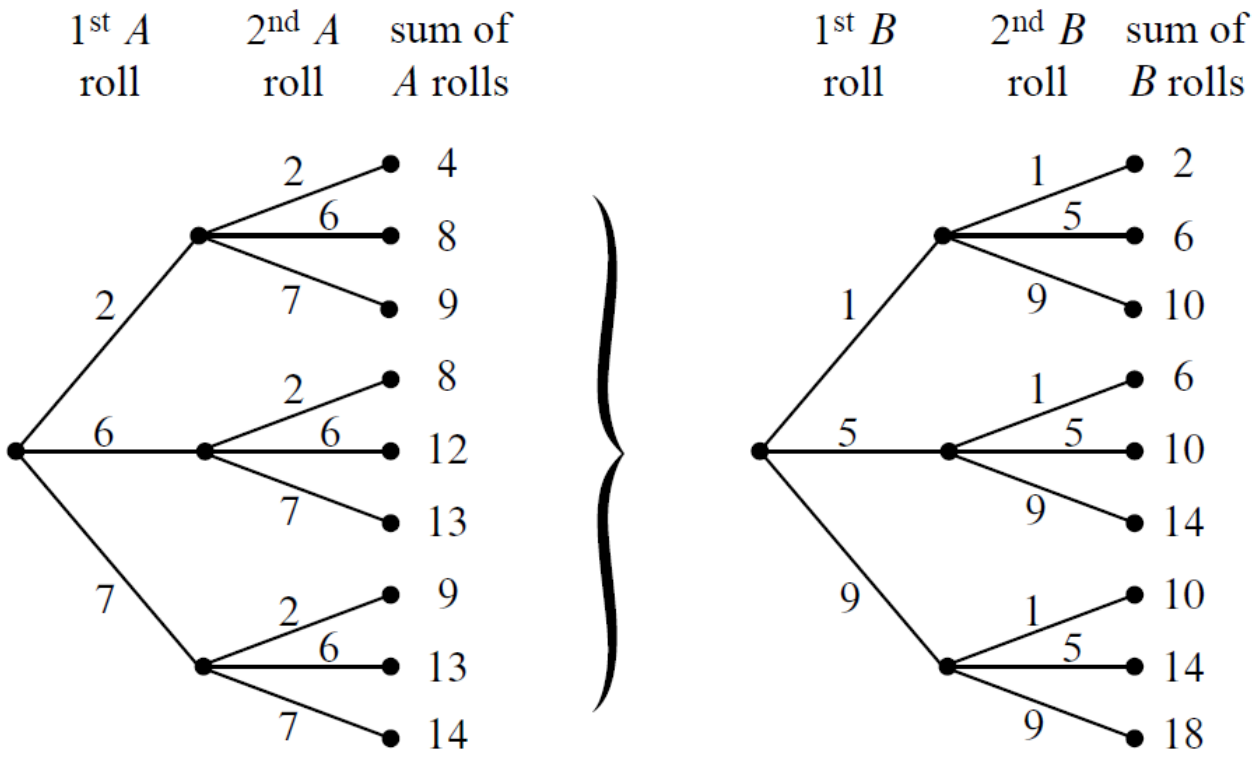


# Rolling Twice

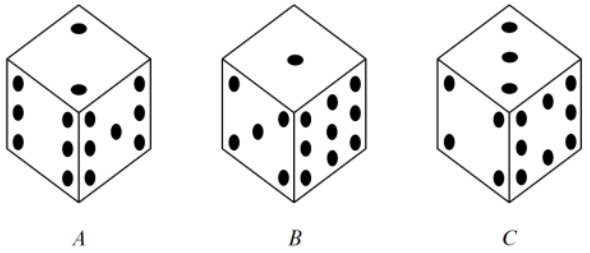



- The other guy chose B, and you chooses A




(4, 8, 8, 9, 9, 12, 13, 13, 14). (2, 6, 6, 10, 10, 10, 14, 14, 18)

# Rolling Twice

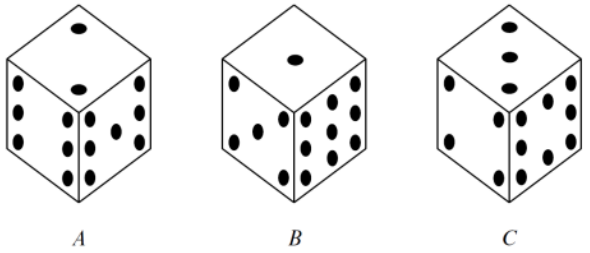



- We can treat the outcome of rolling both dice twice as a pair $(x, y) \in \mathcal{S}_A \times \mathcal{S}_B$, where A wins iff the sum of the two A-rolls of outcome x is larger the sum of the two B-rolls of outcome y.

$\mathcal{S}_A$ $(4, 8, 8, 9, 9, 12, 13, 13, 14)$.

$\mathcal{S}_B$ $(2, 6, 6, 10, 10, 10, 14, 14, 18)$

# Rolling Twice



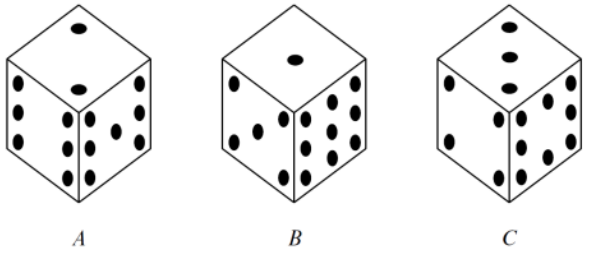

- The number of pairs (x, y) for which the A-sum is larger than the B-sum is

$$1 + 3 + 3 + 3 + 3 + 6 + 6 + 6 + 6 = 37.$$

- A similar count shows that there are 42 pairs for which B-sum is larger than the A-sum, and there are two pairs where the sums are equal, namely, when they both equal 14.

$\mathcal{S}_A$ $(4, 8, 8, 9, 9, 12, 13, 13, 14)$. $\mathcal{S}_B$ $(2, 6, 6, 10, 10, 10, 14, 14, 18)$

# Rolling Twice




- This means that A loses to B with probability 42/81 > 1/2 and ties with probability 2/81.
- Die A wins with probability only 37/81.
- For two rolls,

$$A \prec B \prec C \prec A$$

# The Birthday Principle

- There are 95 students in a class. What is the probability that some birthday is shared by two people?

# The Birthday Principle

- There are 95 students in a class. What is the probability that some birthday is shared by two people?
- Comparing 95 students to the 365 possible birthdays, you might guess the probability lies somewhere around 1/4—but you'd be wrong:
- The probability that there will be two people in the class with matching birthdays is actually more than 0.9999.

# The Birthday Principle

- we'll assume that the probability that a randomly chosen student has a given birthday is ***1/d***.
- We'll also assume that a class is composed of ***n*** randomly and independently selected students.

# Exact Formula for Match Probability

There are $d^n$ sequences of $n$ birthdays, and under our assumptions, these are equally likely. There are $d(d-1)(d-2)\cdots(d-(n-1))$ length $n$ sequences of distinct birthdays. That means the probability that everyone has a different birthday is:[3]

$$\begin{aligned}
&\frac{d(d-1)(d-2)\cdots(d-(n-1))}{d^n} \\
&= \frac{d}{d}\cdot\frac{d-1}{d}\cdot\frac{d-2}{d}\cdots\frac{d-(n-1)}{d} && (16.4)\\
&= \left(1-\frac{0}{d}\right)\left(1-\frac{1}{d}\right)\left(1-\frac{2}{d}\right)\cdots\left(1-\frac{n-1}{d}\right) \\
&< e^0\cdot e^{-1/d}\cdot e^{-2/d}\cdots e^{-(n-1)/d} && (\text{since } 1+x<e^x)\\
&= e^{-\left(\sum_{i=1}^{n-1} i/d\right)} \\
&= e^{-(n(n-1)/2d)}. && (16.5)
\end{aligned}$$

# Exact Formula for Match Probability

- For n = 95 and d = 365, approximate probability value for everyone has different birthday is less than 1/200000, which means the probability of having some pair of matching birthdays actually is more than

**1 – 1/200000 > 0.99999**.

# The Birthday Principle

For $d \leq n^2/2$, the probability of no match turns out to be asymptotically equal to the upper bound (16.5). For $d = n^2/2$ in particular, the probability of no match is asymptotically equal to $1/e$. This leads to a rule of thumb which is useful in many contexts in computer science:

**The Birthday Principle**

If there are $d$ days in a year and $\sqrt{2d}$ people in a room, then the probability that two share a birthday is about $1 - 1/e \approx 0.632$.

For example, the Birthday Principle says that if you have $\sqrt{2 \cdot 365} \approx 27$ people in a room, then the probability that two share a birthday is about 0.632. The actual probability is about 0.626, so the approximation is quite good.

# The Birthday Principle

**The Birthday Principle**

If there are $d$ days in a year and $\sqrt{2d}$ people in a room, then the probability that two share a birthday is about $1 - 1/e \approx 0.632$.

Among other applications, it implies that to use a hash function that maps $n$ items into a hash table of size $d$, you can expect many collisions if $n^2$ is more than a small fraction of $d$.

# Reference


- Eric Lehman, F Thomson Leighton, Albert R Meyer, Mathematics for Computer Science, 1e, MIT, 2010.