Discrete Probability

Warmup

- By the pigeonhole principle we know that it takes 366 students to guarantee that at least two of them have the same birthday
- But how many does it take to guarantee that the probability that at least two of them have the same birthday is at least 90%?

  Note: 90% of 365 is ≈328 days
- What is the probability that in a class of 70 students at least two of them share the same birthday?

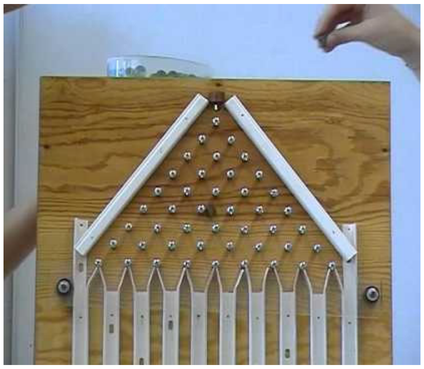
  Note: 70 is ~19% of 365, the number of days in a year
- Not just a toy problem, but a useful model to study practical CS scenarios, e.g., collisions in a hash table

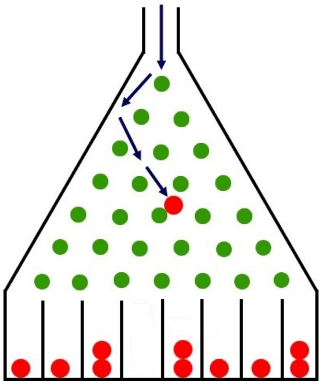
Why Study Probability Theory?

- Probability is used in most areas of science, including computer science
  - One of the most broadly useful part of math
- In algorithmics, it is used to design *randomized algorithms and data structures* that are often simpler and faster than deterministic ones
- In machine learning, it is used in the design of effective learning algorithms
- In signal processing, randomness is used for filtering out noise and compressing data
- In cryptography and digital rights management, probability is used for improving security

Example 1

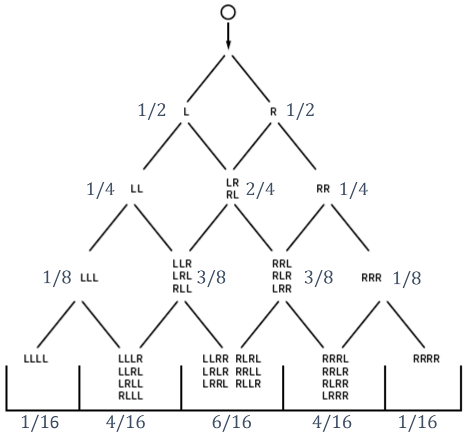
- How will n balls distribute in the following board?

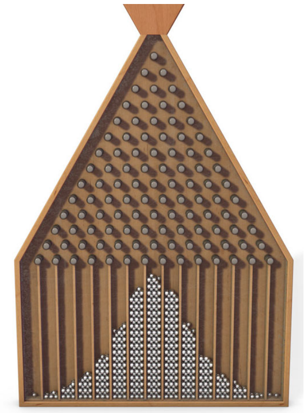




Solution

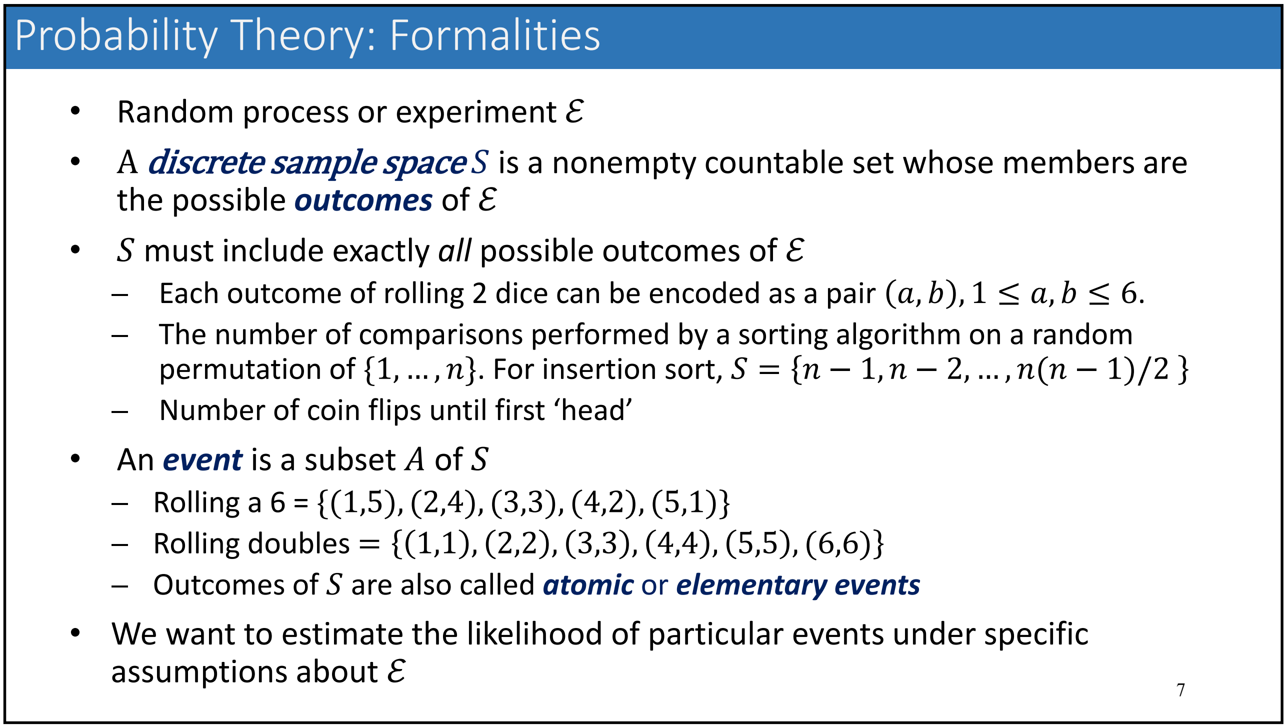
- A ball reaching a peg goes left or right with equal probability





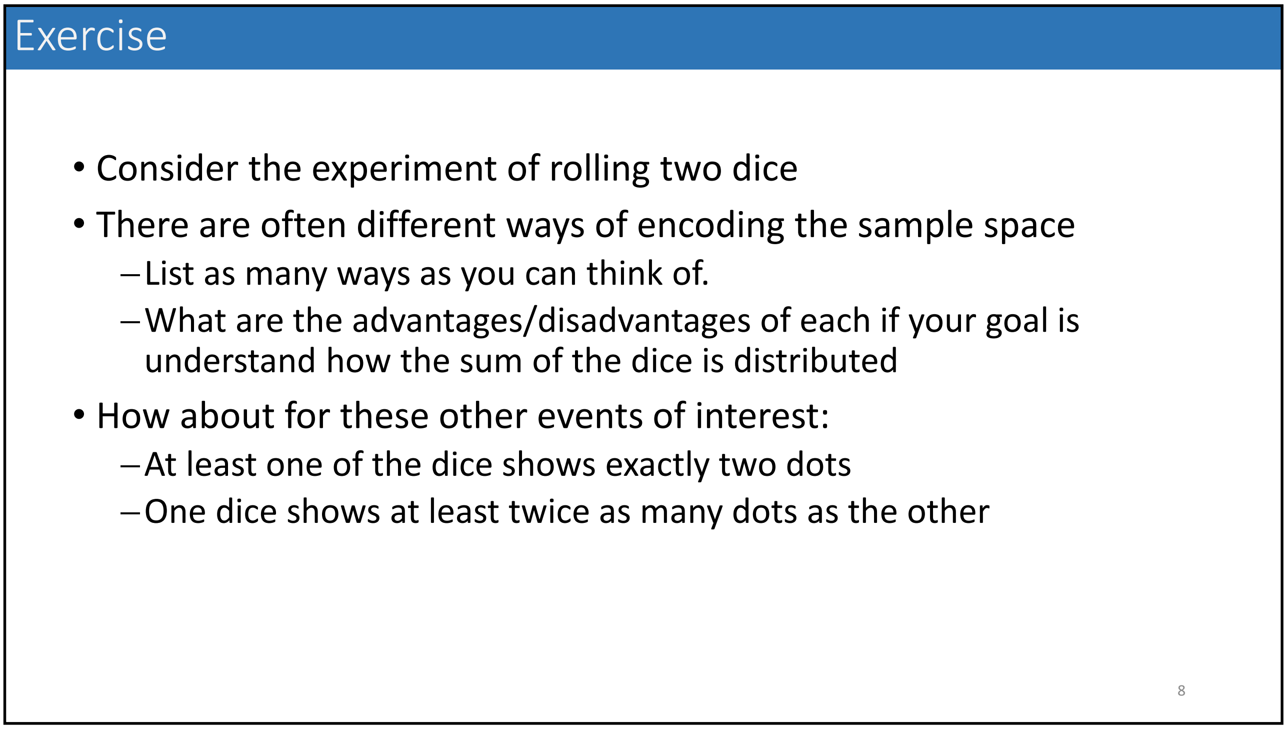
Example 2

- Medical tests are never 100% accurate
- The standard test for tuberculosis attempts to identify carriers (people that have been infected by the tuberculin bacteria)
- Out of 10,000 people, 100 are carriers and 92 of these test positive
- Out of the remaining 9,900 people, 396 test positive
- Furthermore, one out of one hundred people are tuberculosis carriers
- If you test positive, *what is the probability that you are indeed a carrier?*



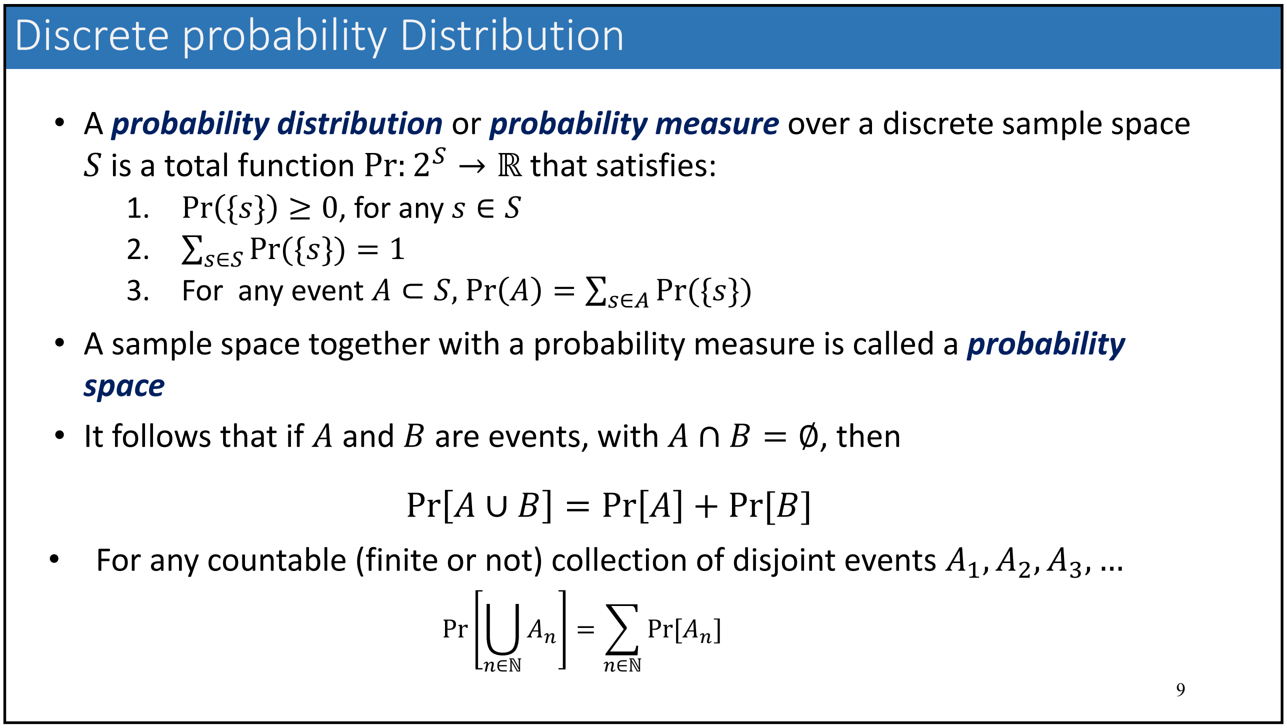
Probability Theory: Formalities

- Random process or experiment $\mathcal{E}$
- A ***discrete sample space*** S is a nonempty countable set whose members are the possible ***outcomes*** of $\mathcal{E}$
- S must include exactly *all* possible outcomes of $\mathcal{E}$
  - Each outcome of rolling 2 dice can be encoded as a pair $(a, b), 1 \leq a, b \leq 6$.
  - The number of comparisons performed by a sorting algorithm on a random permutation of $\{1, \ldots, n\}$. For insertion sort, $S = \{n-1, n-2, \ldots, n(n-1)/2\}$
  - Number of coin flips until first 'head'
- An ***event*** is a subset A of S
  - Rolling a 6 = $\{(1,5), (2,4), (3,3), (4,2), (5,1)\}$
  - Rolling doubles $= \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$
  - Outcomes of S are also called ***atomic*** or ***elementary events***
- We want to estimate the likelihood of particular events under specific assumptions about $\mathcal{E}$



Exercise

- Consider the experiment of rolling two dice
- There are often different ways of encoding the sample space
  - List as many ways as you can think of.
  - What are the advantages/disadvantages of each if your goal is understand how the sum of the dice is distributed
- How about for these other events of interest:
  - At least one of the dice shows exactly two dots
  - One dice shows at least twice as many dots as the other



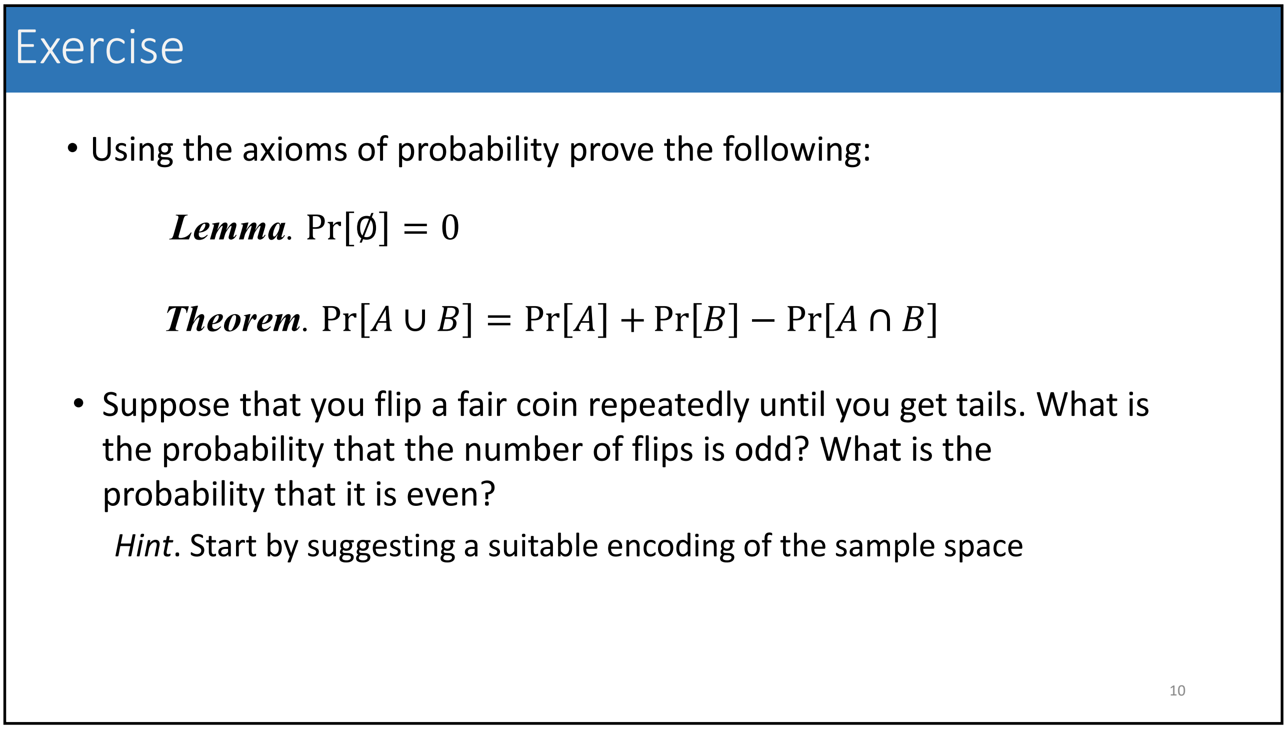
Discrete probability Distribution

- A ***probability distribution*** or ***probability measure*** over a discrete sample space S is a total function $\Pr: 2^S \rightarrow \mathbb{R}$ that satisfies:
  1. $\Pr(\{s\}) \geq 0$, for any $s \in S$
  2. $\sum_{s \in S} \Pr(\{s\}) = 1$
  3. For any event $A \subset S$, $\Pr(A) = \sum_{s \in A} \Pr(\{s\})$
- A sample space together with a probability measure is called a ***probability space***
- It follows that if A and B are events, with $A \cap B = \emptyset$, then

$$\Pr[A \cup B] = \Pr[A] + \Pr[B]$$

- For any countable (finite or not) collection of disjoint events $A_1, A_2, A_3, \ldots$

$$\Pr\left[\bigcup_{n \in \mathbb{N}} A_n\right] = \sum_{n \in \mathbb{N}} \Pr[A_n]$$



Exercise

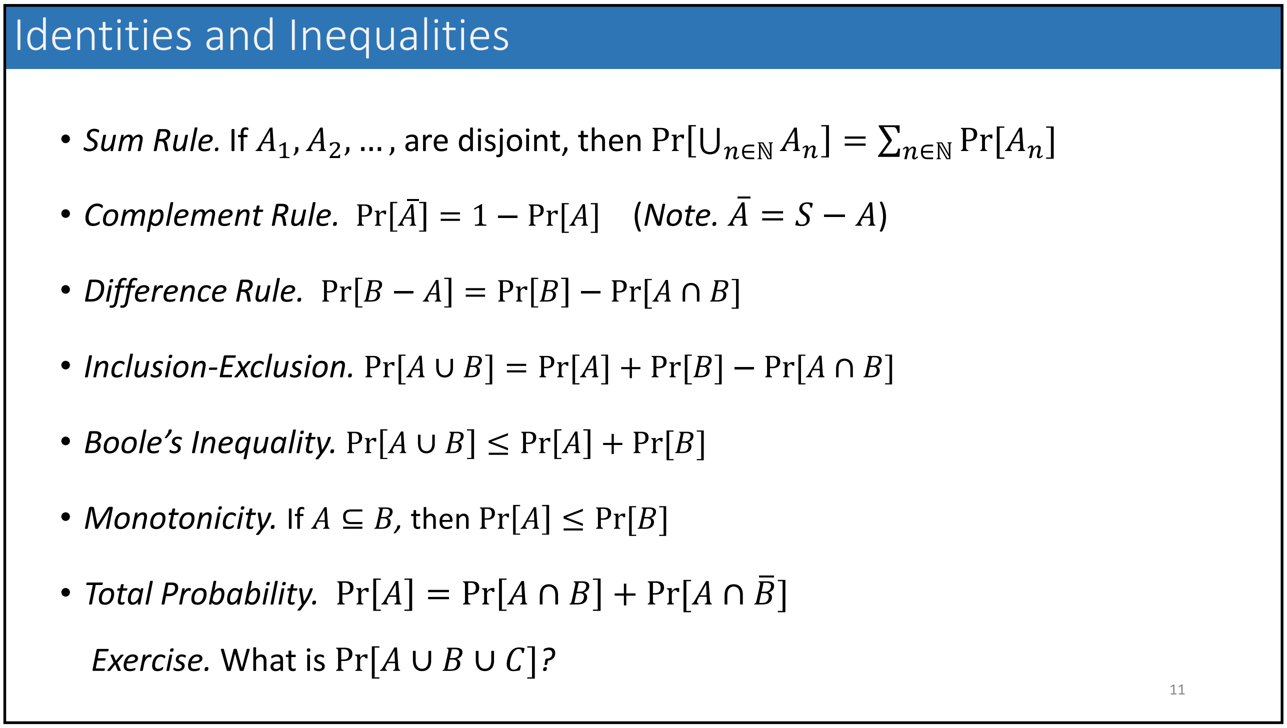
- Using the axioms of probability prove the following:

$$\boldsymbol{Lemma}.\ \Pr[\emptyset] = 0$$

$$\boldsymbol{Theorem}.\ \Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$$

- Suppose that you flip a fair coin repeatedly until you get tails. What is the probability that the number of flips is odd? What is the probability that it is even?

  Hint. Start by suggesting a suitable encoding of the sample space



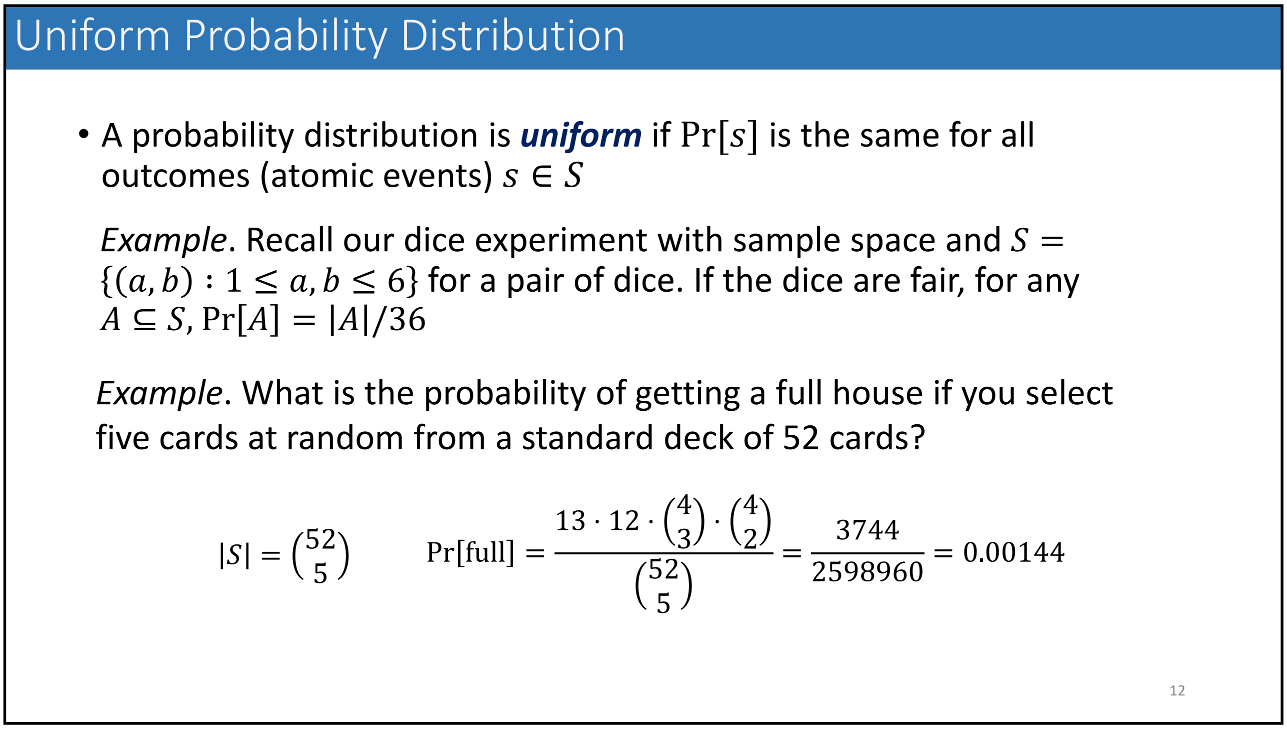
Identities and Inequalities

- *Sum Rule.* If $A_1, A_2, \ldots$, are disjoint, then $\Pr[\bigcup_{n\in\mathbb{N}} A_n] = \sum_{n\in\mathbb{N}} \Pr[A_n]$
- *Complement Rule.* $\Pr[\bar{A}] = 1 - \Pr[A]$ (*Note.* $\bar{A} = S - A$)
- *Difference Rule.* $\Pr[B - A] = \Pr[B] - \Pr[A \cap B]$
- *Inclusion-Exclusion.* $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$
- *Boole's Inequality.* $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$
- *Monotonicity.* If $A \subseteq B$, then $\Pr[A] \leq \Pr[B]$
- *Total Probability.* $\Pr[A] = \Pr[A \cap B] + \Pr[A \cap \bar{B}]$

Exercise. What is $\Pr[A \cup B \cup C]$?

11

11



Uniform Probability Distribution

- A probability distribution is ***uniform*** if $\Pr[s]$ is the same for all outcomes (atomic events) $s \in S$

Example. Recall our dice experiment with sample space and $S = \{(a, b) : 1 \leq a, b \leq 6\}$ for a pair of dice. If the dice are fair, for any $A \subseteq S$, $\Pr[A] = |A|/36$

Example. What is the probability of getting a full house if you select five cards at random from a standard deck of 52 cards?

$$|S| = \binom{52}{5} \qquad \Pr[\text{full}] = \frac{13 \cdot 12 \cdot \binom{4}{3} \cdot \binom{4}{2}}{\binom{52}{5}} = \frac{3744}{2598960} = 0.00144$$

12

12

Exercise

- Consider again the experiment of rolling two fair dice
  - What is the probability of rolling a 4 or doubles?
  - What is the probability that one of the dice shows at least twice as many dots as the other
  - Justify your choice of sample space

- You wish to choose a value in $\{0,1,2,3,4,5,6,7\}$ uniformly at random. To this end you flip a fair coin 7 times and report the number of heads. Does this solve your problem? Explain.

Birthday Problem Revisited

- Given n people and d days in a year there are d^n sequences of n birthdays (n-permutations of $\{\infty \cdot 1, \dots, \infty \cdot 365\}$)
- Of these, $d(d-1)(d-2)\cdots(d-n+1)$ contain different birthdays (n-permutations of $\{1, \dots, 365\}$)
- Under a uniform probability distribution, the probability of no duplicates is:

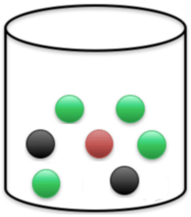
$$\frac{d(d-1)(d-2)\cdots(d-n+1)}{d^n} = \frac{d}{d} \cdot \frac{d-1}{d} \cdot \frac{d-2}{d} \cdots \frac{d-n+1}{d} =$$

$$\left(1-\frac{0}{d}\right)\left(1-\frac{1}{d}\right)\left(1-\frac{2}{d}\right)\cdots\left(1-\frac{n-1}{d}\right) < e^0 \cdot e^{-1/d} \cdot e^{-2/d} \cdots e^{-(n-1)/d} = e^{-\frac{n(n-1)}{2d}}$$

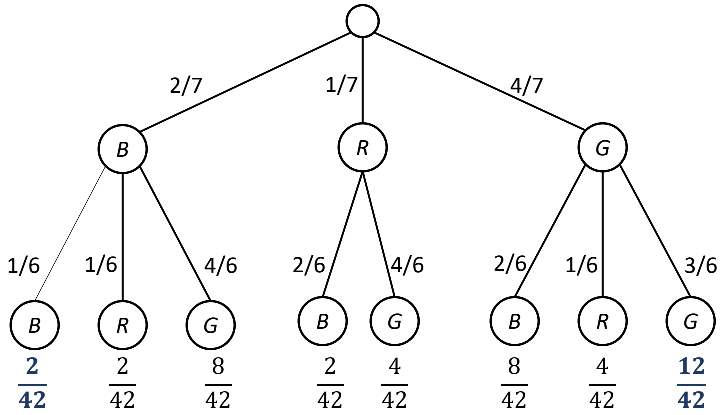
- For $n = 70$, this is $\approx \frac{1}{1200}$

Exercise

- Suppose that an urn contains 2 black balls, 1 red ball, and 4 green balls. Two balls are chosen, without replacement, from the urn. That is, the first ball is chosen, its color is recorded, but the ball is not returned to the urn. Then, the second ball is chosen.
- Assume that all (remaining) balls are equally likely to be chosen.
- What is the probability of choosing two balls of the same color



Tree Diagram



- Initially, urn contains 2 black, 1 red, 4 green
- $\Pr[E] = \frac{2}{42} + \frac{12}{42} = \frac{1}{3}$, where E = draw two balls of same color

Methodology

1. Define the sample space, i.e., all possible outcomes
   - May involve several random choices
   - Model S using a *tree diagram*
   - The sample space is $S = \{BB, BR, BG, RB, RG, GB, GR, GG\}$
2. Define the event of interest
   - $E = \{BB, GG\}$
3. Determine the outcome probabilities
   - Label each branch with a probability and, for each leaf, multiply the probabilities on path from leaf to root
4. Compute the event probabilities
   - Add up the probabilities of outcome nodes in E

An infinite sample space

- Infinite but countable sample spaces are common in discrete probability

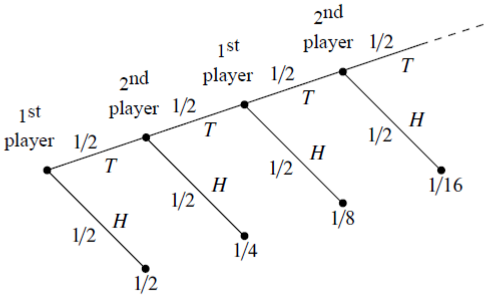
  Example. Two people take turns flipping a fair coin. Whoever flips heads first wins. What is the probability that the first player wins?

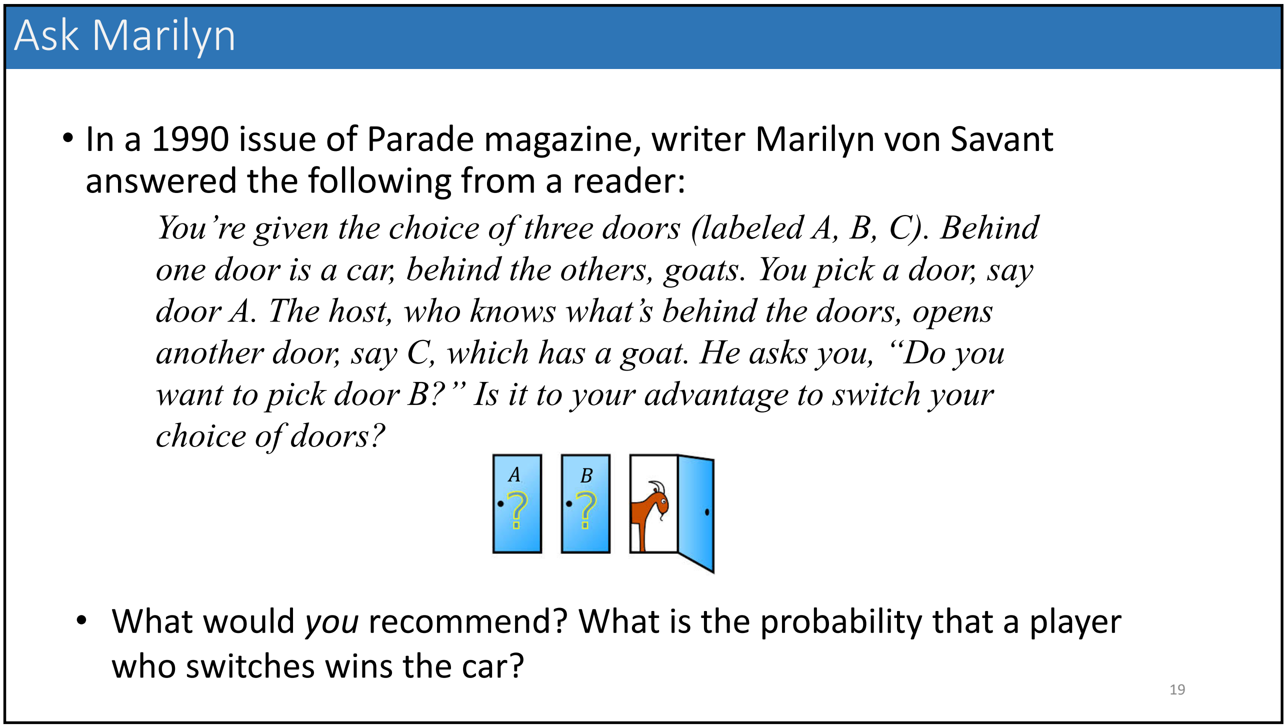
  – Both the sample space S and the event W that the first player wins contain an infinite number of outcomes

$$S = \{H, TH, TTH, TTTH, \dots\}$$

$$W = \{H, TTH, TTTTH, \dots\}$$

$$\Pr[W] = \frac{1}{2}\sum_{i=0}^{\infty}\frac{1}{4^i} = \frac{2}{3}$$



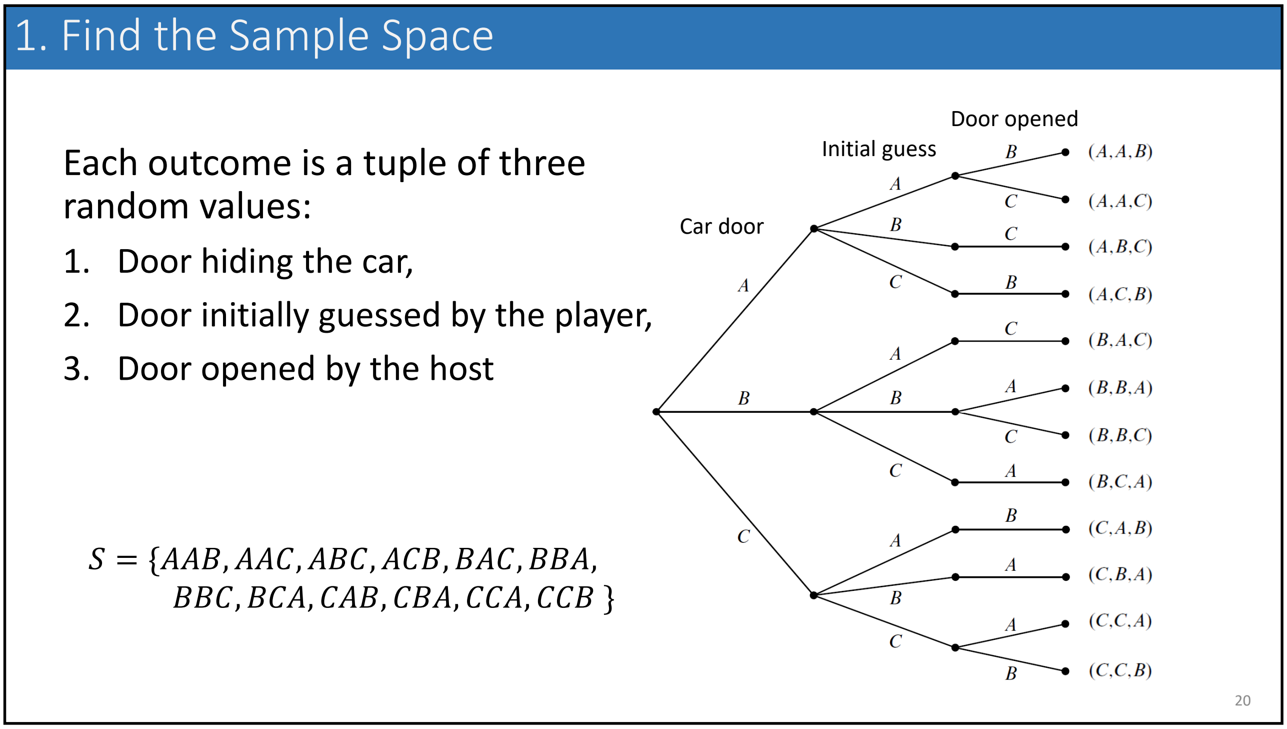


Ask Marilyn

- In a 1990 issue of Parade magazine, writer Marilyn von Savant answered the following from a reader:

> *You're given the choice of three doors (labeled A, B, C). Behind one door is a car, behind the others, goats. You pick a door, say door A. The host, who knows what's behind the doors, opens another door, say C, which has a goat. He asks you, "Do you want to pick door B?" Is it to your advantage to switch your choice of doors?*

- What would *you* recommend? What is the probability that a player who switches wins the car?



1. Find the Sample Space

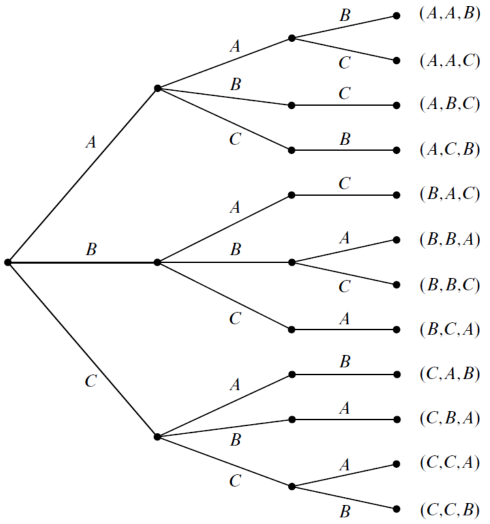
Each outcome is a tuple of three random values:

1. Door hiding the car,
2. Door initially guessed by the player,
3. Door opened by the host

$$S = \{AAB, AAC, ABC, ACB, BAC, BBA, BBC, BCA, CAB, CBA, CCA, CCB\}$$

2. Find the Relevant Events

- The event W we are interested in is “winning by switching”
- When does this happen?

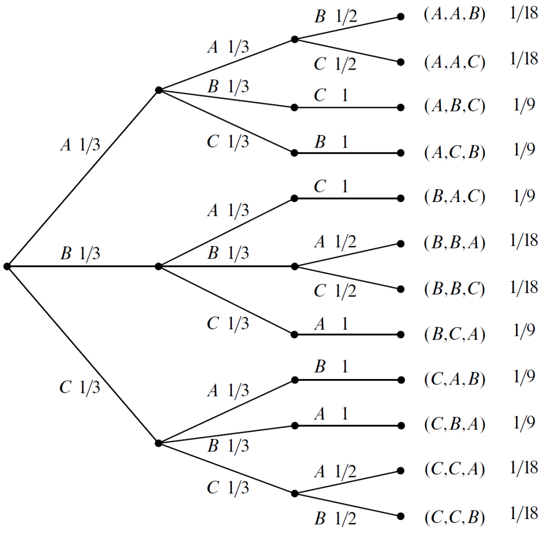


$W = \{ABC, ACB, BAC, BCA, CAB, CBA\}$

3. Determine Outcome Probabilities

- Based on model assumptions, assign a probability to each outcome
  1. Assign *edge probabilities*
  2. Compute *outcome probabilities* by multiplying edge probabilities on path from root to leaves

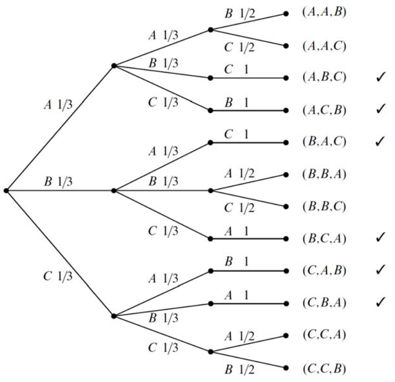
$W = \{ABC, ACB, BAC, BCA, CAB, CBA\}$



4. Compute Event Probabilities

- To compute the probability of an event A we add up the probabilities of all outcomes in A

- Since $W = \{ABC, ACB, BAC, BCA, CAB, CBA\}$,



$$\begin{aligned}\Pr[W] &= \Pr[ABC] + \Pr[ACB] \\ &+ \Pr[BAC] + \Pr[BCA] \\ &+ \Pr[CAB] + \Pr[CBA] \\ &= 6(1/9) = 2/3\end{aligned}$$

Exercise

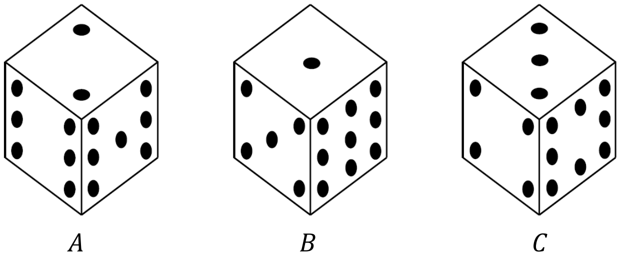
- Consider a set of keys $K \subset U$ of k keys to be stored in a hash table of size m
- A hash function h is chosen at random from the set $\mathcal{H}$ of all possible functions $K \mapsto \{0, \dots, m-1\}$

1. How big is $\mathcal{H}$?
2. What is the probability that 2 given keys collide?
3. What is the probability that all keys hash to the same slot?
4. What is the probability that the hash function is "perfectly balanced" (to simplify, assume $k = c \cdot m$)

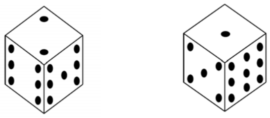
   (Maybe try $k = 4, m = 2$ first)

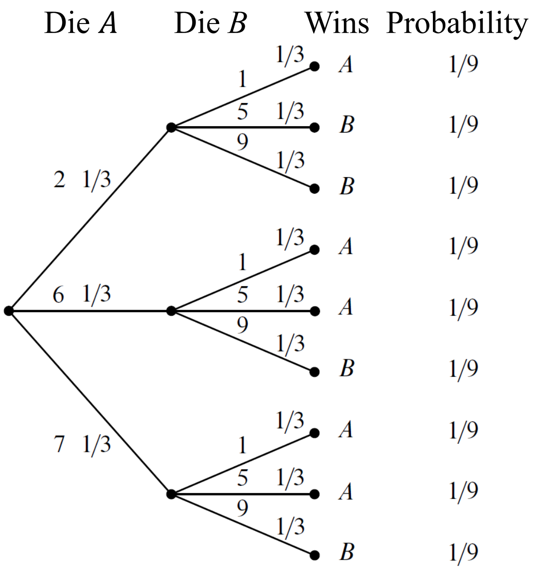
A Dicey Game

- Who is willing to play the following game?
  1. First, you choose one of the three dice below
  2. Second, I choose one of the remaining two dice
  3. We each roll our die
  4. Higher value wins



Die A vs. Die B

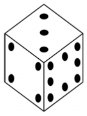


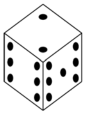


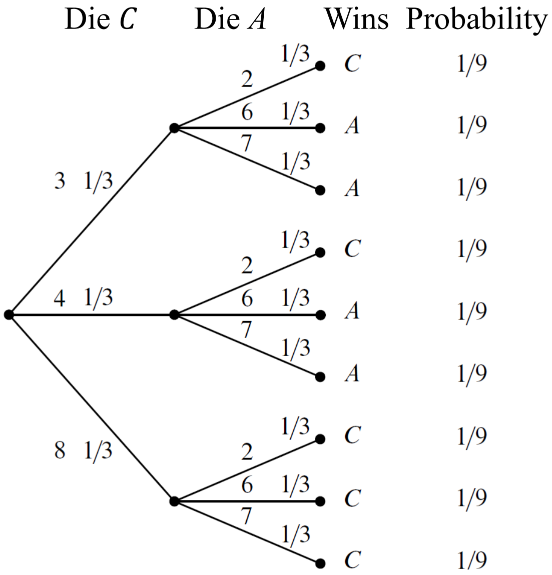
Die A	Die B	Wins	Probability
2 1/3	1 1/3	A	1/9
	5 1/3	B	1/9
	9 1/3	B	1/9
6 1/3	1 1/3	A	1/9
	5 1/3	A	1/9
	9 1/3	B	1/9
7 1/3	1 1/3	A	1/9
	5 1/3	A	1/9
	9 1/3	B	1/9

If you choose B, you lose

Die C vs. Die A



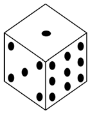


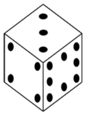


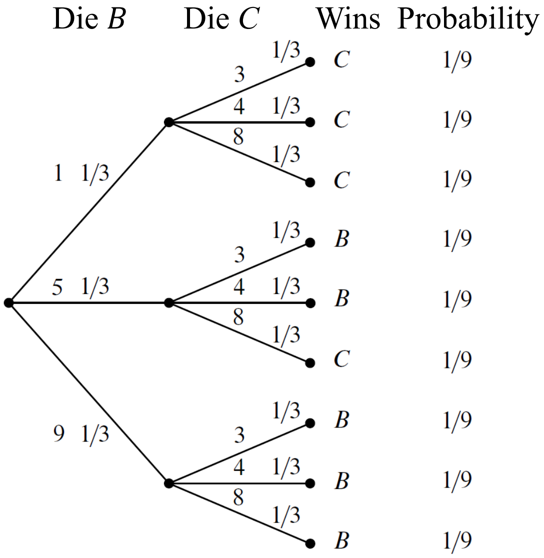
Die C		Die A		Wins	Probability
3	1/3	2	1/3	C	1/9
		6	1/3	A	1/9
		7	1/3	A	1/9
4	1/3	2	1/3	C	1/9
		6	1/3	A	1/9
		7	1/3	A	1/9
8	1/3	2	1/3	C	1/9
		6	1/3	C	1/9
		7	1/3	C	1/9

If you choose A, you lose

Die B vs. Die C







Die B		Die C		Wins	Probability
1	1/3	3	1/3	C	1/9
		4	1/3	C	1/9
		8	1/3	C	1/9
5	1/3	3	1/3	B	1/9
		4	1/3	B	1/9
		8	1/3	C	1/9
9	1/3	3	1/3	B	1/9
		4	1/3	B	1/9
		8	1/3	B	1/9

If you choose C, you lose

Example

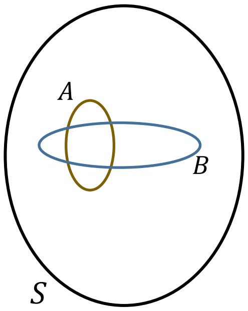
- You are a prisoner sentenced to death. The king offers you a chance to live by playing a simple game. He gives you 15 black balls, 15 white balls, and 2 empty boxes.
- You can distribute the balls between the boxes as you like, provided that no box is empty.
- The king will first pick one of the boxes at random and then pick a random ball from the selected box.
- If the ball is white, you live.

Question. How should you distribute the balls to maximize the probability of survival?

Exercise. (1) Prove that your solution is optimal or find a better solution. (2) What is your probability of survival?

Conditional Probability

- Suppose we uniformly pick a random person p, anywhere in the world. Let A be the event that p is a DU student, and B the event that p lives in Denver, with sample space S denoting all people living in the US



- How big are $\Pr[A]$ and $\Pr[B]$?
  - Compared to S, both $\Pr[A]$ and $\Pr[B]$ are small
- How about the probability of B ***given*** that A happened, denoted $\Pr[B|A]$?
  - Knowing that A happened, makes B much more likely
- The ***given*** is merely a directive to focus on a subset of S, namely A. This implies,

$$\boxed{\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]}}$$

Example

- In a best-2-out-of-three final, the DU hockey team wins the first game with probability 1/2. In the ensuing games, the probability is determined by the outcome of the previous game, as follows. If DU won the previous game, then they win the current game with probability 2/3; else, they win with probability 1/3.
- What is the probability that DU wins the final?
- What is the probability that DU wins the final given that they won the first game?
- What is the sample space S?

$$S = \{WW, WLW, WLL, LWW, LWL, LL\}$$

- Events of interest: F = win final, G_1 = win game 1

Exercise. Given events A, B, C, show that if $A \cap B \subseteq B \cap C$, then $\Pr(A|C) \leq P(B|C)$.

Chaining

- How to you compute the probability of a chain of conjunctions?

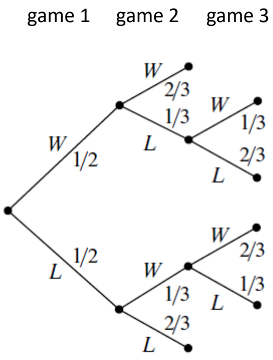
$$\Pr(A \cap B \cap C) = \Pr(A|B \cap C) \cdot \Pr(B|C) \cdot \Pr(C)$$

Proof.

$$\Pr(A \cap B \cap C) = \frac{\Pr(A \cap B \cap C)}{\Pr(B \cap C)} \cdot \frac{\Pr(B \cap C)}{\Pr(C)} \cdot \Pr(C) = \Pr(A|B \cap C) \cdot \Pr(B|C) \cdot \Pr(C)$$

Exercise. Prove that $\Pr(A|B \cap C) = \frac{\Pr(A \cap B|C)}{\Pr(B|C)}$

Tree Diagram

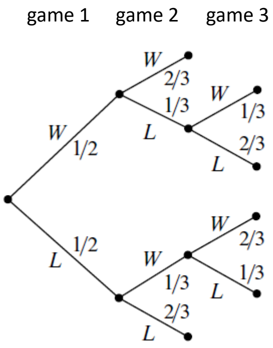


outcome	prob	event F	event G_1
WW	1/3	✓	✓
WLW	1/18	✓	✓
WLL	1/9		✓
LWW	1/9	✓	
LWL	1/18		
LL	1/3		
		$\Pr[F] = 1/2$	$\Pr[G_1] = 1/2$

$$\Pr[F|G_1] = \frac{\Pr[F \cap G_1]}{\Pr[G_1]} = \frac{7/18}{1/2} = \frac{7}{9}$$

A Posteriori Probability

- What is the difference between $P[G_2|G_1]$ and $P[G_1|G_2]$?
- Does $\Pr[G_1|G_2]$ even make sense?



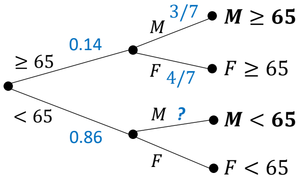
outcome	prob	event G_1	event G_2
WW	1/3	✓	✓
WLW	1/18	✓	
WLL	1/9	✓	
LWW	1/9		✓
LWL	1/18		✓
LL	1/3		
		$\Pr[G_1] = 1/2$	$\Pr[G_2] = 1/2$

$$\Pr[G_2|G_1] = \frac{\Pr[G_1 \cap G_2]}{\Pr[G_1]} = \frac{1/3}{1/2} = \frac{2}{3} \qquad \Pr[G_1|G_2] = \frac{\Pr[G_1 \cap G_2]}{\Pr[G_2]} = \frac{1/3}{1/2} = \frac{2}{3}$$

Exercise

According to Wikipedia, 14% of adults in the US are age 65 and over
Furthermore, the male/female ratio for this group is (3:4)

1. Compute the fraction of males aged 65 and over among the entire population
2. Compute the fraction of males aged less than 65 among the entire population
3. Express the quantities above using the notation of conditional probability



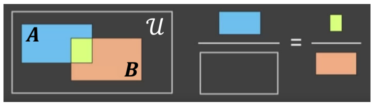
$$\Pr[M \cap \geq 65] = \Pr[\geq \mathbf{65}] \cdot \Pr[\boldsymbol{M} \mid \geq \mathbf{65}]$$
$$= 0.14 \cdot (3/7) = 6\%$$

$$\Pr[\boldsymbol{F} \cap \geq \mathbf{65}] = \Pr[\geq \mathbf{65}] \cdot \Pr[\boldsymbol{F} \mid \geq \mathbf{65}]$$

Independence

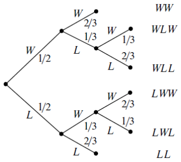
- Let A and B be events with $\Pr[B] \neq 0$. Then A is ***independent*** of B iff

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



Note. An event with probability 0 is *independent* of *every* other event.

Example. In our 2 out of 3 competition example, winning the final and winning the second game are not independent, why?



- *Symmetry*. If A is independent of B then B is independent of A. Why?

- Equivalently, A and B are independent iff

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

Mathematical vs. Everyday Independence

- Beware not to assign additional meaning to the relation *independent-of*

Example. To determine your proficiency in Java, you are given two "*independent tests by two independent committees*" ⇒ no mathematical independence

$$\Pr[\text{passed test 2} \mid \text{passed test 1}] > \Pr[\text{passed test 2}]$$

Exercise. In a single die roll we are interested in the events: roll an even number (*A*), roll a multiple of 3 (*B*), roll a prime number (*C*). Which of these pairs of events, if any, are independent?

Conditional Independence

- Events A and B are conditionally independent on event C if whenever C has happened, knowing whether B happened provides no information about the occurrence of A
- Formally, A and B are conditionally independent on C iff

$$\Pr(A, B|C) = P(A|C) \cdot P(B|C)$$

- Or equivalently, $\Pr(A|B, C) = \Pr(A|C)$

Example. Consider the events *Raining*, *Lightning*, and *Thundering.* We may reasonably claim that while *Raining* is not independent of *Thundering*, they are certainly independent given *Lightning*

Mutually Independent Events

- Informally, a set of events is *mutually independent* if the probability of any subset of them is not affected by which of the other events have occurred
- More formally, events $A_1, \ldots, A_m$ are ***mutually independent*** iff for *all* subsets $S \subseteq \{1, \ldots, m\}$

$$\Pr\left[\bigcap_{i \in S} A_i\right] = \prod_{i \in S} \Pr[A_i]$$

Exercise

- Consider the random experiment of flipping 3 fair coins and define the events
  - A_1: coin 1 matches coin 2
  - A_2: coin 2 matches coin 3
  - A_3: coin 3 matches coin 1
- Are A_1, A_2, A_3 pairwise independent?
- Are they mutually independent?

Bayes’ Theorem: Motivation

- Suppose that you get an email message from your bank, from a foreign address
- Is it reasonable to assume it is a scam? Why?
  - Many scam messages use foreign email addresses
  - Foreign email addresses are not very common
  - Scam emails are common nowadays

$$\Pr[H|E] = \frac{\Pr[H \cap E]}{\Pr[E]} = \frac{\Pr[E|H]}{\Pr[E]}\Pr[H]$$

where H = hypothesis (message is scam) and E = evidence (uses foreign address) are events defined on the same sample space

Bayes’ Theorem

$$\boxed{\mathbf{Pr}[\boldsymbol{H}|\boldsymbol{E}] = \frac{\Pr[H \cap E]}{\Pr[E]} = \left(\frac{\Pr[E|H]}{\Pr[E]}\right)\mathbf{Pr}[\boldsymbol{H}]}$$

- Relates $\mathbf{Pr}[\boldsymbol{H}|\boldsymbol{E}]$ (the *posterior*) to $\mathbf{Pr}[\boldsymbol{H}]$ (the *prior*)
  - How does the likelihood of H change in light of evidence E?
  - Evidence E multiplies the probability of H by a factor that measures how much condition H affects the probability of E
- *H*: message is scam, E: message uses a foreign address
- $\Pr[H|E]$ is high (compared to $\Pr[H]$) because
  - $\Pr[E|H]$ is very high
  - $\Pr[E]$ is low
  - $\Pr[H]$ is medium, but largely irrelevant in our reasoning
- Foreign address makes the scam hypothesis much more probable because it appears in scam messages more often than in general

Example

- Suppose that 1% of Americans suffer from disease *D*.
- A laboratory test for this disease is known to give the wrong result 10% of the time. How much should you worry if you test positive?

1. Model the problem using the language of conditional probability
2. Compute the probability that you have the disease
3. What is the probability that you are healthy if the test result is negative

Notation. D = have the disease, $\overline{D}$ = do not have the disease
T^+ = tested positive, T^- = tested negative

Example…

Notation: D: got the disease, T^+: tested positive

1. Model the problem: tested positive,
   $\Pr[D] = 1\%$, $\Pr[\text{false } +] = \Pr[\text{false } -] = 10\%$, want $\Pr[D|T^+]$
2. Compute the probability that you are ill (D) given T^+

$$\Pr[D|T^+] = \frac{\Pr[D \cap T^+]}{\Pr[T^+]} = \frac{\Pr[T^+|D]\Pr[D]}{\Pr[T^+ \cap D] + \Pr[T^+ \cap \overline{D}]}$$

$$= \frac{(0.01)(0.9)}{(0.01)(0.9) + (0.99)(0.1)} = 0.083$$

3. Compute the probability that you are healthy given T^-

Exercise

- Using Bayes Theorem, comment informally on the soundness of the following arguments:

1. Look, you can count up to 10; for sure you are a mathematician
2. Look, you speak a foreign language in a country where few people know a foreign language; you must be a spy

Random Variables

- A ***random variable*** X is a function $S \to \mathbb{R}$, i.e., a mapping from outcomes to real numbers

  Example. flip 3 fair coins. Define random variables $A = \#$ heads, $B = 1$ if all coins agree, 0 otherwise. Then,

A	
$HHH \longmapsto 3$	$THH \longmapsto 2$
$HHT \longmapsto 2$	$THT \longmapsto 1$
$HTH \longmapsto 2$	$TTH \longmapsto 1$
$HTT \longmapsto 1$	$TTT \longmapsto 0$

B	
$HHH \longmapsto 1$	$THH \longmapsto 0$
$HHT \longmapsto 0$	$THT \longmapsto 0$
$HTH \longmapsto 0$	$TTH \longmapsto 0$
$HTT \longmapsto 0$	$TTT \longmapsto 1$

- A random variable X is ***discrete*** if it takes countably many values $\{x_1, x_2, \dots\}$
  - A random variable over a discrete sample space is always discrete
  - A random variable over an uncountable set may or may not be discrete

Events on random variables

- We extend the notion of events to discrete random variables
- If X is a random variable, the event $\{X = x\}$ is defined as $\{X = x\} \coloneqq \{s \in S, X(s) = x\}$

Example. flip 3 fair coins. Define random variables $A = \#$ heads, $B = 1$ if all coins agree, 0 otherwise.

A	B
$HHH \longmapsto 3$	$HHH \longmapsto 1$
$HHT \longmapsto 2$	$HHT \longmapsto 0$
$HTH \longmapsto 2$	$HTH \longmapsto 0$
$HTT \longmapsto 1$	$HTT \longmapsto 0$
$THH \longmapsto 2$	$THH \longmapsto 0$
$THT \longmapsto 1$	$THT \longmapsto 0$
$TTH \longmapsto 1$	$TTH \longmapsto 0$
$TTT \longmapsto 0$	$TTT \longmapsto 1$

$\{A = 0\} \equiv \{TTT\}$

$\{A = 1\} \equiv \{HTT, THT, TTH\}$

$\{A = 3\} \equiv \{HHH\}$

$\{B = 1\} \equiv \{HHH, TTT\}$

Independence

- Two random variables X and Y are *independent* iff *for all* values x and y the events $\{X = x\}$ and $\{Y = y\}$ are independent
- If X and Y are independent random variables, then

$$\Pr[X = x, Y = y] = \Pr[X = x] \cdot \Pr[Y = y]$$

Example. Let X denote the sum of rolling two dice, one after the other, and let X_i be the value on die i. Since X_1 and X_2 are independent, then,

$$\Pr[X = 12] = \Pr[X_1 = 6, X_2 = 6] = \Pr[X_1 = 6] \cdot \Pr[X_2 = 6]$$

$$= (1/6)(1/6) = 1/36$$

Exercise

- In our experiment of flipping three coins, are A and B independent random variables?
  - $A = \#$ of heads, with sample space $\{0,1,2,3\}$
  - $B = 1$ if all coins match, else 0, with sample space $\{0,1\}$

Hint. Can you find $a, b \in \mathbb{R}$, such that

$$\Pr[A = a \ \wedge B = b] \neq \Pr[A = a] \cdot \Pr[B = b]?$$

Expected Value

- If X is a random variable then the ***expected value*** of X, denoted $E(X)$ is defined as

$$E(X) = \sum_{x} x \cdot \Pr[X = x]$$

Example. When rolling 2 dice, let $X = a + b$

$$E(X) = 2(1/36){+}3(2/36){+}4(3/36){+}\ldots{+}12(1/36){=}\ 7$$

- *Linearity.* $E[aX + y] = aE[X] + E[Y]$, constant a

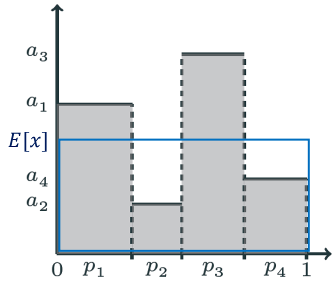
Example: X_1 = number on die 1, X_2 = number on die 2

$$E[X] = E[X_1 + X_2] = E[X_1] + E[X_2] = 2 \times (1 + 2 + \cdots + 6)/6$$

- *Independence.* If X and Y are independent random variables, then $E[XY] = E[X] \cdot E[Y]$

A Geometric Interpretation

- Let x be a random variable with sample space $\{a_1, a_2, a_3, a_4\}$ and probabilities $\Pr[x = a_i] = p_i$
- Then $E[x] = a_1p_1 + a_2p_2 + a_3p_3 + a_4p_4$

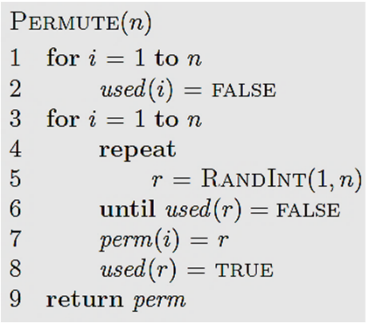


Example. Alice and Bob play a game with funny dice (2,2,2,2,3,3) and (1,1,1,1,6,6), respectively. Higher roll wins that value, paid by the losing player.

- What is the expected value of each dice?
- Which player is more likely to win? Which player would you rather be?

Exercise

- Here is an algorithm to generate a random permutation



```
PERMUTE($n$)
for $i = 1$ to $n$
    $used(i) = $ FALSE
for $i = 1$ to $n$
    repeat
        $r = $ RANDINT$(1, n)$
    until $used(r) = $ FALSE
    $perm(i) = r$
    $used(r) = $ TRUE
return $perm$
```

- Prove that the algorithm chooses a permutation uniformly at random
- What is the expected number of calls to RANDINT?

Indicator Variables

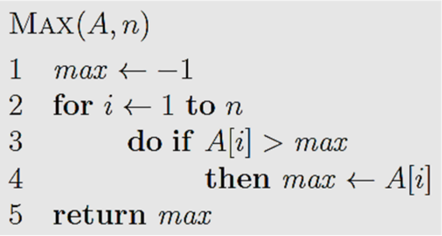
- An ***indicator variable*** is a random variable with sample space $\{0,1\}$
- *Notation*. For event A, define

$$\mathbb{1}_A = I_A = I(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not} \end{cases}$$

- What is the expected value of an indicator variable?
- What is the expected value of a sum of indicator variables?

Example

- Consider the algorithm below operating on a random permutation of an array $A[1{:}n]$ of positive numbers



```
MAX($A, n$)
1  $max \leftarrow -1$
2  for $i \leftarrow 1$ to $n$
3          do if $A[i] > max$
4                  then $max \leftarrow A[i]$
5  return $max$
```

- The number X of executions of line 4 is a random variable
- What is the sample space of X? What is $E[X]$?

Finding Max on Random Array

- Without indicator variables

  Let X = # executions of line 4

$$E[X] = \sum_{x=1}^{n} x \cdot \Pr[X = x]$$

- With indicator variables:

  Let $X_i = I\{\text{line 4 is executed in iteration } i\}$

  Then $X = X_1 + X_2 + \cdots + X_n$

  What is $E[X_i]$? What is $E[X]$?

  $E[X_i] = 1/i$ and $E[X] = H_n$ where $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$

  Claim. $\ln(n+1) < H_n < 1 + \ln n$ (will prove in next module)

Exercise

- Consider a set of $n = 28$ people chosen at random. Show that the expected number of pairs of people that have the same birthday is greater than 1.

Markov's Inequality

Theorem. If X is a non-negative random variable then $P[X \geq t] \leq E[X]/t$

*Proof***.** Define an indicator variable $Y = I\{X \geq t\}$. Then, $\Pr[X \geq t] = E[Y]$. Since $Y \leq X/t$ for all t, then $\Pr[X \geq t] = E[Y] \leq E[X/t] = E[X]/t$

- This inequality is often used in the context of chance "games" in order to find an upper bound for the probability of losing or for the probability that an algorithm takes longer than certain acceptable limit

*Example***.** Surveys show that DU students carry an average of $20 in cash. If you meet a student at random, estimate the chance that they are carrying less than $80.

Probability Mass Function

- The ***probability mass function*** $f: \mathbb{R} \to [0,1]$ of a discrete random variable X is simply its probability distribution, defined as

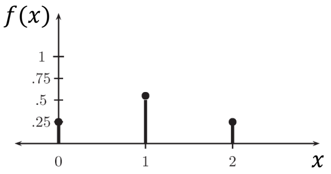
$$f(x) = \Pr[X = x]$$

- The ***cumulative distribution function*** $F: \mathbb{R} \to [0,1]$ of a discrete random variable X is defined as:

$$F(x) = \Pr(X \leq x) = \sum_{z \leq x} f(z)$$

Example. Let X be the number of heads after flipping a fair coin twice. Then

$$f(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \end{cases}$$



Some special random variables

- *Bernoulli.* $X \sim \text{Bernoulli}(p)$ with pmf $f(x) = p^x(1-p)^{1-x}, x = 0,1$

- *Binomial.* $X \sim \text{Binomial}(n, p)$ with pmf $f(x) = \binom{n}{x}p^x(1-p)^{n-x}, x = 0,1,\dots,n$

- *Geometric.* $X \sim \text{Geom}(p)$ with pmf $f(x) = p(1-p)^{x-1}, x = 1,2,3,\dots$

- *Poisson.* $X \sim \text{Poisson}(\lambda)$ with pmf $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, x = 0,1,2,\dots$

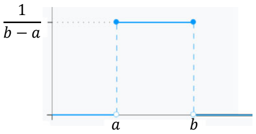
Side note: probability density function

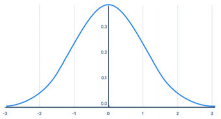
- Some random processes are best modelled using uncountable sample spaces, e.g., height/weight of a person, highest temperature tomorrow, etc.
- A random variable X is ***continuous*** if there exists a non-negative f (called the ***probability density function***) such that $\int_{-\infty}^{+\infty} f(x)dx = 1$ and for every $a \leq b$:

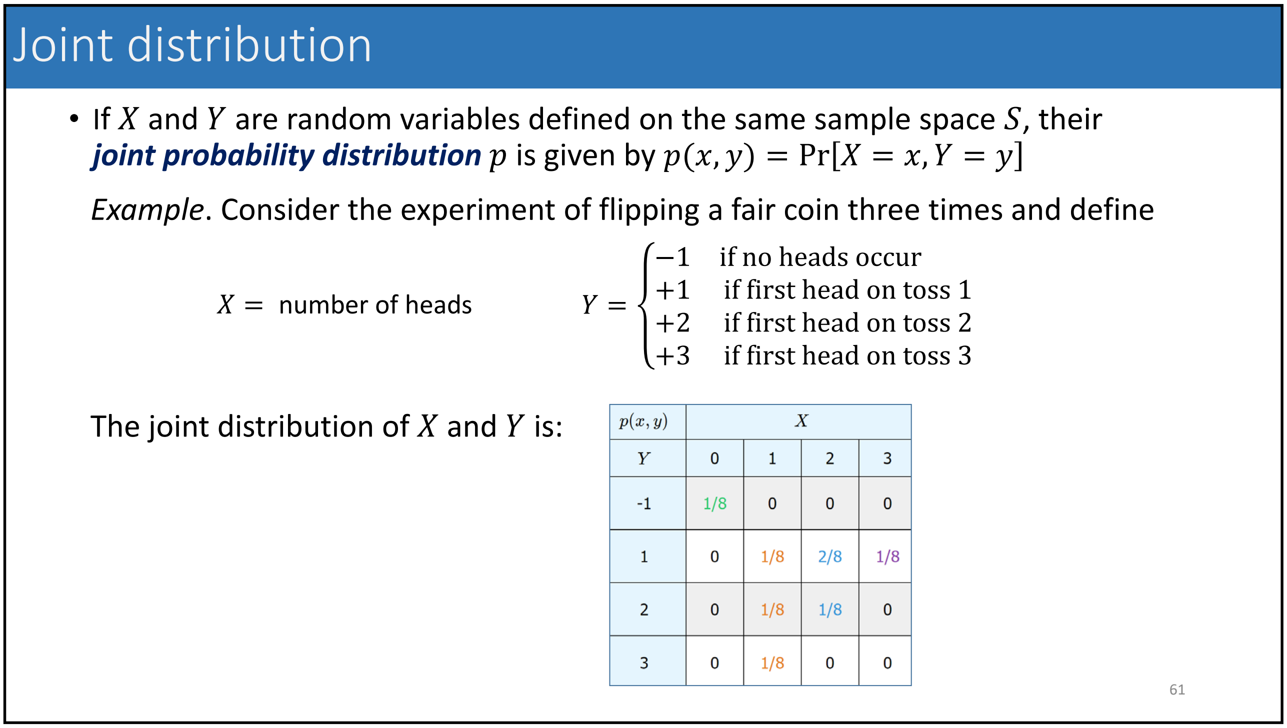
$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

Examples. (a) standard uniform and (b) standard normal distributions

(a) $f(x) = \begin{cases} 1 & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$ (b) $f(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}, x \in \mathbb{R}$







Joint distribution

- If X and Y are random variables defined on the same sample space S, their ***joint probability distribution*** p is given by $p(x, y) = \Pr[X = x, Y = y]$

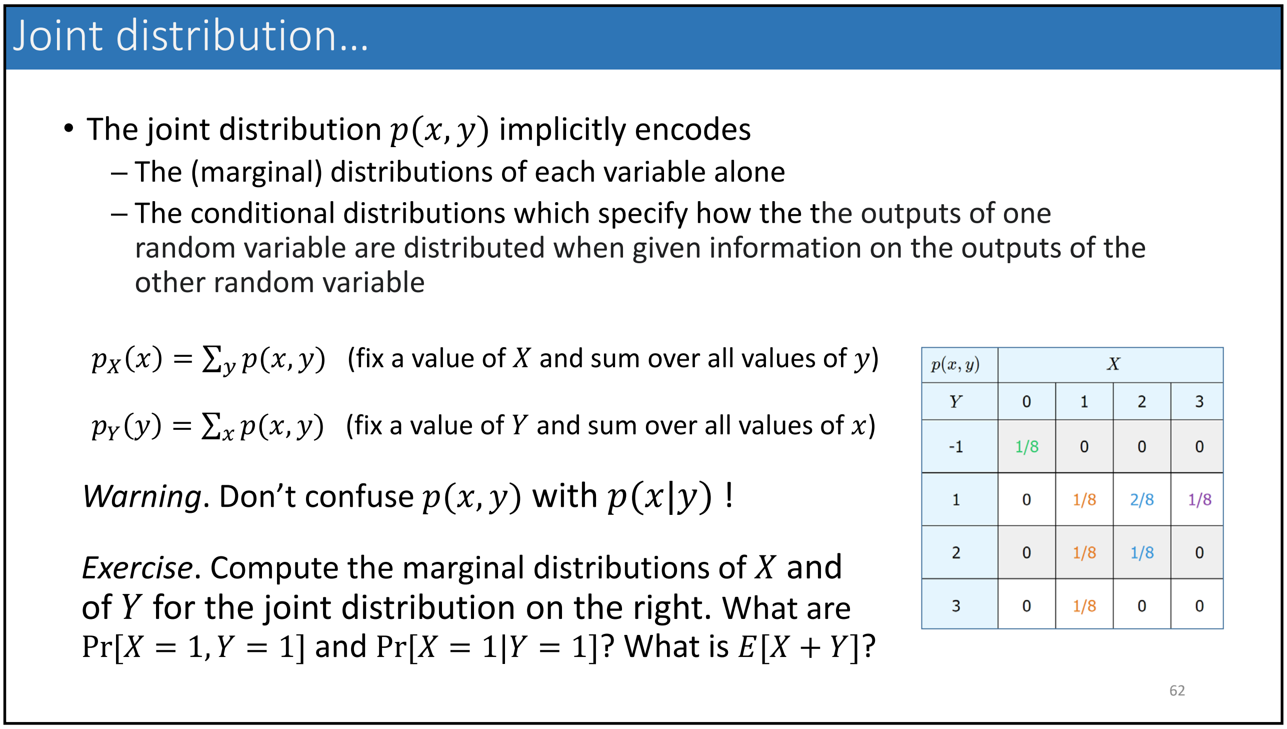
Example. Consider the experiment of flipping a fair coin three times and define

$X =$ number of heads

$$Y = \begin{cases} -1 & \text{if no heads occur} \\ +1 & \text{if first head on toss 1} \\ +2 & \text{if first head on toss 2} \\ +3 & \text{if first head on toss 3} \end{cases}$$

The joint distribution of X and Y is:

$p(x, y)$	X			
Y	0	1	2	3
-1	1/8	0	0	0
1	0	1/8	2/8	1/8
2	0	1/8	1/8	0
3	0	1/8	0	0



Joint distribution…

- The joint distribution $p(x, y)$ implicitly encodes
  - The (marginal) distributions of each variable alone
  - The conditional distributions which specify how the the outputs of one random variable are distributed when given information on the outputs of the other random variable

$p_X(x) = \sum_y p(x, y)$ (fix a value of X and sum over all values of y)

$p_Y(y) = \sum_x p(x, y)$ (fix a value of Y and sum over all values of x)

Warning. Don't confuse $p(x, y)$ with $p(x|y)$!

Exercise. Compute the marginal distributions of X and of Y for the joint distribution on the right. What are $\Pr[X = 1, Y = 1]$ and $\Pr[X = 1|Y = 1]$? What is $E[X + Y]$?

$p(x, y)$	X			
Y	0	1	2	3
-1	1/8	0	0	0
1	0	1/8	2/8	1/8
2	0	1/8	1/8	0
3	0	1/8	0	0

Expectation

- Let $f(x,y)$ be a function of random variables X and Y with joint distribution $p(x,y)$. Then, the *expected value* of $f(x,y)$ is given by

$$E[f(X,Y)] = \sum_x \sum_y f(x,y)\cdot p(x,y)$$

Example. In our example of flipping a coin 3 times

$$E[XY] = \sum_x \sum_y x\cdot y\cdot p(x,y) = (0)(-1)\left(\frac{1}{8}\right) + (1)(1)\left(\frac{1}{8}\right) + (1)(2)\left(\frac{1}{8}\right) + (1)(3)\left(\frac{1}{8}\right) + (2)(1)\left(\frac{2}{8}\right) + (2)(2)\left(\frac{1}{8}\right) + (3)(1)\left(\frac{1}{8}\right) = \frac{17}{8} = 2.125$$

$p(x,y)$	X			
Y	0	1	2	3
-1	1/8	0	0	0
1	0	1/8	2/8	1/8
2	0	1/8	1/8	0
3	0	1/8	0	0

Maximum likelihood

- *Maximum likelihood estimation* (MLE) is a method for finding the parameter(s) of a given probability mass/density function that best justify a set of observed data
- This is achieved by maximizing a *likelihood function* so that, under the assumed statistical distribution, the observed data is most likely

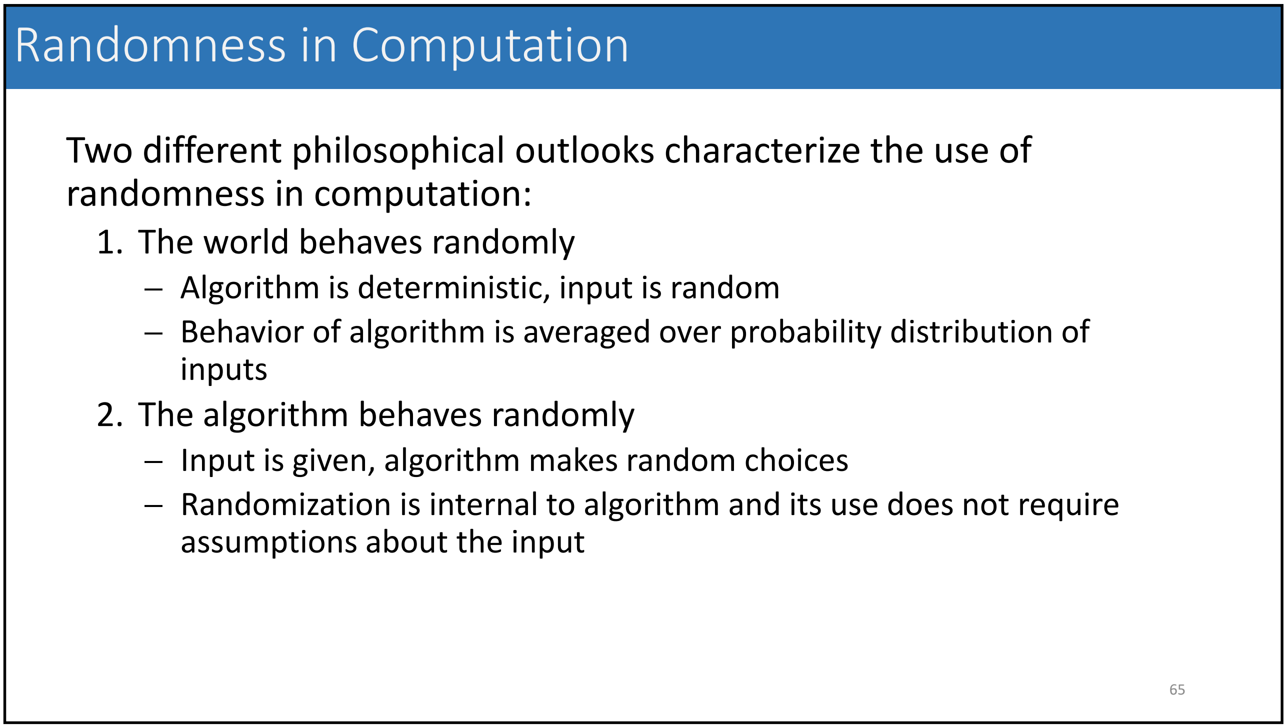
Example. The probability w that a random person carries a specific gene can be modeled with a Bernoulli variate, with mass function $p(x) = w^x(1-w)^{1-x}$, for $x \in \{0,1\}$. Given three iid samples 1,0,1 what is the most likely value of w?

The likelihood function is:

$$\mathcal{L}(w\ ) = w^{x_1}(1-w)^{1-x_1}\cdot w^{x_2}(1-w)^{1-x_2}\cdot w^{x_3}(1-w)^{1-x_3} = w(1-w)w$$

The likelihood $\mathcal{L}(w) = w^2 - w^3$ is maximized when $\mathcal{L}'(w) = 0$

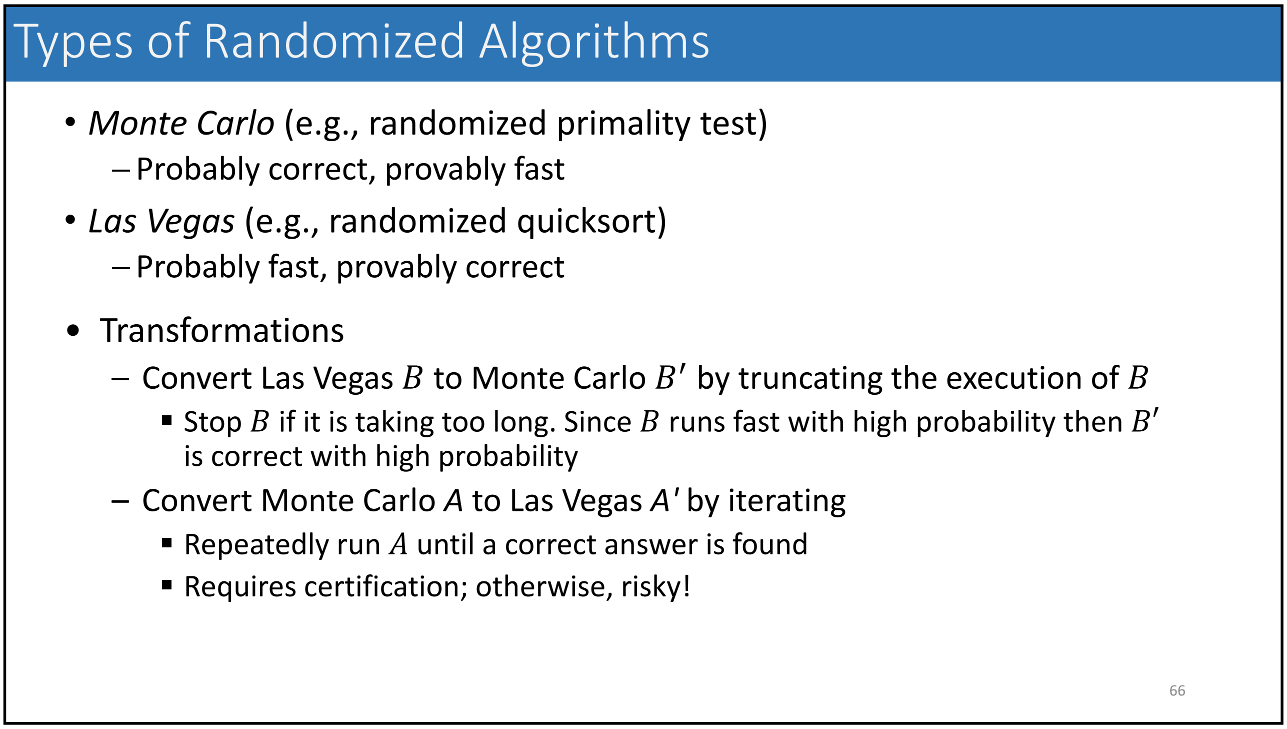
Since $\mathcal{L}'(w) = 2w - 3w^2 = 0$ when $2w = 3w^2 \Rightarrow$
$w = 2/3$ is the parameter value that best justifies the data



Randomness in Computation

Two different philosophical outlooks characterize the use of randomness in computation:

1. The world behaves randomly
   - Algorithm is deterministic, input is random
   - Behavior of algorithm is averaged over probability distribution of inputs
2. The algorithm behaves randomly
   - Input is given, algorithm makes random choices
   - Randomization is internal to algorithm and its use does not require assumptions about the input



Types of Randomized Algorithms

- *Monte Carlo* (e.g., randomized primality test)
  - Probably correct, provably fast
- *Las Vegas* (e.g., randomized quicksort)
  - Probably fast, provably correct

- Transformations
  - Convert Las Vegas B to Monte Carlo B' by truncating the execution of B
    - Stop B if it is taking too long. Since B runs fast with high probability then B' is correct with high probability
  - Convert Monte Carlo A to Las Vegas A' by iterating
    - Repeatedly run A until a correct answer is found
    - Requires certification; otherwise, risky!

Example

- A top school is evaluating n applicants for admission, using a test consisting of q questions (e.g., $q = 100$)
- Each candidate is either an *Ace* or a *Dud.* Duds answer less than 70% of the questions correctly, while each Ace must answer at least 70% of the questions correctly. Goal is to separate the Aces from the Duds
- A *deterministic* algorithm grades, for each student, the first $0.7q$ questions of the exam. Declare Ace if all correct; else a Dud
- A *Monte Carlo* algorithm, grades 10 random questions for each student. Label as Ace if at least 7 are correct; else a Dud
- A *Las Vegas* algorithm tests all questions in random order until more than 30% are wrongly answered (then Dud) or 70% of q questions answered correctly (then Ace)

From Las Vegas to Montecarlo

- Suppose you have a Las Vegas algorithm A whose expected running time $E[T]$ is $0.1n^2$ (sometimes it takes more time, sometimes it takes less time)
- Create a new algorithm B that simply runs A for at most $10n^2$ units of time and stops with a random answer (e.g., whatever it has computed so far) if computation is not finished
- How likely is it that B will return an incomplete answer?

$$\Pr[T > 10n^2] \leq \frac{E[T]}{10n^2} = 0.01$$