A Short Advertisement

Why you should all participate in class!

- You'll learn more
  - Active participation is likely to increase your recall
- You'll force me to explain things that you're not sure of

A Short Advertisement

Why you should all participate in class!

- You'll learn more
  - Active participation is likely to increase your recall
- You'll force me to explain things that you're not sure of
- The pragmatic reason: I'll get to know you and will be able to write letters of reference for you

A Short Advertisement

Why you should all participate in class!

- You'll learn more
  - Active participation is likely to increase your recall
- You'll force me to explain things that you're not sure of
- The pragmatic reason: I'll get to know you and will be able to write letters of reference for you

Why you should read the course notes after every class:

- You'll learn the material better
- You can come to the next class asking me to explain what you're confused about!

A Short Advertisement

Why you should all participate in class!

- You'll learn more
  - Active participation is likely to increase your recall
- You'll force me to explain things that you're not sure of
- The pragmatic reason: I'll get to know you and will be able to write letters of reference for you

Why you should read the course notes after every class:

- You'll learn the material better
- You can come to the next class asking me to explain what you're confused about!

Why you should read the homework solutions:

- Even if you got them right, the solutions may point out subtleties or a better way of doing the problem.
- The homework is a good guide to the prelim/final.

Probability

Life is full of uncertainty.

Probability is the best way we currently have to quantify it.

Probability

Life is full of uncertainty.

Probability is the best way we currently have to quantify it.

Applications of probability arise everywhere:

- Should you guess in a multiple-choice test with five choices?
  - What if you're not penalized for guessing?
  - What if you're penalized $1/4$ for every wrong answer?
  - What if you can eliminate two of the five possibilities?

Probability

Life is full of uncertainty.

Probability is the best way we currently have to quantify it.

Applications of probability arise everywhere:

- Should you guess in a multiple-choice test with five choices?
  - What if you're not penalized for guessing?
  - What if you're penalized 1/4 for every wrong answer?
  - What if you can eliminate two of the five possibilities?
- Suppose that an AIDS test guarantees 99% accuracy:
  - of every 100 people who have AIDS, the test returns positive 99 times (very few *false negative*);
  - of every 100 people who don't have AIDS, the test returns negative 99 times (very few *false positives*)

  Suppose you test positive. How likely are you to have AIDS?
  - Hint: the probability is *not* .99

Probability

Life is full of uncertainty.

Probability is the best way we currently have to quantify it.

Applications of probability arise everywhere:

- Should you guess in a multiple-choice test with five choices?
  - What if you're not penalized for guessing?
  - What if you're penalized 1/4 for every wrong answer?
  - What if you can eliminate two of the five possibilities?
- Suppose that an AIDS test guarantees 99% accuracy:
  - of every 100 people who have AIDS, the test returns positive 99 times (very few *false negative*);
  - of every 100 people who don't have AIDS, the test returns negative 99 times (very few *false positives*)

  Suppose you test positive. How likely are you to have AIDS?
  - Hint: the probability is *not* .99
- How do you compute the average-case running time of an algorithm?
- Is it worth buying a $1 lottery ticket?
  - Probability isn't enough to answer this question

Probability

Life is full of uncertainty.

Probability is the best way we currently have to quantify it.

Applications of probability arise everywhere:

- Should you guess in a multiple-choice test with five choices?
  - What if you're not penalized for guessing?
  - What if you're penalized 1/4 for every wrong answer?
  - What if you can eliminate two of the five possibilities?
- Suppose that an AIDS test guarantees 99% accuracy:
  - of every 100 people who have AIDS, the test returns positive 99 times (very few *false negative*);
  - of every 100 people who don't have AIDS, the test returns negative 99 times (very few *false positives*)

  Suppose you test positive. How likely are you to have AIDS?
  - Hint: the probability is *not* .99
- How do you compute the average-case running time of an algorithm?
- Is it worth buying a $1 lottery ticket?
  - Probability isn't enough to answer this question

(I think) everybody ought to know something about probability.

Interpreting Probability

Probability can be a subtle.

The first (philosophical) question is "What does probability mean?"

- What does it mean to say that "The probability that the coin landed (will land) heads is 1/2"?

Interpreting Probability

Probability can be a subtle.

The first (philosophical) question is "What does probability mean?"

- What does it mean to say that "The probability that the coin landed (will land) heads is 1/2"?

Two standard interpretations:

- Probability is *subjective*: This is a subjective statement describing an individual's feeling about the coin landing heads
  - This feeling can be quantified in terms of betting behavior
- Probability is an *objective* statement about frequency

Both interpretations lead to the same mathematical notion.

The text discusses these issues in Section 18.4.6.

- You should read it!

Formalizing Probability

What do we assign probability to?

Intuitively, we assign them to possible *events* (things that might happen, *outcomes* of an experiment)

Formalizing Probability

What do we assign probability to?
Intuitively, we assign them to possible *events* (things that might happen, *outcomes* of an experiment)

Formally, we take a *sample space* to be a *set*.

- Intuitively, the sample space is the set of possible outcomes, or possible ways the world could be.

An *event* is a subset of a sample space.

We assign probability to events: that is, to subsets of a sample space.

Formalizing Probability

What do we assign probability to?
Intuitively, we assign them to possible *events* (things that might happen, *outcomes* of an experiment)

Formally, we take a *sample space* to be a *set*.

- Intuitively, the sample space is the set of possible outcomes, or possible ways the world could be.

An *event* is a subset of a sample space.

We assign probability to events: that is, to subsets of a sample space.

Sometimes the hardest thing to do in a problem is to decide what the sample space should be.

- There's often more than one choice
- A good thing to do is to try to choose the sample space so that all outcomes (i.e., elements) are equally likely
  - This is not always possible or reasonable

Choosing the Sample Space

Example 1: We toss a coin. What's the sample space?

- Most obvious choice: $\{\text{heads, tails}\}$
- Should we bother to model the possibility that the coin lands on edge?
- What about the possibility that somebody snatches the coin before it lands?
- What if the coin is biased?

Choosing the Sample Space

Example 1: We toss a coin. What's the sample space?

- Most obvious choice: $\{\text{heads, tails}\}$
- Should we bother to model the possibility that the coin lands on edge?
- What about the possibility that somebody snatches the coin before it lands?
- What if the coin is biased?

Example 2: We toss a die. What's the sample space?

Example 3: Two distinguishable dice are tossed together. What's the sample space?

Choosing the Sample Space

Example 1: We toss a coin. What's the sample space?

- Most obvious choice: $\{\text{heads, tails}\}$
- Should we bother to model the possibility that the coin lands on edge?
- What about the possibility that somebody snatches the coin before it lands?
- What if the coin is biased?

Example 2: We toss a die. What's the sample space?

Example 3: Two distinguishable dice are tossed together. What's the sample space?

- (1,1), (1,2), (1,3), . . . , (6,1), (6,2), . . . , (6,6)

What if the dice are indistinguishable?

Choosing the Sample Space

Example 1: We toss a coin. What's the sample space?

- Most obvious choice: {heads, tails}
- Should we bother to model the possibility that the coin lands on edge?
- What about the possibility that somebody snatches the coin before it lands?
- What if the coin is biased?

Example 2: We toss a die. What's the sample space?

Example 3: Two distinguishable dice are tossed together. What's the sample space?

- (1,1), (1,2), (1,3), . . . , (6,1), (6,2), . . . , (6,6)

What if the dice are indistinguishable?

Example 4: You're a doctor examining a seriously ill patient, trying to determine the probability that he has cancer. What's the sample space?

Example 5: You're an insurance company trying to insure a nuclear power plant. What's the sample space?

The text gives a systematic way of generating a sample space that's very useful in many cases; we'll come back to that.

Probability Measures

A *probability measure* assigns a real number between 0 and 1 to every subset of (event in) a sample space.

- Intuitively, the number measures how likely that event is.
- Probability 1 says it's certain to happen; probability 0 says it's certain not to happen
- Probability acts like a *weight* or *measure*. The probability of separate things (i.e., disjoint sets) adds up.

Formally, a probability measure Pr on S is a function mapping subsets of S to real numbers such that:

1. For all $A \subseteq S$, we have $0 \leq \Pr(A) \leq 1$
2. $\Pr(\emptyset) = 0$; $\Pr(S) = 1$
3. If A and B are disjoint subsets of S (i.e., $A \cap B = \emptyset$), then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

It follows by induction that if $A_1, \ldots, A_k$ are pairwise disjoint, then

$$\Pr(\cup_{i=1}^{k} A_i) = \sum_{i}^{k} \Pr(A_i).$$

- This is called *finite additivity*; it's actually more standard to assume a countable version of this, called *countable additivity*

In particular, this means that if $A = \{e_1, \ldots, e_k\}$, then

$$\Pr(A) = \sum_{i=1}^{k} \Pr(e_i).$$

In finite spaces, the probability of a set is determined by the probability of its elements.

The text defines a probability measure on S to be a function $\Pr : S \to \boldsymbol{R}$ such that

(a) $\Pr(s) \geq 0$ for all $s \in S$

(b) $\sum_{s \in S} \Pr(s) = 1$.

- Notice that in the text's definition, the domain of Pr is S, not 2^S. They then *define* $\Pr(A) = \sum_{s \in A} \Pr(s)$ for $A \subseteq S$.
- The text's definition is equivalent to the one on the previous slide if S is finite.
  - The definition on the previous slide generalizes better to infinite domains (e.g., to probability measures on $[0, 1]$).

What does it mean when the weatherman says there a 25% chance of rain tomorrow:

- It will rain tomorrow for 25% of the day

What does it mean when the weatherman says there a 25% chance of rain tomorrow:

- It will rain tomorrow for 25% of the day
- It will rain in 25% of the Ithaca region

What does it mean when the weatherman says there a 25% chance of rain tomorrow:

- It will rain tomorrow for 25% of the day
- It will rain in 25% of the Ithaca region
- On 25% of the days on which a prediction like this is made, it rains the next day?
- What count as rain?
  - A light sprinkle
  - A heavy downpour

What does it mean when the weatherman says there a 25% chance of rain tomorrow:

- It will rain tomorrow for 25% of the day
- It will rain in 25% of the Ithaca region
- On 25% of the days on which a prediction like this is made, it rains the next day?
- What count as rain?
  - A light sprinkle
  - A heavy downpour

If I'm trying to figure out whether to take an umbrella tomorrow, I want to know!

What does it mean when the weatherman says there a 25% chance of rain tomorrow:

- It will rain tomorrow for 25% of the day
- It will rain in 25% of the Ithaca region
- On 25% of the days on which a prediction like this is made, it rains the next day?
- What count as rain?
  - A light sprinkle
  - A heavy downpour

If I'm trying to figure out whether to take an umbrella tomorrow, I want to know!

How can we use different sample spaces to make this precise?

Equiprobable Measures

Suppose S has n elements, and we want Pr to make each element equally likely.

- Then each element gets probability $1/n$
- $\Pr(A) = |A|/n$

In this case, Pr is called an *equiprobable* or *uniform* measure.

- Not all probability measures are uniform!

Equiprobable Measures

Suppose S has n elements, and we want Pr to make each element equally likely.

- Then each element gets probability $1/n$
- $\Pr(A) = |A|/n$

In this case, Pr is called an *equiprobable* or *uniform* measure.

- Not all probability measures are uniform!

Example 1: In the coin example, if you think the coin is fair, and the only outcomes are heads and tails, then we can take
$S = \{\text{heads,tails}\}$, and
$\Pr(\text{heads}) = \Pr(\text{tails}) = 1/2$.

Equiprobable Measures

Suppose S has n elements, and we want Pr to make each element equally likely.

- Then each element gets probability $1/n$
- $\Pr(A) = |A|/n$

In this case, Pr is called an *equiprobable* or *uniform* measure.

- Not all probability measures are uniform!

Example 1: In the coin example, if you think the coin is fair, and the only outcomes are heads and tails, then we can take
$S = \{\text{heads,tails}\}$, and
$\Pr(\text{heads}) = \Pr(\text{tails}) = 1/2$.

Example 2: In the two-dice example where the dice are distinguishable, if you think both dice are fair, then we can take
$\Pr((i, j)) = 1/36$.

- Should it make a difference if the dice are indistinguishable?

Equiprobable measures on infinite sets

Defining an equiprobable measure on an infinite set can be tricky.

Theorem: There is no equiprobable measure on the positive integers.

Proof: By contradiction. Suppose Pr is an equiprobable measure on the positive integers, and $\Pr(1) = \epsilon > 0$.

There must be some N such that $\epsilon > 1/N$.
Since $\Pr(1) = \cdots = \Pr(N) = \epsilon$, we have

$$\Pr(\{1, \ldots, N\}) = N\epsilon > 1 \text{ — a contradiction}$$

But if $\Pr(1) = 0$, then $\Pr(S) = \Pr(1) + \Pr(2) + \cdots = 0$.

Some basic results

How are the probability of E and $\overline{E}$ related?

- How does the probability that the dice lands either 2 or 4 (i.e., $E = \{2, 4\}$) compare to the probability that the dice lands 1, 3, 5, or 6 ($\overline{E} = \{1, 3, 5, 6\}$)

Theorem 1: $\Pr(\overline{E}) = 1 - \Pr(E)$.

Proof: E and $\overline{E}$ are disjoint, so that

$$\Pr(E \cup \overline{E}) = \Pr(E) + \Pr(\overline{E}).$$

But $E \cup \overline{E} = S$, so $\Pr(E \cup \overline{E}) = 1$.
Thus $\Pr(E) + \Pr(\overline{E}) = 1$, so

$$\Pr(\overline{E}) = 1 - \Pr(E).$$

Theorem 2: $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Theorem 2: $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

$$\begin{aligned}
A &= (A - B) \cup (A \cap B) \\
B &= (B - A) \cup (A \cap B) \\
A \cup B &= (A - B) \cup (B - A) \cup (A \cap B)
\end{aligned}$$

So

$$\begin{aligned}
\Pr(A) &= \Pr(A - B) + \Pr(A \cap B) \\
\Pr(B) &= \Pr(B - A) + \Pr(A \cap B) \\
\Pr(A \cup B) &= \Pr(A - B) + \Pr(B - A) + \Pr(A \cap B)
\end{aligned}$$

The result now follows.

Theorem 2: $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

$$\begin{aligned} A &= (A - B) \cup (A \cap B) \\ B &= (B - A) \cup (A \cap B) \\ A \cup B &= (A - B) \cup (B - A) \cup (A \cap B) \end{aligned}$$

So

$$\begin{aligned} \Pr(A) &= \Pr(A - B) + \Pr(A \cap B) \\ \Pr(B) &= \Pr(B - A) + \Pr(A \cap B) \\ \Pr(A \cup B) &= \Pr(A - B) + \Pr(B - A) + \Pr(A \cap B) \end{aligned}$$

The result now follows.

Remember the Inclusion-Exclusion Rule?

$$|A \cup B| = |A| + |B| - |A \cap B|$$

This follows easily from Theorem 2, if we take Pr to be an equiprobable measure. We can also generalize to arbitrary unions.

For homework, you'll prove other standard properties of probability, like monotonicity:

- If $A \subseteq B$, then $\Pr(A) \leq \Pr(B)$.

Disclaimer

- Probability is a well defined mathematical theory.
- Applications of probability theory to "real world" problems is not.
- Choosing the sample space, the events and the probability function requires a "leap of faith".
- We cannot prove that we chose the right model but we can argue for that.
- Some examples are easy some are not:
  - Flipping a coin or rolling a die.
  - Playing a lottery game.
  - Guessing in a multiple choice test.
  - Determining whether or not the patient has AIDS based on a test.
  - Does the patient have cancer?

Conditional Probability

One of the most important features of probability is that there is a natural way to *update* it.

Example: Bob draws a card from a 52-card deck. Initially, Alice considers all cards equally likely, so her probability that the ace of spades was drawn is $1/52$. Her probability that the card drawn was a spade is $1/4$.

Conditional Probability

One of the most important features of probability is that there is a natural way to *update* it.

Example: Bob draws a card from a 52-card deck. Initially, Alice considers all cards equally likely, so her probability that the ace of spades was drawn is $1/52$. Her probability that the card drawn was a spade is $1/4$.

Then she sees that the card is black. What should her probability now be that

- the card is the ace of spades?
- the card is a spade?

A reasonable approach:

- Start with the original sample space
- Eliminate all outcomes (elements) that you now consider impossible, based on the observation (i.e., assign them probability 0).
- Keep the relative probability of everything else the same.
  - Renormalize to get the probabilities to sum to 1.

What should the probability of B be, given that you've observed A? According to this recipe, it's

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

$\Pr(\mathrm{A}\spadesuit \mid \mathrm{black}) = (1/52)/(1/2) = 1/26$
$\Pr(\mathrm{spade} \mid \mathrm{black}) = (1/4)/(1/2) = 1/2.$

What should the probability of B be, given that you've observed A? According to this recipe, it's

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

$\Pr(\mathrm{A}\spadesuit \mid \mathrm{black}) = (1/52)/(1/2) = 1/26$
$\Pr(\mathrm{spade} \mid \mathrm{black}) = (1/4)/(1/2) = 1/2.$

A subtlety:

- What if Alice doesn't completely trust Bob? How do you take this into account?

What should the probability of B be, given that you've observed A? According to this recipe, it's

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

$\Pr(\mathrm{A}\spadesuit \mid \mathrm{black}) = (1/52)/(1/2) = 1/26$
$\Pr(\mathrm{spade} \mid \mathrm{black}) = (1/4)/(1/2) = 1/2.$

A subtlety:

- What if Alice doesn't completely trust Bob? How do you take this into account?

  Two approaches:

  (1) Enlarge sample space to allow more observations.
  (2) Jeffrey's rule:

$$\Pr(\mathrm{A}\spadesuit \mid \mathrm{black}) \cdot \Pr(\text{Bob telling the truth}) + \Pr(\mathrm{A}\spadesuit \mid \mathrm{red}) \cdot \Pr(\text{Bob lying}).$$

What do you condition on?

In general, figuring out what to condition on can be subtle.

- See Steve Strogatz's article wonderful article "Chances Are" in the NYTimes (also discussed in MCS): `opinionator.blogs.nytimes.com/2010/04/25/chances-are/?_r=0`

Example from the O.J. Simpson trial:

- The prosecution argued that OJ had a pattern of violent behavior towards his wife
  - E.g., he would slap her, throw her against walls
- The defense argued that all this was irrelevant
  - Fewer than 1 out 2500 men who slap/beat their wives go on to murder them.

Who was right?

What do you condition on?

In general, figuring out what to condition on can be subtle.

- See Steve Strogatz's article wonderful article "Chances Are" in the NYTimes (also discussed in MCS): `opinionator.blogs.nytimes.com/2010/04/25/chances-are/?_r=0`

Example from the O.J. Simpson trial:

- The prosecution argued that OJ had a pattern of violent behavior towards his wife
  - E.g., he would slap her, throw her against walls
- The defense argued that all this was irrelevant
  - Fewer than 1 out 2500 men who slap/beat their wives go on to murder them.

Should we be interested in

(a) Pr(someone murdered his wife | he previously battered her).

(b) Pr(someone murdered his wife | he previously battered her *and she was murdered*).

(c) neither.

The second-ace puzzle

Alice gets two cards from a deck with four cards:
A♠, 2♠, A♡, 2♡.

A♠ A♡	A♠ 2♠	A♠ 2♡
A♡ 2♠	A♡ 2♡	2♠ 2♡

The probability that Alice has both aces is 1/6.

The second-ace puzzle

Alice gets two cards from a deck with four cards:
A♠, 2♠, A♡, 2♡.

A♠ A♡	A♠ 2♠	A♠ 2♡
A♡ 2♠	A♡ 2♡	2♠ 2♡

The probability that Alice has both aces is 1/6.

Alice then tells Bob "I have an ace".

- What's the probability that Alice has both aces? 1/5

The second-ace puzzle

Alice gets two cards from a deck with four cards:
A♠, 2♠, A♡, 2♡.

A♠ A♡	A♠ 2♠	A♠ 2♡
A♡ 2♠	A♡ 2♡	2♠ 2♡

The probability that Alice has both aces is 1/6.

Alice then tells Bob "I have an ace".

- What's the probability that Alice has both aces? 1/5

She then says "I have the ace of spades".

- Now what's the probability that Alice has both aces?

The second-ace puzzle

Alice gets two cards from a deck with four cards:
A♠, 2♠, A♡, 2♡.

A♠ A♡	A♠ 2♠	A♠ 2♡
A♡ 2♠	A♡ 2♡	2♠ 2♡

The probability that Alice has both aces is 1/6.

Alice then tells Bob "I have an ace".

- What's the probability that Alice has both aces? 1/5

She then says "I have the ace of spades".

- Now what's the probability that Alice has both aces? 1/3

What if Alice had said "I have the ace of hearts" instead?

The second-ace puzzle

Alice gets two cards from a deck with four cards:
A♠, 2♠, A♡, 2♡.

A♠ A♡	A♠ 2♠	A♠ 2♡
A♡ 2♠	A♡ 2♡	2♠ 2♡

The probability that Alice has both aces is 1/6.

Alice then tells Bob "I have an ace".

- What's the probability that Alice has both aces? 1/5

She then says "I have the ace of spades".

- Now what's the probability that Alice has both aces? 1/3

What if Alice had said "I have the ace of hearts" instead?

- Also 1/3

But then why did Bob need Alice?

- Bob knows she has an ace. Whichever ace she has, the probability that she has both aces is 1/3.
- So he knows it's 1/3 even without Alice saying anything??!!

Is the probability that Alice has both aces $1/3$?

(a) Yes

(b) No

(c) I have no idea

The Monty Hall Puzzle

- You're on a game show and given a choice of three doors.
  - Behind one is a car; behind the others are goats.
- You pick door 1.
- Monty Hall opens door 2, which has a goat.
- He then asks you if you still want to take what's behind door 1, or to take what's behind door 3 instead.

Should you switch?

The Monty Hall Puzzle: Two Arguments

Here's the argument for not switching:

- The car is equally likely to be behind each door. After you learn it's not behind door 2, you condition on that fact. Now it's still equally likely to be behind door 1 and door 3. There's no point in switching.

The Monty Hall Puzzle: Two Arguments

Here's the argument for not switching:

- The car is equally likely to be behind each door. After you learn it's not behind door 2, you condition on that fact. Now it's still equally likely to be behind door 1 and door 3. There's no point in switching.

Here's the argument for switching:

- With probability 1/3 you picked the door with a car; with probability 2/3 you picked a door with a goat.
  - If you picked the door with a car, you lose by switching: you definitely get a goat.
  - If you picked a door with with a goat, you win by switching; the car is behind door 3 (the goats are behind door 1 and 2).

  So it seems that switching gains with probability 2/3.

The Monty Hall Puzzle: Two Arguments

Here's the argument for not switching:

- The car is equally likely to be behind each door. After you learn it's not behind door 2, you condition on that fact. Now it's still equally likely to be behind door 1 and door 3. There's no point in switching.

Here's the argument for switching:

- With probability 1/3 you picked the door with a car; with probability 2/3 you picked a door with a goat.
  - If you picked the door with a car, you lose by switching: you definitely get a goat.
  - If you picked a door with with a goat, you win by switching; the car is behind door 3 (the goats are behind door 1 and 2).

  So it seems that switching gains with probability 2/3.

Which argument is right?

(a) 2/3
(b) 1/2
(c) both
(d) neither

The Monty Hall Puzzle: Two Arguments

Here's the argument for not switching:

- The car is equally likely to be behind each door. After you learn it's not behind door 2, you condition on that fact. Now it's still equally likely to be behind door 1 and door 3. There's no point in switching.

Here's the argument for switching:

- With probability 1/3 you picked the door with a car; with probability 2/3 you picked a door with a goat.
  - If you picked the door with a car, you lose by switching: you definitely get a goat.
  - If you picked a door with with a goat, you win by switching; the car is behind door 3 (the goats are behind door 1 and 2).

  So it seems that switching gains with probability 2/3.

Which argument is right?

- If you think it's 2/3, what's wrong with conditioning?
  - Do we condition only in some cases and not in others?
  - If so, when?

The Protocol Matters

Conditioning is always the right thing to do, but you have to use the right sample space to get the result.

- The right sample space includes the protocol!

For the second-ace puzzle, suppose Alice's protocol says that at the first step, she'll tell Bob whether she has an ace. At the second step, she'll tell Bob which ace she has.

- But what does she do if she both aces? Which ace does she tell Bob about?
  - Protocol #1: she says "ace of hearts" whenever she has a choice.
    - In that case, the probability that she has both aces if she says "ace of spades" is 0, not 1/3!
    - the probability that she has both aces if she says "ace of hearts" is 1/3.

- Possibility #2: she randomizes when she has a choice (says "Ace of hearts" with probability 1/2 and "ace of spades" with probability 1/2).
  - Now the sample space has to include how Alice's coin that determines what she says in this case landed.
    - There are 7 elements in the sample space, not 6!

  An easy calculation (done in class) shows that the probability that she has both aces if she says "ace of spades" is 1/5, not 1/3.

Back to Monty Hall

Again, what Monty does is determined if the there's a goat behind door 1

- He opens the other door that has a goat behind it
- *Assuming that he necessarily opens a door—see below.*

But which door does Monty open if door 1 has a car?

- if he definitely opens door 2, then switching doesn't help.

Back to Monty Hall

Again, what Monty does is determined if the there's a goat behind door 1

- He opens the other door that has a goat behind it
- *Assuming that he necessarily opens a door—see below.*

But which door does Monty open if door 1 has a car?

- if he definitely opens door 2, then switching doesn't help.
- if he randomizes between door 2 and door 3, then you gain by switching. Here's the calculation:
  - The probability space has four elements: $(C1, D2)$ (the car is behind door 1 and he opens door 2), $(C1, D3)$, $(C2, D3)$, and $(C3, D2)$.
    - The first two each have probability $1/6$; the last two each have probability $1/3$.
    - An easy calculation shows that $\Pr(C1 \mid D2) = 1/3$ and $\Pr(C3 \mid D2) = 2/3$, so you gain by switching

Back to Monty Hall

Again, what Monty does is determined if the there's a goat behind door 1

- He opens the other door that has a goat behind it
- *Assuming that he necessarily opens a door—see below.*

But which door does Monty open if door 1 has a car?

- if he definitely opens door 2, then switching doesn't help.
- if he randomizes between door 2 and door 3, then you gain by switching. Here's the calculation:
  - The probability space has four elements: $(C1, D2)$ (the car is behind door 1 and he opens door 2), $(C1, D3)$, $(C2, D3)$, and $(C3, D2)$.
    - The first two each have probability $1/6$; the last two each have probability $1/3$.
    - An easy calculation shows that $\Pr(C1 \mid D2) = 1/3$ and $\Pr(C3 \mid D2) = 2/3$, so you gain by switching

But what if Monty's protocol is to open door 2 only if door 1 has the car behind it?

- Then switching is a terrible idea!

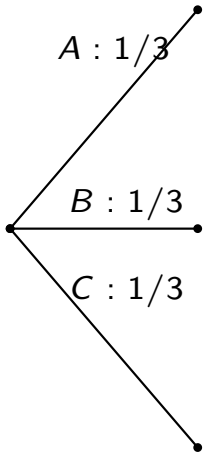
Using Protocols to Generate Tree Diagrams

We can use protocols to generate a *tree diagram* or *probability tree* that determines the sample space.

- Each non-leaf node in the tree corresponds to an uncertain choice.
- The edges leading from that node correspond to ways of resolving the uncertainty.
- The edges are labeled by the probability of making that choice.

Consider Monty Hall.

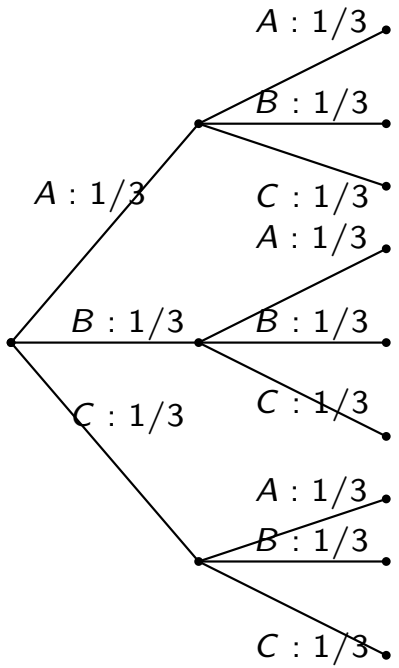
- The first choice is where the car is.
  - There are three possibilities: door A, B, or C.
  - By assumption, these are all equally likely.



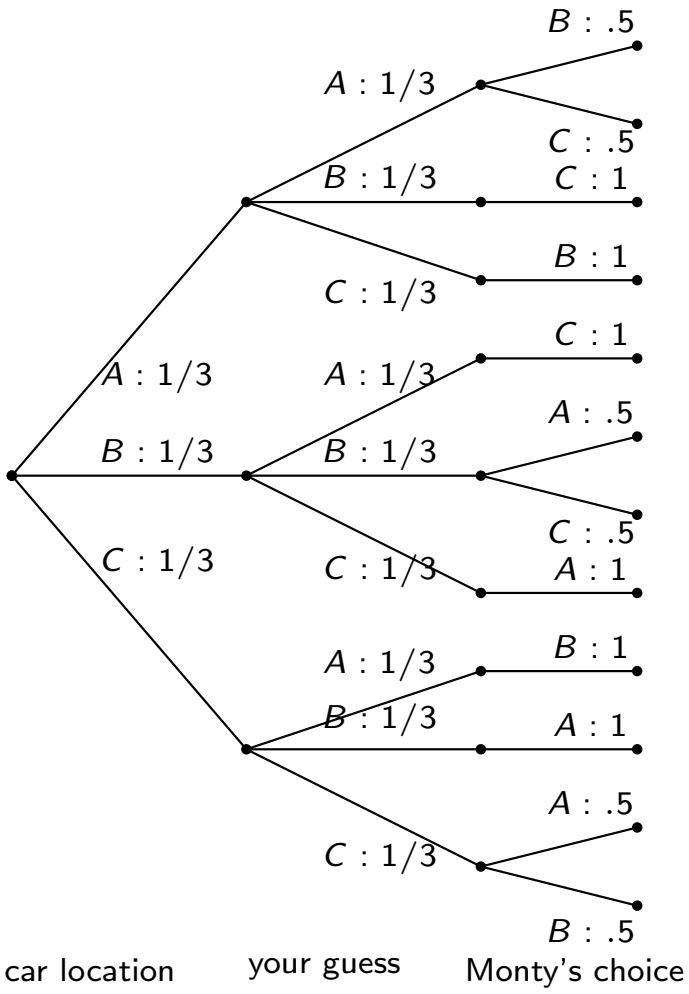
Consider Monty Hall.

- The first choice is where the car is.
  - There are three possibilities: door A, B, or C.
  - By assumption, these are all equally likely.

- Next you point to a door, again with uniform probability



Finally, Monty opens a door (choosing at random when he has a choice):



- Each path in the tree defines an outcome (an element of the sample space)
- The probability of the outcome is the product of the probabilities on the edges of the path

Probability Trees

Probability trees are useful for describing sequential decision, randomized algorithms, . . .

One more example:

Suppose that the probability of rain tomorrow is .7. If it rains, then the probability that the game will be cancelled is .8; if it doesn't rain, then the probability that it will be cancelled is .1. What is the probability that the game will be played?

The situation can be described by a tree:

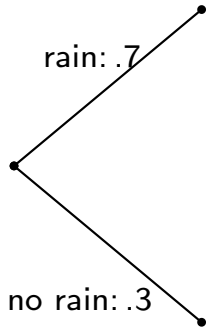
Probability Trees

Probability trees are useful for describing sequential decision, randomized algorithms, . . .

One more example:

Suppose that the probability of rain tomorrow is .7. If it rains, then the probability that the game will be cancelled is .8; if it doesn't rain, then the probability that it will be cancelled is .1. What is the probability that the game will be played?

The situation can be described by a tree:



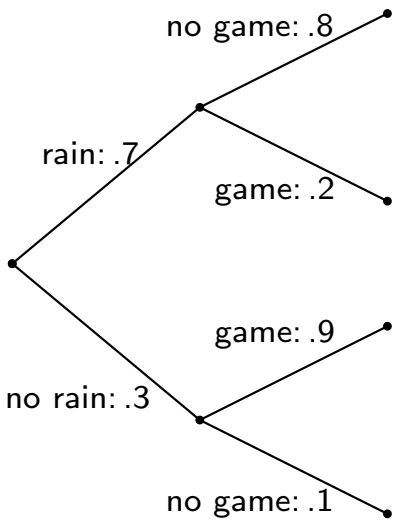
Probability Trees

Probability trees are useful for describing sequential decision, randomized algorithms, . . .

One more example:

Suppose that the probability of rain tomorrow is .7. If it rains, then the probability that the game will be cancelled is .8; if it doesn't rain, then the probability that it will be cancelled is .1. What is the probability that the game will be played?

The situation can be described by a tree:



Why Does This Work?

Probability trees provide a great systematic way to generate a probability space.

But why do they work?

- Why is the probability of a path the product of the probability of the edges?
- Going back to the Monty Hall tree, the path $C; C; B$ is the outcome where the car is behind door C, you chose door C, and Monty opened door B.
  - But what does it mean that the bottom-most edge is labeled "$B : .5$"?
    - Exactly what probability is it that's .5?
    - Is it the probability of Monty opening door B?

Why Does This Work?

Probability trees provide a great systematic way to generate a probability space.

But why do they work?

- Why is the probability of a path the product of the probability of the edges?
- Going back to the Monty Hall tree, the path $C; C; B$ is the outcome where the car is behind door C, you chose door C, and Monty opened door B.
  - But what does it mean that the bottom-most edge is labeled "$B : .5$"?
    - Exactly what probability is it that's .5?
    - Is it the probability of Monty opening door B?

This is a conditional probability!

- The probability that Monty opens door B given that the car is behind door C and you pointed to door C

Observe that $\Pr(A_1 \cap A_2) = \Pr(A_1 \mid A_2) \times \Pr(A_2)$.
Taking $X; Y; Z$ to denote the path where the car is behind door X, you point to door Y, and Monty opens door Z, we have

$$\begin{aligned}
\Pr(A; C; B) &= \Pr(A; C) \times \Pr(B \mid A; C) \\
&= \Pr(A; C) \times 1 \\
&= \Pr(A \mid C) \times \Pr(C) \times 1/2 \\
&= 1/3 \times 1/3 \times 1
\end{aligned}$$

That's why the probability of a path is the product of the edge probabilities.

Observe that $\Pr(A_1 \cap A_2) = \Pr(A_1 \mid A_2) \times \Pr(A_2)$.
Taking $X; Y; Z$ to denote the path where the car is behind door X, you point to door Y, and Monty opens door Z, we have

$$\begin{aligned}
\Pr(A; C; B) &= \Pr(A; C) \times \Pr(B \mid A; C) \\
&= \Pr(A; C) \times 1 \\
&= \Pr(A \mid C) \times \Pr(C) \times 1/2 \\
&= 1/3 \times 1/3 \times 1
\end{aligned}$$

That's why the probability of a path is the product of the edge probabilities.

More generally,
$\Pr(A_1 \cap A_2 \cap A_3) = \Pr(A_1) \Pr(A_2|A_1) \Pr(A_3|A_1 \cap A_2)$.

There's an obvious generalization to $\Pr(A_1 \cap \ldots \cap A_n)$.

The book has lots of great examples of the use of probability trees.

- It's part of what they call the *four-step method*
  - Identify the sample space: the outcomes that you care about
  - Identify the events of interest (and how they arise over time)
    - E.g., the car and goats are placed behind a door, then you point to a door, then Monty opens a door: this determines an outcome
  - Then compute the probability of events
    - The events + the probabilities determine the probability tree

I probably won't cover the example in Section 17.3 ("Strange Dice"), but you should read it!

Independence

Intuitively, events A and B are independent if they have no effect on each other.

This means that observing A should have no effect on the likelihood we ascribe to B, and similarly, observing B should have no effect on the likelihood we ascribe to A.

Thus, if $\Pr(A) \neq 0$ and $\Pr(B) \neq 0$ and A is independent of B, we would expect

$$\Pr(B|A) = \Pr(B) \text{ and } \Pr(A|B) = \Pr(A).$$

Interestingly, one implies the other.

Independence

Intuitively, events A and B are independent if they have no effect on each other.

This means that observing A should have no effect on the likelihood we ascribe to B, and similarly, observing B should have no effect on the likelihood we ascribe to A.

Thus, if $\Pr(A) \neq 0$ and $\Pr(B) \neq 0$ and A is independent of B, we would expect

$$\Pr(B|A) = \Pr(B) \text{ and } \Pr(A|B) = \Pr(A).$$

Interestingly, one implies the other.

$\Pr(B|A) = \Pr(B)$ iff $\Pr(A \cap B)/\Pr(A) = \Pr(B)$ iff

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B).$$

Formally, we say A and B are *(probabilistically) independent* if

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B).$$

This definition makes sense even if $\Pr(A) = 0$ or $\Pr(B) = 0$.

Mutual vs. Pairwise Independence

What should it mean for 4 events $A_1, \ldots, A_4$ to be independent?

- *pairwise independence*: A_i is independent of A_j for $i \neq j$
- *mutual independence*:
  - A_i is independent of A_j for $i \neq j$ (pairwise independence)
  - *3-way independence*:
    - $\Pr(A_1 \cap A_2 \cap A_3) = \Pr(A_1)\Pr(A_2)\Pr(A_3)$
    - $\Pr(A_1 \cap A_2 \cap A_4) = \Pr(A_1)\Pr(A_2)\Pr(A_4)$
    - $\Pr(A_1 \cap A_3 \cap A_4) = \Pr(A_1)\Pr(A_3)\Pr(A_4)$
    - $\Pr(A_2 \cap A_3 \cap A_4) = \Pr(A_2)\Pr(A_3)\Pr(A_4)$
  - *4-way independence*:
    - $\Pr(A_1 \cap A_2 \cap A_3 \cap A_4) = \Pr(A_1)\Pr(A_2)\Pr(A_3)\Pr(A_4)$

Mutual independence obviously requires much more than just pairwise independence.

Example: Suppose A and B are 0 or 1 with probability 1/2, and $C = A \oplus B$.

- Then we have pairwise independence but not mutual independence
  - E.g., knowing A tells you nothing about either B or C
- However, they're not mutually independent
  - Knowing A and B determines C!
- (This fact is used in cryptographic protocols.)

This issue also arises in legal cases ...

Example: DNA Testing

In a jury trial, you hear things like "We did DNA testing and found a match. The probability of such a match is 1 in 170 million."

- Where did those numbers come from?

Genes have *markers*. Suppose we have statistics like

- 1 person in 100 has marker A
- 1 person in 50 has marker B
- 1 person in 40 has marker C
- 1 person in 5 has marker D
- 1 person in 170 has marker E

The witness has all five markers, and so does the blood sample at the crime scene. What's the probability of this happening?

If we assume that the markers are *mutually* independent, then the probability of a match is

$$\frac{1}{100} \times \frac{1}{50} \times \frac{1}{40} \times \frac{1}{5} \times \frac{1}{170} \approx \frac{1}{170,000,000}.$$

But is mutual independence a reasonable assumption?

Example: DNA Testing

Genes have *markers*. Suppose we have statistics like

- 1 person in 100 has marker A
- 1 person in 50 has marker B
- 1 person in 40 has marker C
- 1 person in 5 has marker D
- 1 person in 170 has marker E

If we assume that the markers are *mutually* independent, then the probability of a match is $\frac{1}{100} \times \frac{1}{50} \times \frac{1}{40} \times \frac{1}{5} \times \frac{1}{170} \approx \frac{1}{170{,}000{,}000}$.

What if they're only pairwise independent?

- E.g., if you have markers A and B, you're likely to C too

What can you conclude in that case?

(a) Nothing
(b) The probability of a match is still 1 in 170,000,000
(c) The probability of a match could be as high as 1 in 17,000
(d) The probability of a match could be as high as 1 in 200
(e) No idea

Example: DNA Testing

Genes have *markers*. Suppose we have statistics like

- 1 person in 100 has marker A
- 1 person in 50 has marker B
- 1 person in 40 has marker C
- 1 person in 5 has marker D
- 1 person in 170 has marker E

If we assume that the markers are *mutually* independent, then the probability of a match is $\frac{1}{100} \times \frac{1}{50} \times \frac{1}{40} \times \frac{1}{5} \times \frac{1}{170} \approx \frac{1}{170{,}000{,}000}$.

What if they're only pairwise independent?

- E.g., if you have markers A and B, you're likely to C too

What if they weren't independent at all?

Independence, pairwise independence, and mutual independence are typically assumptions made based on our understanding of the science, and not from the data. We need to think about how reasonable they are in practice ...

This issue arose in a real-life court case:

> *On 24 March 2003, Lucia De Berk, a Dutch nurse, was sentenced by the court in The Hague to life imprisonment for the murder of four patients and the attempted murder of three others. The verdict depended in part on a statistical calculation, according to which the probability was allegedly only 1 in 342 million that a nurse's shifts would coincide with so many of the deaths and resuscitations purely by chance.*

This issue arose in a real-life court case:

> *On 24 March 2003, Lucia De Berk, a Dutch nurse, was sentenced by the court in The Hague to life imprisonment for the murder of four patients and the attempted murder of three others. The verdict depended in part on a statistical calculation, according to which the probability was allegedly only 1 in 342 million that a nurse's shifts would coincide with so many of the deaths and resuscitations purely by chance.*

- Statisticians said the the probabilistic reasoning that led to the conviction was seriously flawed
- Case was reheard, and de Berk declared not guilty in 2010

Bayes' Theorem

Suppose that you have a barrel-full of coins. One coin in the barrel is double-headed; all the rest are fair. You draw a coin from the barrel at random, and toss it 10 times. It lands heads each time. What's the probability that it's double headed?

(a) $1/2^{10}$
(b) I have no idea

Bayes' Theorem

Suppose that you have a barrel-full of coins. One coin in the barrel is double-headed; all the rest are fair. You draw a coin from the barrel at random, and toss it 10 times. It lands heads each time. What's the probability that it's double headed?

(a) $1/2^{10}$

(b) I have no idea

The right answer is (b). What else do you need to know to figure out the true probability?

Bayes' Theorem

Suppose that you have a barrel-full of coins. One coin in the barrel is double-headed; all the rest are fair. You draw a coin from the barrel at random, and toss it 10 times. It lands heads each time. What's the probability that it's double headed?

(a) $1/2^{10}$
(b) I have no idea

The right answer is (b). What else do you need to know to figure out the true probability?

Suppose we have a test that is 99% effective against AIDS.

- The probability of a *false positive*—the test is positive although you don't have AIDS—is 1%.
- The probability of a *false negative*—the test is negative although you have AIDS—is 1%.

Suppose that you test positive. What's the probability that you have AIDS?

(a) .99
(b) it depends . . .

The Law of Total Probability

The first step in addressing this formally is the *law of total probability*:

$$\Pr(A) = \Pr(A \mid E)\Pr(E) + \Pr(A \mid \overline{E})\Pr(\overline{E}).$$

Why is this true?

The Law of Total Probability

The first step in addressing this formally is the *law of total probability*:

$$\Pr(A) = \Pr(A \mid E)\Pr(E) + \Pr(A \mid \overline{E})\Pr(\overline{E}).$$

Why is this true?

$$\begin{aligned}\Pr(A) &= \Pr(A \cap E) + \Pr(A \cap \overline{E}) \\ &= \Pr(A \mid E)\Pr(E)\Pr(A \mid \overline{E}) + \Pr(\overline{E}).\end{aligned}$$

Example: You first toss a fair coin. If it comes up heads, you roll a fair die and win if it comes up 1 or 2. If it comes up tails, you roll the die and win if it comes up 3. What's the probability of winning?

You could easily draw the probability tree and do the calculations. But let's use the law of total probability:

$$\begin{aligned}\Pr(win) &= \Pr(win \mid heads)\Pr(heads) + \Pr(win \mid tails)\Pr(tails) \\ &= 1/3 \times 1/2 + 1/6 \times 1/2.\end{aligned}$$

Generalized law of total probability

If $E_1, \ldots, E_n$ are pairwise disjoint ($E_i \cap E_j = \emptyset$ for all i, j) and $E_1 \cup \ldots \cup E_n = S$, then

$$Pr(A) = \Pr(A \mid E_1)\Pr(E_1) + \cdots + \Pr(A \mid E_n)\Pr(E_n).$$

Bayes' Theorem

Bayes Theorem: Let $A_1, \ldots, A_n$ be mutually exclusive and exhaustive events in a sample space S.

- That means $A_1 \cup \ldots \cup A_n = S$, and the A_i's are pairwise disjoint: $A_i \cap A_j = \emptyset$ if $i \neq j$.

Let B be any other event in S. Then

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

Bayes' Theorem

Bayes Theorem: Let $A_1, \ldots, A_n$ be mutually exclusive and exhaustive events in a sample space S.

- That means $A_1 \cup \ldots \cup A_n = S$, and the A_i's are pairwise disjoint: $A_i \cap A_j = \emptyset$ if $i \neq j$.

Let B be any other event in S. Then

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

Proof: $\Pr(A_i \mid B) = \frac{\Pr(A_i \cap B)}{\Pr(B)}$.

We have seen that $\Pr(A_i \cap B) = \Pr(A_i \mid B)\Pr(B)$.

By the (generalized) law of total probability:

$$\Pr(B) = \sum_{j=1}^{n} \Pr(B \mid A_j)\Pr(A_j).$$

Example

In a certain county, 60% of registered voters are Republicans, 30% are Democrats, and 10% are Independents. 40% of Republicans oppose increased military spending, while 65% of the Democrats and 55% of the Independents oppose it. A registered voter writes a letter to the county paper, arguing against increased military spending. What is the probability that this voter is a Democrat?

$S = \{\text{registered voters}\}$
$A_1 = \{\text{registered Republicans}\}$
$A_2 = \{\text{registered Democrats}\}$
$A_3 = \{\text{registered independents}\}$
$B = \{\text{voters who oppose increased military spending}\}$

We want to know $\Pr(A_2|B)$.

We have

$$\begin{array}{lll} \Pr(A_1) = .6 & \Pr(A_2) = .3 & \Pr(A_3) = .1 \\ \Pr(B|A_1) = .4 & \Pr(B|A_2) = .65 & \Pr(B|A_3) = .55 \end{array}$$

Using Bayes' Theorem, we have:

$$\begin{aligned}
\Pr(A_2|B) &= \frac{\Pr(B|A_2)\times\Pr(A_2)}{\Pr(B|A_1)\times\Pr(A_1)+\Pr(B|A_2)\times\Pr(A_2)+\Pr(B|A_3)\times\Pr(A_3)} \\
&= \frac{.65\times.3}{(.4\times.6)+(.65\times.3)+(.55\times.1)} \\
&= \frac{.195}{.49} \\
&\approx .398
\end{aligned}$$

AIDS

Suppose we have a test that is 99% effective against AIDS. Suppose we also know that .3% of the population has AIDS. What is the probability that you have AIDS if you test positive?

$S = \{\text{all people}\}$ (in North America??)
$A_1 = \{\text{people with AIDS}\}$
$A_2 = \{\text{people who don't have AIDS}\}$ ($A_2 = \overline{A_1}$)
$B = \{\text{people who test positive}\}$

$$\Pr(A_1) = .003 \qquad \Pr(A_2) = .997$$

Since the test is 99% effective:

$$\Pr(B|A_1) = .99 \qquad \Pr(B|A_2) = .01$$

Using Bayes' Theorem again:

$$\begin{aligned}\Pr(A_1|B) &= \frac{.99\times.003}{(.99\times.003)+(.01\times.997)} \\ &\approx \frac{.003}{.003+.01} \\ &\approx .23\end{aligned}$$

Averaging and Expectation

Suppose you toss a coin that's biased towards heads ($\Pr(\text{heads}) = 2/3$) twice. How many heads do you expect to get?

- In mathematics-speak:
  What's the *expected number* of heads?

What about if you toss the coin k times?

What's the average weight of the people in this classroom?

- That's easy: add the weights and divide by the number of people in the class.

But what about if I tell you I'm going to toss a coin to determine which person in the class I'm going to choose; if it lands heads, I'll choose someone at random from the first aisle, and otherwise I'll choose someone at random from the last aisle.

- What's the expected weight?

Averaging makes sense if you use a uniform distribution; in general, we need to talk about *expectation*.

Random Variables

To deal with expectation, we formally associate with every element of a sample space a real number.

Definition: A *random variable* on sample space S is a function from S to some codomain,usually the real numbers.

- It's not random and it's not a variable!

Example: Suppose we toss a biased coin ($\Pr(h) = 2/3$) twice. The sample space is:

- hh - Probability $4/9$
- ht - Probability $2/9$
- th - Probability $2/9$
- tt - Probability $1/9$

If we're interested in the number of heads, we would consider a random variable $\#H$ that counts the number of heads in each sequence:

$$\#H(hh) = 2; \quad \#H(ht) = \#H(th) = 1; \quad \#H(tt) = 0$$

Example: If we're interested in weights of people in the class, the sample space is people in the class, and we could have a random variable that associates with each person his or her weight.

Important Example: An *indicator* or *binary random variable* maps every element of the sample space to either 0 or 1.

- Given a subset $A \subseteq S$, the indicator random variable I_A maps $s \in A$ to 1 and $s \notin A$ to 0

Indicator random variables turn out to be quite useful. (More examples coming.)

Random Variables and Events

Given a real-valued random variable X whose domain is a sample space S, and real number c, $X = c$ is an *event*: a subset of S.

- Which event is it?

Random Variables and Events

Given a real-valued random variable X whose domain is a sample space S, and real number c, $X = c$ is an *event*: a subset of S.

- Which event is it? $X = c$ is the event $\{s \in S : X(s) = c\}$

  Similarly, $X \leq c$ is the event $\{s \in S : X(s) \leq c\}$.

Probability Distributions

If X is a real-valued random variable on sample space S, then the probability that X takes on the value c is

$$\Pr(X = c) = \Pr(\{s \in S \mid X(s) = c\})$$

Similarly,

$$\Pr(X \leq c) = \Pr(\{s \in S \mid X(s) \leq c\}).$$

- $\Pr(X \leq c)$ makes sense since $X \leq c$ is an event (a subset of S)
  - We can talk about the probability only of events
- $\{s \in S \mid X(s) \leq c\}$ makes sense because X is a function whose range is the real numbers.

Probability Distributions

If X is a real-valued random variable on sample space S, then the probability that X takes on the value c is

$$\Pr(X = c) = \Pr(\{s \in S \mid X(s) = c\})$$

Similarly,

$$\Pr(X \leq c) = \Pr(\{s \in S \mid X(s) \leq c\}).$$

- $\Pr(X \leq c)$ makes sense since $X \leq c$ is an event (a subset of S)
  - We can talk about the probability only of events
- $\{s \in S \mid X(s) \leq c\}$ makes sense because X is a function whose range is the real numbers.

Example: In the coin example,

$$\Pr(\#H = 2) = 4/9 \text{ and } \Pr(\#H \leq 1) = 5/9$$

Given a probability measure Pr on a sample space S and a random variable X, the *probability distribution* associated with X is $PDF_X(x) = \Pr(X = x)$.

- PDF_X is a probability measure on the real numbers.

The *cumulative distribution* associated with X is $CDF_X(x) = \Pr(X \leq x)$.

An Example With Dice

Suppose S is the sample space corresponding to tossing a pair of fair dice: $\{(i,j) \mid 1 \leq i,j \leq 6\}$.

Let X be the random variable that gives the sum:

- $X(i,j) = i + j$

$PDF_X(2) = \Pr(X = 2) = \Pr(\{(1,1)\}) = 1/36$
$PDF_X(3) = \Pr(X = 3) = \Pr(\{(1,2),(2,1)\}) = 2/36$
$\vdots$
$PDF_X(7) = \Pr(X = 7) = \Pr(\{(1,6),(2,5),\ldots,(6,1)\}) = 6/36$
$\vdots$
$PDF_X(12) = \Pr(X = 12) = \Pr(\{(6,6)\}) = 1/36$

An Example With Dice

Suppose S is the sample space corresponding to tossing a pair of fair dice: $\{(i,j) \mid 1 \leq i,j \leq 6\}$.

Let X be the random variable that gives the sum:

- $X(i,j) = i + j$

$PDF_X(2) = \Pr(X = 2) = \Pr(\{(1,1)\}) = 1/36$
$PDF_X(3) = \Pr(X = 3) = \Pr(\{(1,2),(2,1)\}) = 2/36$
$\vdots$
$PDF_X(7) = \Pr(X = 7) = \Pr(\{(1,6),(2,5),\ldots,(6,1)\}) = 6/36$
$\vdots$
$PDF_X(12) = \Pr(X = 12) = \Pr(\{(6,6)\}) = 1/36$

Can similarly compute the cumulative distribution:

$$CDF_X(2) = PDF_X(2) = 1/36$$
$$CDF_X(3) = PDF_X(2) + PDF_X(3) = 3/36$$
$$\vdots$$
$$CDF_X(12) = 1$$

The Finite Uniform Distribution

Some PDFs/CDFs come up so often that they get a special name:

If X is a random variable that takes on each value in its range with equal probability, then PDF_X is the *uniform distibution* on the range of X:

If the range of X is $\{x_1, \ldots, x_n\}$, where $x_1 < x_2 < \ldots < x_n$, then:

$$PDF_X(x_k) = 1/n$$

$$CDF_X(x_k) = k/n$$

The Binomial Distribution

Suppose there is an experiment with probability p of success and thus probability $q = 1 - p$ of failure.

- For example, consider tossing a biased coin, where $\Pr(h) = p$. Getting "heads" is success, and getting tails is failure.

Suppose the experiment is repeated independently n times.

- For example, the coin is tossed n times.

This is called a sequence of *Bernoulli trials*.

Key features:

- Only two possibilities: success or failure.
- Probability of success does not change from trial to trial.
- The trials are independent.

What is the probability of k successes in n trials?

- Let $\#H_{n,p}$ count the number of heads (successes) in n trials, where the probability of heads is p
- we are interested in $\Pr(\#H_{n,p} = k)$.

Suppose $n = 5$ and $k = 3$. How many sequences of 5 coin tosses have exactly three heads?

- $hhhtt$, $hhtht$, $hhtth$, ...

What is the probability of k successes in n trials?

- Let $\#H_{n,p}$ count the number of heads (successes) in n trials, where the probability of heads is p
- we are interested in $\Pr(\#H_{n,p} = k)$.

Suppose $n = 5$ and $k = 3$. How many sequences of 5 coin tosses have exactly three heads?

- $hhhtt$, $hhtht$, $hhtth$, ...
- $C(5, 3)$

What is the probability of each one?

What is the probability of k successes in n trials?

- Let $\#H_{n,p}$ count the number of heads (successes) in n trials, where the probability of heads is p
- we are interested in $\Pr(\#H_{n,p} = k)$.

Suppose $n = 5$ and $k = 3$. How many sequences of 5 coin tosses have exactly three heads?

- *hhhtt*, *hhtht*, *hhtth*, ...
- $C(5, 3)$

What is the probability of each one?

$$p^3(1-p)^2$$

Therefore, $\Pr(\#H_{5,p} = 3) = C(5, 3)p^3(1-p)^2$.

Let $B_{n,p}(k)$ be the probability of getting k successes in n Bernoulli trials with probability p of success.

$$B_{n,p}(k) = \Pr(\#H_{n,p} = k) = C(n, k)p^k(1-p)^{n-k}$$

Not surprisingly, $B_{n,p}$ is called the *Binomial Distribution*.

- $B_{n,p}$ is the PDF of the random variable $\#H_{n,p}$

New Distributions from Old

If X and Y are random variables on a sample space S, then so are $X + Y$, $X + 2Y$, XY, $\sin(X)$, etc.

For example,

- $(X + Y)(s) = X(s) + Y(s)$.
- $\sin(X)(s) = \sin(X(s))$

Note $\sin(X)$ is a random variable: a function from the sample space to the reals.

Example 1: Two dice are rolled. Let X be the number that comes up on the first dice, and Y the number that comes up on the second.

- Formally, $X((i,j)) = i$, $Y((i,j)) = j$.

The random variable $X + Y$ is the total number showing.

Example 2: Suppose we toss a biased coin n times (more generally, we perform n Bernoulli trials). Let X_k describe the outcome of the kth coin toss: $X_k = 1$ if the kth coin toss is heads, and 0 otherwise.

- X_k is an indicator random variable.

How do we formalize this?

- What's the sample space?

Example 2: Suppose we toss a biased coin n times (more generally, we perform n Bernoulli trials). Let X_k describe the outcome of the kth coin toss: $X_k = 1$ if the kth coin toss is heads, and 0 otherwise.

- X_k is an indicator random variable.

How do we formalize this?

- What's the sample space?

Notice that $\sum_{k=1}^{n} X_k$ describes the number of successes of n Bernoulli trials.

- If the probability of a single success is p, then $\sum_{k=1}^{n} X_k$ has distribution $B_{n,p}$
  - The binomial distribution is the sum of Bernoullis

Independent random variables

In a roll of two dice, let X and Y record the numbers on the first and second die respectively.

- What can you say about the events $X = 3$, $Y = 2$?
- What about $X = i$ and $Y = j$?

Definition: The random variables X and Y are independent if for every x and y the events $X = x$ and $Y = y$ are independent.

Example: X and Y above are independent.

Independent random variables

In a roll of two dice, let X and Y record the numbers on the first and second die respectively.

- What can you say about the events $X = 3$, $Y = 2$?
- What about $X = i$ and $Y = j$?

Definition: The random variables X and Y are independent if for every x and y the events $X = x$ and $Y = y$ are independent.

Example: X and Y above are independent.

Definition: The random variables $X_1, X_2, \ldots, X_n$ are *mutually independent* if, for every $x_1, x_2 \ldots, x_n$

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) = \Pr(X_1 = x_1) \ldots \Pr(X_n = x_n)$$

Example: X_k, the success indicators in n Bernoulli trials, are independent.

Expected Value

Suppose we toss a biased coin, with $\Pr(h) = 2/3$. If the coin lands heads, you get \$1; if the coin lands tails, you get \$3. What are your expected winnings?

- $2/3$ of the time you get \$1;
  $1/3$ of the time you get \$3
- $(2/3 \times 1) + (1/3 \times 3) = 5/3$

What's a good way to think about this? We have a random variable W (for winnings):

- $W(h) = 1$
- $W(t) = 3$

The expectation of W is

$$\begin{aligned} E(W) &= \Pr(h)W(h) + \Pr(t)W(t) \\ &= \Pr(W = 1) \times 1 + \Pr(W = 3) \times 3 \end{aligned}$$

Expected Value

Suppose we toss a biased coin, with $\Pr(h) = 2/3$. If the coin lands heads, you get \$1; if the coin lands tails, you get \$3. What are your expected winnings?

- $2/3$ of the time you get \$1;
  $1/3$ of the time you get \$3
- $(2/3 \times 1) + (1/3 \times 3) = 5/3$

What's a good way to think about this? We have a random variable W (for winnings):

- $W(h) = 1$
- $W(t) = 3$

The expectation of W is

$$\begin{aligned} E(W) &= \Pr(h)W(h) + \Pr(t)W(t) \\ &= \Pr(W = 1) \times 1 + \Pr(W = 3) \times 3 \end{aligned}$$

More generally, the *expected value* of random variable X on sample space S is

$$E(X) = \sum_x x \Pr(X = x)$$

Example: What is the expected count when two dice are rolled?

Let X be the count:

$$\begin{aligned} & E(X) \\ = & \sum_{i=2}^{12} i \Pr(X = i) \\ = & 2\tfrac{1}{36} + 3\tfrac{2}{36} + 4\tfrac{3}{36} + \cdots + 7\tfrac{6}{36} + \cdots + 12\tfrac{1}{36} \\ = & \tfrac{252}{36} \\ = & 7 \end{aligned}$$

An Alternative Definition of Expectation

We defined $E(X) = \sum_x x \Pr(X = x)$.

Let $E'(X) = \sum_{s \in S} X(s) \Pr(s)$.

The two definitions are equivalent:

Theorem: $E(X) = E'(X)$

An Alternative Definition of Expectation

We defined $E(X) = \sum_x x \Pr(X = x)$.

Let $E'(X) = \sum_{s \in S} X(s) \Pr(s)$.

The two definitions are equivalent:

Theorem: $E(X) = E'(X)$

Proof:

$$
\begin{aligned}
E'(X) &= \textstyle\sum_{s \in S} X(s) \Pr(s) \\
&= \textstyle\sum_x \sum_{\{s \in S : X(s) = x\}} X(s) \Pr(s) && \text{[partition the sum by } x\text{]} \\
&= \textstyle\sum_x \sum_{\{s \in S : X(s) = x\}} x \Pr(s) \\
&= \textstyle\sum_x x \sum_{\{s \in S : X(s) = x\}} \Pr(s) && [x \text{ a constant}] \\
&= \textstyle\sum_x x \Pr(\{s : X(s) = x\}) \\
&= \textstyle\sum_x x \Pr(\{X = x\}) && \text{[by definition, } X = x \\
& && \text{is the event } \{s : X(s) = x\}] \\
&= E(X)
\end{aligned}
$$

Expectation of Indicator Variables

What is the expected value of the indicator variable I_A? Recall

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Thus, $E(I_A) = 1\Pr(I_A = 1) + 0\Pr(I_A = 0) = \Pr(A)$,

- Since $\{s : I_A(s) = 1\} = A$.

Expectation of Binomials

What is $E(B_{n,p})$, the expectation for the binomial distribution $B_{n,p}$

- How many heads do you expect to get after n tosses of a biased coin with $\Pr(h) = p$?

Expectation of Binomials

What is $E(B_{n,p})$, the expectation for the binomial distribution $B_{n,p}$

- How many heads do you expect to get after n tosses of a biased coin with $\Pr(h) = p$?

Method 1: Use the definition and crank it out:

$$E(B_{n,p}) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

This looks awful, but it can be calculated ...

Expectation of Binomials

What is $E(B_{n,p})$, the expectation for the binomial distribution $B_{n,p}$

- How many heads do you expect to get after n tosses of a biased coin with $\Pr(h) = p$?

Method 1: Use the definition and crank it out:

$$E(B_{n,p}) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

This looks awful, but it can be calculated ...

Method 2: Use Induction; break it up into what happens on the first toss and on the later tosses.

- On the first toss you get heads with probability p and tails with probability $1-p$. On the last $n-1$ tosses, you expect $E(B_{n-1,p})$ heads. Thus, the expected number of heads is:

$$\begin{aligned} E(B_{n,p}) &= p(1 + E(B_{n-1,p})) + (1-p)(E(B_{n-1,p})) \\ &= p + E(B_{n-1,p}) \\ E(B_{1,p}) &= p \end{aligned}$$

Now an easy induction shows that $E(B_{n,p}) = np$.

Expectation is Linear

Theorem: $E(X+Y) = E(X) + E(Y)$

Proof: Recall that

$$E(X) = \sum_{s \in S} \Pr(s) X(s)$$

Thus,

$$\begin{aligned} E(X+Y) &= \textstyle\sum_{s \in S} \Pr(s)(X+Y)(s) \\ &= \textstyle\sum_{s \in S} \Pr(s) X(s) + \sum_{s \in S} \Pr(s) Y(s) \\ &= E(X) + E(Y). \end{aligned}$$

This is true even if X and Y aren't independent!

Expectation is Linear

Theorem: $E(X+Y) = E(X) + E(Y)$

Proof: Recall that

$$E(X) = \sum_{s \in S} \Pr(s) X(s)$$

Thus,

$$\begin{aligned} E(X+Y) &= \textstyle\sum_{s \in S} \Pr(s)(X+Y)(s) \\ &= \textstyle\sum_{s \in S} \Pr(s) X(s) + \sum_{s \in S} \Pr(s) Y(s) \\ &= E(X) + E(Y). \end{aligned}$$

This is true even if X and Y aren't independent!

Theorem: $E(aX) = aE(X)$

Proof:

$$E(aX) = \sum_{s \in S} \Pr(s)(aX)(s) = a \sum_{s \in S} \Pr(s) X(s) = aE(X).$$

Example 1: Back to the expected value of tossing two dice: Let X_1 be the count on the first die, X_2 the count on the second die, and let X be the total count.

Notice that

$$E(X_1) = E(X_2) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$$

$$E(X) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$$

Example 1: Back to the expected value of tossing two dice:
Let X_1 be the count on the first die, X_2 the count on the second die, and let X be the total count.

Notice that

$$E(X_1) = E(X_2) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$$

$$E(X) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$$

Example 2: Back to the expected value of $B_{n,p}$.

Let X be the total number of successes and let X_k be the outcome of the kth experiment, $k = 1, \ldots, n$:

$$E(X_k) = p \cdot 1 + (1 - p) \cdot 0 = p$$

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

Therefore

$$E(X) = E(X_1) + \cdots + E(X_n) = np.$$

$E(XY)$

In general, $E(XY) \neq E(X)E(Y)$.
A simple counterexample: Suppose that X is Bernoulli with success probability p.

- $E(X) = p$ (X is 1 with probability p, 0 with probability $1 - p$)
- $X^2 = X$
- Thus, $E(X^2) = E(X) \neq (E(X))^2$.

$E(XY)$

In general, $E(XY) \neq E(X)E(Y)$.
A simple counterexample: Suppose that X is Bernoulli with success probability p.

- $E(X) = p$ (X is 1 with probability p, 0 with probability $1 - p$)
- $X^2 = X$
- Thus, $E(X^2) = E(X) \neq (E(X))^2$.

Theorem: If X and Y are independent, then $E(XY) = E(X)E(Y)$.
Proof:

$$\begin{aligned} E(XY) &= \textstyle\sum_z z \Pr(XY = z) \\ &= \textstyle\sum_{xy} xy \Pr(X = x \cap Y = y) \\ &= \textstyle\sum_{xy} xy \Pr(X = x) \Pr(Y = y) \\ &= \textstyle\sum_x x \Pr(X = x) \sum_y y \Pr(Y = y) \\ &= E(X)E(Y) \end{aligned}$$

Conditional Expectation

$E(X \mid A)$ is the *conditional expectation* of X given A.

$$\begin{aligned} E(X \mid A) &= \textstyle\sum_x x \Pr(X = x \mid A) \\ &= \textstyle\sum_x x \Pr(X = x \cap A) / \Pr(A) \end{aligned}$$

Theorem: For all events A such that $\Pr(A), \Pr(\overline{A}) > 0$:

$$E(X) = E(X \mid A) \Pr(A) + E(X \mid \overline{A}) \Pr(\overline{A})$$

Conditional Expectation

$E(X \mid A)$ is the *conditional expectation* of X given A.

$$\begin{aligned} E(X \mid A) &= \textstyle\sum_x x \Pr(X = x \mid A) \\ &= \textstyle\sum_x x \Pr(X = x \cap A) / \Pr(A) \end{aligned}$$

Theorem: For all events A such that $\Pr(A), \Pr(\overline{A}) > 0$:

$$E(X) = E(X \mid A) \Pr(A) + E(X \mid \overline{A}) \Pr(\overline{A})$$

Proof:

$$\begin{aligned} & E(X) \\ = \; & \textstyle\sum_x x \Pr(X = x) \\ = \; & \textstyle\sum_x x(\Pr((X = x) \cap A) + \Pr((X = x) \cap \overline{A})) \\ = \; & \textstyle\sum_x x(\Pr(X = x \mid A) \Pr(A) + \Pr(X = x \mid \overline{A}) \Pr(\overline{A})) \\ = \; & \textstyle\sum_x (x \Pr(X = x \mid A) \Pr(A)) + (x \Pr(X = x \mid \overline{A}) \Pr(\overline{A})) \\ = \; & E(X \mid A) \Pr(A) + E(X \mid \overline{A}) \Pr(\overline{A}) \end{aligned}$$

Example: I toss a fair die. If it lands with 3 or more, I toss a coin with bias p_1 (towards heads). If it lands with less than 3, I toss a coin with bias p_2. What is the expected number of heads?

Let A be the event that the die lands with 3 or more.

$$\Pr(A) = 2/3$$

$$\begin{aligned} E(\#H) &= E(\#H \mid A)\Pr(A) + E(\#H \mid \overline{A})\Pr(\overline{A}) \\ &= p_1 \tfrac{2}{3} + p_2 \tfrac{1}{3} \end{aligned}$$

The Power of Randomization

Suppose we play the following game:

- Step 1: We have two envelopes. I put different integers between 0 and 100 in each one.
- Step 2: An envelope is chosen at random and the number inside is revealed.
- Step 3: You choose an envelope. You win if you choose the envelope with the larger number.

If you just choose an envelope at random (ignoring what you saw at step 2), you will win with probability 1/2.

- Can you do better (assuming that I know your strategy)?
  (a) Yes
  (b) No
  (c) ??

A potentially better strategy

Here's a strategy:

- Fix a number between 0 and 100; e.g., 60.
- If the number in the envelope that you see is ≥ 60, stick with that envelope; otherwise switch.

Call this strategy $S(60)$: stick with 60

- Can similarly define $S(k)$ for $1 \leq k \leq 100$

Analysis of $S(60)$

Recall that with $S(60)$, you stick with the opened envelope if you see a number ≥ 60; otherwise you switch.

- Case A: If the numbers in both envelopes are ≥ 60, you win with probability $1/2$.
  - whichever envelope you open, you'll stick with it.
  - $\Pr(\text{You open the envelope with the bigger number}) = 1/2$.
- Case B: If the numbers in both envelopes are < 60, you win with probability $1/2$.
  - whichever envelope you open, you'll switch.
  - $\Pr(\text{You open the envelope with the smaller number}) = 1/2$.
- Case C: If the number in one envelope is < 60 and the number in the other one is ≥ 60, you're guaranteed to win!
  - if you open the envelope with the bigger number, you stick with it (it's ≥ 60); if you open the envelope with the smaller number, you switch.

Thus, the expected probability that you'll win is

$$\begin{aligned} & 1/2\Pr(A) + 1/2\Pr(B) + \Pr(C) \\ = \; & 1/2(\Pr(A) + \Pr(B) + \Pr(C)) + 1/2\Pr(C) \\ = \; & 1/2 + 1/2\Pr(C) \quad [\text{since } \Pr(A) + \Pr(B) + \Pr(C) = 1] \end{aligned}$$

You win with probability $= 1/2$ iff $\Pr(C) = 0$:

- i.e., if I never put numbers in the envelope where one is ≥ 60 and the other is < 60.

So how can you ensure that you win?

Note that $S(k)$ will win with probability $> 1/2$ as long as I put a number $\geq k$ in one envelope and a number $< k$ in the other with positive probability.

- I can beat any strategy $S(k)$, but I can't beat them all.
  - If you choose k between 1 and 100 at random (i.e., with probability $1/100$) and play $S(k)$, you're guaranteed to win with probability $> 1/2$!

You win by randomizing!

A CS Application: Primality Testing

Key number theory result: There is an easily computable (deterministic) test $T(b, n)$ such that

- $T(b, n) = 1$ (for all b) if n is prime.
- There are lots of bs for which $T(b, n) = 0$ if n is not prime.
  - In fact, for at least $1/3$ of the the bs between 1 and n, $T(b, n) = 0$ if n is composite.

So heres a primality-testing algorithm:

Input n [the number you want to test for primality]
For k from 1 to 100 **do**
  Choose b at random between 1 and n
  If $T(b, n) = 0$ **return** "n is not prime"
EndFor
return "n is prime"

Probabilistic Primality Testing: Analysis

If n is composite, what is the probability that algorithm returns "n is prime":

Probabilistic Primality Testing: Analysis

If n is composite, what is the probability that algorithm returns "n is prime":

$$(2/3)^{100} < (.2)^{25} = 10^{-70}$$

- I wouldn't lose sleep over mistakes!

If 10^{-70} is unacceptable, try 200 random choices.

Probabilistic Primality Testing: Analysis

If n is composite, what is the probability that algorithm returns "n is prime":

$$(2/3)^{100} < (.2)^{25} = 10^{-70}$$

- I wouldn't lose sleep over mistakes!

If 10^{-70} is unacceptable, try 200 random choices.

How long will it take until we find a witness?

Probabilistic Primality Testing: Analysis

If n is composite, what is the probability that algorithm returns "n is prime":

$$(2/3)^{100} < (.2)^{25} = 10^{-70}$$

- I wouldn't lose sleep over mistakes!

If 10^{-70} is unacceptable, try 200 random choices.

How long will it take until we find a witness?

- Expected number of steps is ≤ 3

What is the probability that it takes k steps to find a witness?

Probabilistic Primality Testing: Analysis

If n is composite, what is the probability that algorithm returns "n is prime":

$$(2/3)^{100} < (.2)^{25} = 10^{-70}$$

- I wouldn't lose sleep over mistakes!

If 10^{-70} is unacceptable, try 200 random choices.

How long will it take until we find a witness?

- Expected number of steps is ≤ 3

What is the probability that it takes k steps to find a witness?

- $(2/3)^{k-1}(1/3)$
- That's the probability of not finding a witness for the first $k-1$ steps ($(2/3)^{k-1}$) then finding a witness the kth step ($1/3$)

Bottom line: the algorithm is extremely fast and almost certainly gives the right results.

Probability Surprises I: The Birthday Paradox

Probability occasionally behaves in counterintuitive ways. Let's start with the birthday paradox.

- How likely is it that two people in this class have the same birthday?

Number the people in the class 1, . . . , 48.

- Let X_{ij} be 1 if person i and j have the same birthday, 0 otherwise.
- What's the probability that $X_{ij} = 1$?

Probability Surprises I: The Birthday Paradox

Probability occasionally behaves in counterintuitive ways. Let's start with the birthday paradox.

- How likely is it that two people in this class have the same birthday?

Number the people in the class 1, . . . , 48.

- Let X_{ij} be 1 if person i and j have the same birthday, 0 otherwise.
- What's the probability that $X_{ij} = 1$?
- $1/365$.
- $X = \sum_{ij:i<j} X_{ij}$ is the total number of pairs of people who share a birthday.
- What's $E(X)$?
- Since there are 48 people in this class, for this class, it's $C(48, 2)/365 \approx 3.1$

Probability Surprises I: The Birthday Paradox

Probability occasionally behaves in counterintuitive ways. Let's start with the birthday paradox.

- How likely is it that two people in this class have the same birthday?

Number the people in the class 1, . . . , 48.

- Let X_{ij} be 1 if person i and j have the same birthday, 0 otherwise.
- What's the probability that $X_{ij} = 1$?
- $1/365$.
- $X = \sum_{ij:i<j} X_{ij}$ is the total number of pairs of people who share a birthday.
- What's $E(X)$?
- Since there are 48 people in this class, for this class, it's $C(48, 2)/365 \approx 3.1$
- And indeed, we have three pairs of people with shared birthdays!

How many people do you need before you get a shared birthday?

How many people do you need before you get a shared birthday?

- You want the expected number of matches to be at least 1, so we want $C(n, 2) \geq 365$
- This happens if $n \geq 28$.
  - So with 28 people you are likely to have a shared birthday!

More voter fraud?

It has been claimed that there were over 3,00,000 fraudulent votes (out of the ≈130,000,000 cast in the presidential election).

- What's a fraudulent vote?
- Apparently, it's the same voter registered in different states.
- But how do they know it's the same voter?

More voter fraud?

It has been claimed that there were over 3,00,000 fraudulent votes (out of the ≈130,000,000 cast in the presidential election).

- What's a fraudulent vote?
- Apparently, it's the same voter registered in different states.
- But how do they know it's the same voter?
  - Same name and same birthday
  - Must be fraudulent!
- To address this carefully, we should leave out obvious errors
  - e.g., lots of people are listed as having a birthday of Jan. 1.
- That leaves about 750,000 votes
- How many will have the same birthday?
- By the pigeonhole principle, the expected number of people with the same birthday $> 2,000$
- What about same names?
  - There are lots of common names!
  - Moreover, names and birthdays are not uniformly distributed
    - Carol, Christine, and Jesus are more likely on 25 Dec.

Bottom line: there's no evidence of fraud!

For more on this, check out:

- https://www.thisamericanlife.org/630/things-i-mean-to-know
- http://www.slate.com/articles/news_and_politics/ jurisprudence/2016/11/we_looked_at_130_million_ballots_from_the_2012_election_and_found_zero_fraud.html

Deviation from the Mean

Expectation summarizes a lot of information about a random variable as a single number. But no single number can tell it all.

Compare these two distributions:

- Distribution 1:

$$\Pr(49) = \Pr(51) = 1/4; \quad \Pr(50) = 1/2.$$

- Distribution 2: $\Pr(0) = \Pr(50) = \Pr(100) = 1/3$.

Both have the same expectation: 50. But the first is much less "dispersed" than the second. We want a measure of *dispersion*.

- One measure of dispersion is how far things are from the mean, on average.

Given a random variable X, $(X(s) - E(X))^2$ measures how far the value of s is from the mean value (the expectation) of X. Define the *variance* of X to be

$$\mathrm{Var}(X) = E((X - E(X))^2) = \sum_{s \in S} \Pr(s)(X(s) - E(X))^2$$

Standard Deviation

The *standard deviation* of X is

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\sum_{s \in S} \Pr(s)(X(s) - E(X))^2}$$

Why not use $|X(s) - E(X)|$ as the measure of distance instead of variance?

- $(X(s) - E(X))^2$ turns out to have nicer mathematical properties.
- In R^n, the distance between $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ is $\sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$

Why not use $|X(s) - E(X)|$ as the measure of distance instead of variance?

- $(X(s) - E(X))^2$ turns out to have nicer mathematical properties.
- In R^n, the distance between $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ is $\sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$

Example:

- The variance of distribution 1 is

$$\frac{1}{4}(51 - 50)^2 + \frac{1}{2}(50 - 50)^2 + \frac{1}{4}(49 - 50)^2 = \frac{1}{2}$$

- The variance of distribution 2 is

$$\frac{1}{3}(100 - 50)^2 + \frac{1}{3}(50 - 50)^2 + \frac{1}{3}(0 - 50)^2 = \frac{5000}{3}$$

Expectation and variance are two ways of compactly describing a distribution.

- They don't completely describe the distribution
- But they're still useful!

Variance: Example

Let X be Bernoulli, with probability p of success. $E(X) = p$, so

$$\begin{aligned}\mathrm{Var}(X) &= (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p(1-p)[p + (1-p)] \\ &= p(1-p)\end{aligned}$$

Variance Properties

Theorem 1: $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

Proof:

$$\begin{aligned} E((X - E(X))^2) &= E(X^2 - 2E(X)X + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

Variance Properties

Theorem 1: $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

Proof:

$$\begin{aligned} E((X - E(X))^2) &= E(X^2 - 2E(X)X + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

Example: Suppose X is the outcome of a roll of a fair die.

- Recall $E(X) = 7/2$.
- $E(X^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \cdots + 6^2 \cdot \frac{1}{6} = \frac{91}{6}$
- So $\mathrm{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$.

Theorem 2: $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

Theorem 2: $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

Theorem 3: If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof:

$$
\begin{aligned}
& \mathrm{Var}(X+Y) \\
= \; & E((X+Y)^2) - (E(X+Y))^2 \\
= \; & E(X^2) + E(Y^2) + 2E(XY) - (E(X)^2 + 2(E(X)E(Y) + E(Y)^2) \\
= \; & E(X^2) + E(Y^2) + 2E(X)E(Y) - E(X)^2 - 2(E(X)E(Y) - E(Y)^2 \\
& \text{[since } X \text{ and } Y \text{ are independent]} \\
= \; & E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 \\
= \; & \mathrm{Var}(X) + \mathrm{Var}(Y)
\end{aligned}
$$

Theorem 2: $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

Theorem 3: If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof:

$$\begin{aligned}
& \mathrm{Var}(X+Y) \\
= \; & E((X+Y)^2) - (E(X+Y))^2 \\
= \; & E(X^2) + E(Y^2) + 2E(XY) - (E(X)^2 + 2(E(X)E(Y) + E(Y)^2) \\
= \; & E(X^2) + E(Y^2) + 2E(X)E(Y) - E(X)^2 - 2(E(X)E(Y) - E(Y)^2 \\
& \text{[since } X \text{ and } Y \text{ are independent]} \\
= \; & E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 \\
= \; & \mathrm{Var}(X) + \mathrm{Var}(Y)
\end{aligned}$$

Example: Variance of $B_{n,p}$;
Recall that $B_{n,p} = X_1 + \cdots + X_n$, where X_i is Bernoulli with probability p of success.

$\mathrm{Var}(X_i) = p(1-p)$, so $\mathrm{Var}(B_{n,p}) = np(1-p)$.

Markov's Inequality

How likely is it that things are far from the mean? Markov's inequality gives one estimate:

Theorem: Suppose that X is a nonnegative random variable and $\alpha > 0$. Then $\Pr(X \geq \alpha) \leq \frac{E(X)}{\alpha}$.

Proof:

$$\begin{aligned} E(X) &= \textstyle\sum_x x \cdot \Pr(X = x) \\ &\geq \textstyle\sum_{x \geq \alpha} x \cdot \Pr(X = x) \quad [X \text{ is nonnegative}] \\ &\geq \textstyle\sum_{x \geq \alpha} \alpha \cdot \Pr(X = x) \\ &= \alpha \textstyle\sum_{x \geq \alpha} \Pr(X = x) \\ &= \alpha \cdot \Pr(X \geq \alpha) \end{aligned}$$

Markov's Inequality

How likely is it that things are far from the mean? Markov's inequality gives one estimate:

Theorem: Suppose that X is a nonnegative random variable and $\alpha > 0$. Then $\Pr(X \geq \alpha) \leq \frac{E(X)}{\alpha}$.

Proof:

$$\begin{aligned} E(X) &= \textstyle\sum_x x \cdot \Pr(X = x) \\ &\geq \textstyle\sum_{x \geq \alpha} x \cdot \Pr(X = x) \quad [X \text{ is nonnegative}] \\ &\geq \textstyle\sum_{x \geq \alpha} \alpha \cdot \Pr(X = x) \\ &= \alpha \textstyle\sum_{x \geq \alpha} \Pr(X = x) \\ &= \alpha \cdot \Pr(X \geq \alpha) \end{aligned}$$

Example: If X is $B_{100,1/2}$, then

$$\Pr(X \geq 100) \leq 50/100 = 1/2.$$

This is not a particularly useful estimate. In fact, $Pr(X \geq 100) = 2^{-100} \sim 10^{-30}$.

Chebyshev's Inequality

Theorem: If X is a random variable and $\beta > 0$, then

$$\Pr(|X - E(X)| \geq \beta) \leq \frac{\mathrm{Var}(X)}{\beta^2}.$$

Proof: Let $Y = (X - E(X))^2$. Then

$$|X - E(X)| \geq \beta \text{ iff } Y \geq \beta^2.$$

That is, $\{s : |X(s) - E(X)| \geq \beta\} = \{s : Y(s) \geq \beta^2\}$.
Thus

$$\Pr(|X - E(X)| \geq \beta) = \Pr(Y \geq \beta^2).$$

Since $Y \geq 0$, by Markov's inequality,

$$\Pr(Y \geq \beta^2) \leq \frac{E(Y)}{\beta^2}.$$

Finally, note that $E(Y) = E[(X - E(X))^2] = \mathrm{Var}(X)$.

- Statement equivalent to Chebyshev's inequality:

$$\Pr(|X - E(X)| \geq \beta\sigma_X) \leq \frac{1}{\beta^2}.$$

- Intuitively, the probability of a random variable being k standard deviations from the mean is $\leq 1/k^2$.
- Chebyshev's inequality gives a better estimate of how far things are from the mean than Markov's inequality, although Markov's inequality is used to prove it.
- If we have more information, we can do even better.
  - See the discussion of Chernoff bounds in the text

Chebyshev's Inequality: Example

Chebyshev's inequality gives a lower bound on how well X is concentrated about its mean.

- Suppose that X is $B_{100,1/2}$ and we want a lower bound on $\Pr(40 < X < 60)$.
- $E(X) = 50$ and $40 < X < 60$ iff $|X - 50| < 10$ so

$$\begin{aligned} \Pr(40 < X < 60) &= \Pr(|X - 50| < 10) \\ &= 1 - \Pr(|X - 50| \geq 10). \end{aligned}$$

  Now

$$\begin{aligned} \Pr(|X - 50| \geq 10) &\leq \frac{\mathrm{Var}(X)}{10^2} \\ &= \frac{100 \cdot (1/2)^2}{100} \\ &= \frac{1}{4}. \end{aligned}$$

  So $\Pr(40 < X < 60) \geq 1 - 1/4 = 3/4$.

This is not too bad: the correct answer is ~ 0.9611.

Probability Surprises II: Simpson's Paradox

Derek Jeter and David Justice were two well-known baseball players in the 1990s. Here are their batting average in 1995, 1996, and 1997.

- batting average = #hits/#times at bat

	1995	1996	1997
David Justice	.253	.321	.329
Derek Jeter	.250	.314	.291

Probability Surprises II: Simpson's Paradox

Derek Jeter and David Justice were two well-known baseball players in the 1990s. Here are their batting average in 1995, 1996, and 1997.

- batting average = #hits/#times at bat

	1995	1996	1997	Overall
David Justice	.253	.321	.329	.298
Derek Jeter	.250	.314	.291	.300

Probability Surprises II: Simpson's Paradox

Derek Jeter and David Justice were two well-known baseball players in the 1990s. Here are their batting average in 1995, 1996, and 1997.

- batting average = #hits/#times at bat

	1995	1996	1997	Overall
David Justice	.253	.321	.329	.298
Derek Jeter	.250	.314	.291	.300

How could this happen??

Probability Surprises II: Simpson's Paradox

Derek Jeter and David Justice were two well-known baseball players in the 1990s. Here are their batting average in 1995, 1996, and 1997.

- batting average = #hits/#times at bat

	1995	1996	1997	Overall
David Justice	104/411 = .253	45/140 = .321	163/495 = .329	312/1046 = .298
Derek Jeter	12/48 = .250	183/582 = .314	190/654 = .291	385/1284 = .300

Probability Surprises II: Simpson's Paradox

Derek Jeter and David Justice were two well-known baseball players in the 1990s. Here are their batting average in 1995, 1996, and 1997.

- batting average = #hits/#times at bat

	1995	1996	1997	Overall
David Justice	104/411 = .253	45/140 = .321	163/495 = .329	312/1046 = .298
Derek Jeter	12/48 = .250	183/582 = .314	190/654 = .291	385/1284 = .300

- The year that Derek Jeter had a poor batting average, he didn't bat much, so it didn't affect his 3-year average much; the
- The year that David Justice had a poor batting average, he batted a lot.

So is it OK to consider 3-year average in this case?

How about this case?

We want to check the affect of a drug on heart disease, so we give the treatment group the drug, and also consider a control group who get a placebo:

	Control Group		Treatment group	
	Heart attack	OK	Heart attack	OK
Female	1	19	3	37
Male	12	28	8	12

The drug is

- bad for women: $1/19 \to 3/37$
- bad for men: $12/28 \to 8/12$

How about this case?

We want to check the affect of a drug on heart disease, so we give the treatment group the drug, and also consider a control group who get a placebo:

	Control Group		Treatment group	
	Heart attack	OK	Heart attack	OK
Female	1	19	3	37
Male	12	28	8	12
Overall	13	47	11	49

The drug is

- bad for women: $1/19 \to 3/37$
- bad for men: $12/28 \to 8/12$
- good for people: $13/47 \to 11/49$!

Why might this have happened?

How about this case?

We want to check the affect of a drug on heart disease, so we give the treatment group the drug, and also consider a control group who get a placebo:

	Control Group		Treatment group	
	Heart attack	OK	Heart attack	OK
Female	1	19	3	37
Male	12	28	8	12
Overall	13	47	11	49

The drug is

- bad for women: $1/19 \to 3/37$
- bad for men: $12/28 \to 8/12$
- good for people: $13/47 \to 11/49$!

Why might this have happened?

- Gender affects the propensity to heart attack
- Data shouldn't be combined in that case!

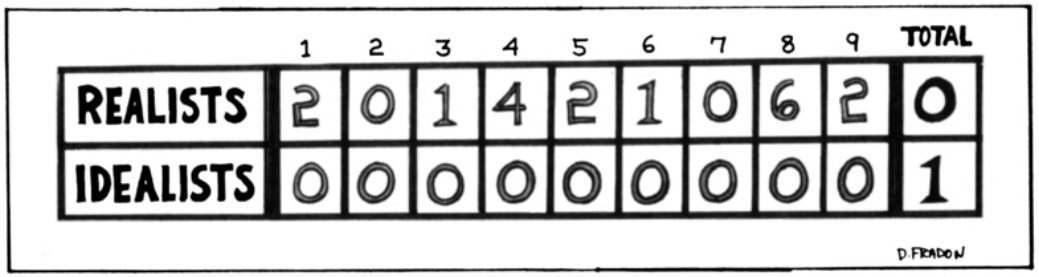
There are a number of real-world instances of Simpson's paradox:

- admissions at Berkeley (discussed in text)
- medical drug tests
- ...

The question of whether to combine the data or keep it separate is arguably best understood in terms of causality:

- see Judea Pearl's books *Causality* or *The Book of Why*

One more example of Simpson's Paradox . . .



	1	2	3	4	5	6	7	8	9	TOTAL
REALISTS	2	0	1	4	2	1	0	6	2	0
IDEALISTS	0	0	0	0	0	0	0	0	0	1

D. FRADON