Massachusetts Institute of Technology
6.042J/18.062J, Fall '05: Mathematics for Computer Science
Prof. Albert R. Meyer and Prof. Ronitt Rubinfeld

Course Notes, Week 12
November 21
revised November 29, 2005, 1285 minutes

Introduction to Probability

Probability is the last topic in this course and perhaps the most important. Many algorithms rely on randomization. Investigating their correctness and performance requires probability theory. Moreover, many aspects of computer systems, such as memory management, branch prediction, packet routing, and load balancing are designed around probabilistic assumptions and analyses. Probability also comes up in information theory, cryptography, artificial intelligence, and game theory. Beyond these engineering applications, an understanding of probability gives insight into many everyday issues, such as polling, DNA testing, risk assessment, investing, and gambling.

So probability is good stuff.

1 Monty Hall

In the September 9, 1990 issue of *Parade* magazine, the columnist Marilyn vos Savant responded to this letter:

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

The letter roughly describes a situation faced by contestants on the 1970's game show *Let's Make a Deal*, hosted by Monty Hall and Carol Merrill. Marilyn replied that the contestant should indeed switch. But she soon received a torrent of letters— many from mathematicians— telling her that she was wrong. The problem generated thousands of hours of heated debate.

Yet this is is an elementary problem with an elementary solution. Why was there so much dispute? Apparently, most people *believe* they have an intuitive grasp of probability. (This is in stark contrast to other branches of mathematics; few people believe they have an intuitive ability to compute integrals or factor large integers!) Unfortunately, approximately 100% of those people are *wrong*. In fact, everyone who has studied probability at length can name a half-dozen problems in which their intuition led them astray— often embarassingly so.

The way to avoid errors is to distrust informal arguments and rely instead on a rigorous, systematic approach. In short: intuition *bad*, formalism *good*. If you insist on relying on intuition, then there are lots of compelling financial deals we'd love to offer you!

1.1 The Four-Step Method

Every probability problem involves some sort of randomized experiment, process, or game. And each such problem involves two distinct challenges:

1. How do we model the situation mathematically?
2. How do we solve the resulting mathematical problem?

In this section, we introduce a four-step approach to questions of the form, "What is the probability that —– ?" In this approach, we build a probabilistic model step-by-step, formalizing the original question in terms of that model. Remarkably, the structured thinking that this approach imposes reduces many famously-confusing problems to near triviality. For example, as you'll see, the four-step method cuts through the confusion surrounding the Monty Hall problem like a Ginsu knife. However, more complex probability questions may spin off challenging counting, summing, and approximation problems— which, fortunately, you've already spent weeks learning how to solve!

1.2 Clarifying the Problem

Craig's original letter to Marilyn vos Savant is a bit vague, so we must make some assumptions in order to have any hope of modeling the game formally:

1. The car is equally likely to be hidden behind each of the three doors.
2. The player is equally likely to pick each of the three doors, regardless of the car's location.
3. After the player picks a door, the host *must* open a different door with a goat behind it and offer the player the choice of staying with the original door or switching.
4. If the host has a choice of which door to open, then he is equally likely to select each of them.

In making these assumptions, we're reading a lot into Craig Whitaker's letter. Other interpretations are at least as defensible, and some actually lead to different answers. But let's accept these assumptions for now and address the question, "What is the probability that a player who switches wins the car?"

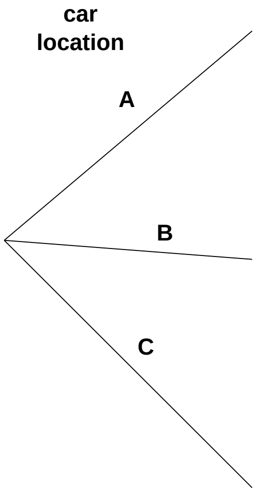
1.3 Step 1: Find the Sample Space

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

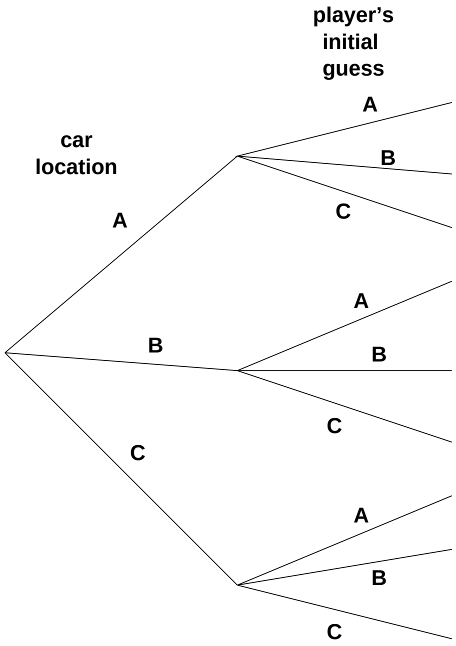
1. The door concealing the car.
2. The door initially chosen by the player.
3. The door that the host opens to reveal a goat.

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

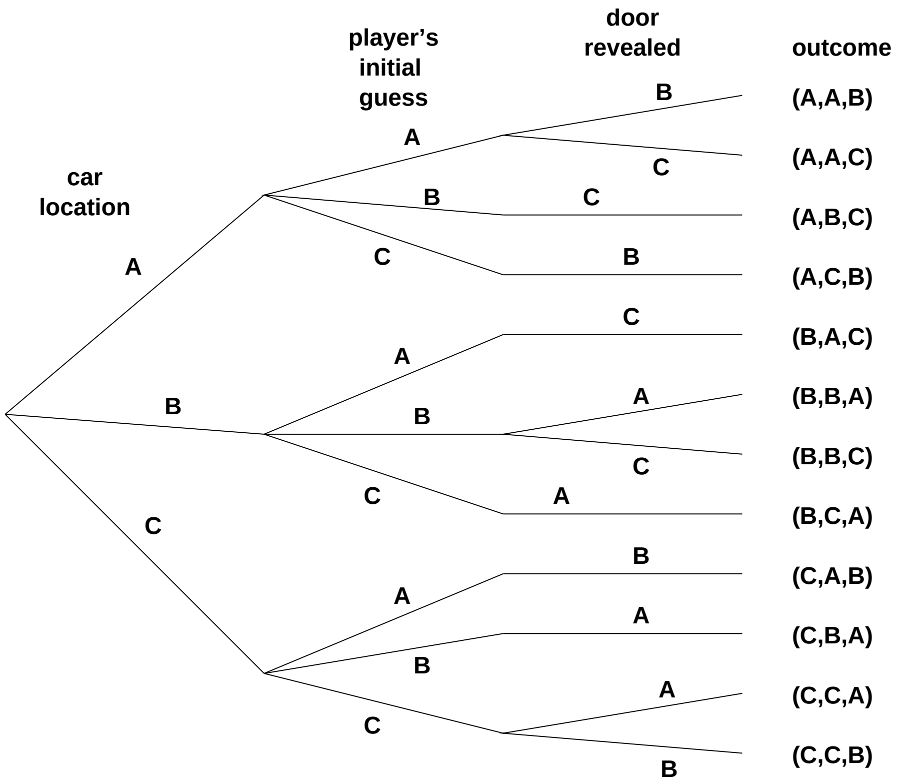
A ***tree diagram*** is a graphical tool that can help us work through the four-step approach when the number of outcomes is not too large or the problem is nicely structured. In particular, we can use a tree diagram to help understand the sample space of an experiment. The first randomly-determined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



In this diagram, the doors are called A, B, and C instead of 1, 2, and 3 because we'll be adding a lot of other numbers to the picture later. Now, for each possible location of the prize, the player could initially choose any of the three doors. We represent this by adding a second layer to the tree.



Finally, the host opens a door to reveal a goat. The host has either one choice or two, depending on the position of the car and the door initially selected by the player. For example, if the prize is behind door A and the player picks door B, then the host must open door C. However, if the prize is behind door A and the player picks door A, then the host could open either door B or door C. All of these possibilities are worked out in a third layer of the tree:



Now let's relate this picture to the terms we introduced earlier: the leaves of the tree represent *outcomes* of the experiment, and the set of all leaves represents the *sample space*. Thus, for this experiment, the sample space consists of 12 outcomes. For reference, we've labeled each outcome with a triple of doors indicating:

$$(\text{door concealing prize},\ \text{door initially chosen},\ \text{door opened to reveal a goat})$$

In these terms, the sample space is the set:

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

The tree diagram has a broader interpretation as well: we can regard the whole experiment as "walk" from the root down to a leaf, where the branch taken at each stage is randomly determined. Keep this interpretation in mind; we'll use it again later.

1.4 Step 2: Define Events of Interest

Our objective is to answer questions of the form "What is the probability that —— ?", where the horizontal line stands for some phrase such as "the player wins by switching", "the player initially picked the door concealing the prize", or "the prize is behind door

C". Almost any such phrase can be modeled mathematically as an ***event***, which is defined to be a subset of the sample space.

For example, the event that the prize is behind door C is the set of outcomes:

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

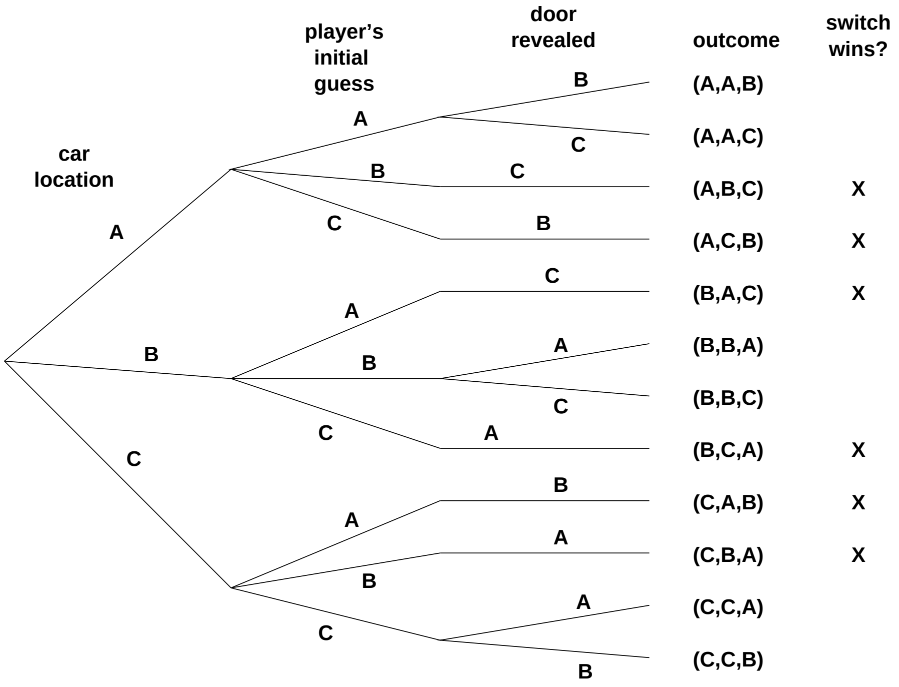
The event that the player initially picked the door concealing the prize is the set of outcomes:

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

And what we're really after, the event that the player wins by switching, is the set of outcomes:

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

Let's annotate our tree diagram to indicate the outcomes in this event.



Notice that exactly half of the outcomes are marked, meaning that the player wins by switching in half of all outcomes. You might be tempted to conclude that a player who switches wins with probability $1/2$. *This is wrong.* The reason is that these outcomes are not all equally likely, as we'll see shortly.

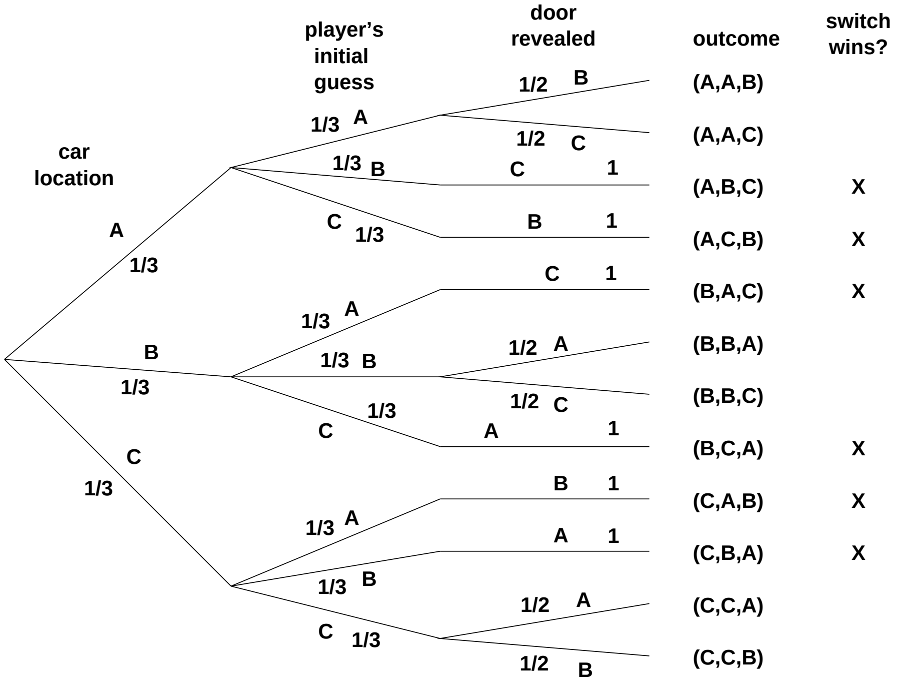
1.5 Step 3: Determine Outcome Probabilities

So far we've enumerated all the possible outcomes of the experiment. Now we must start assessing the likelihood of those outcomes. In particular, the goal of this step is to assign each outcome a probability, which is a real number between 0 and 1. The sum of all outcome probabilities must be 1, reflecting the fact that exactly one outcome must occur.

Ultimately, outcome probabilities are determined by the phenomenon we're modeling and thus are not quantities that we can derive mathematically. However, mathematics can help us compute the probability of every outcome *based on fewer and more elementary modeling decisions.* In particular, we'll break the task of determining outcome probabilities into two stages.

1.5.1 Step 3a: Assign Edge Probabilities

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



1.5.2 Step 3b: Compute Outcome Probabilities

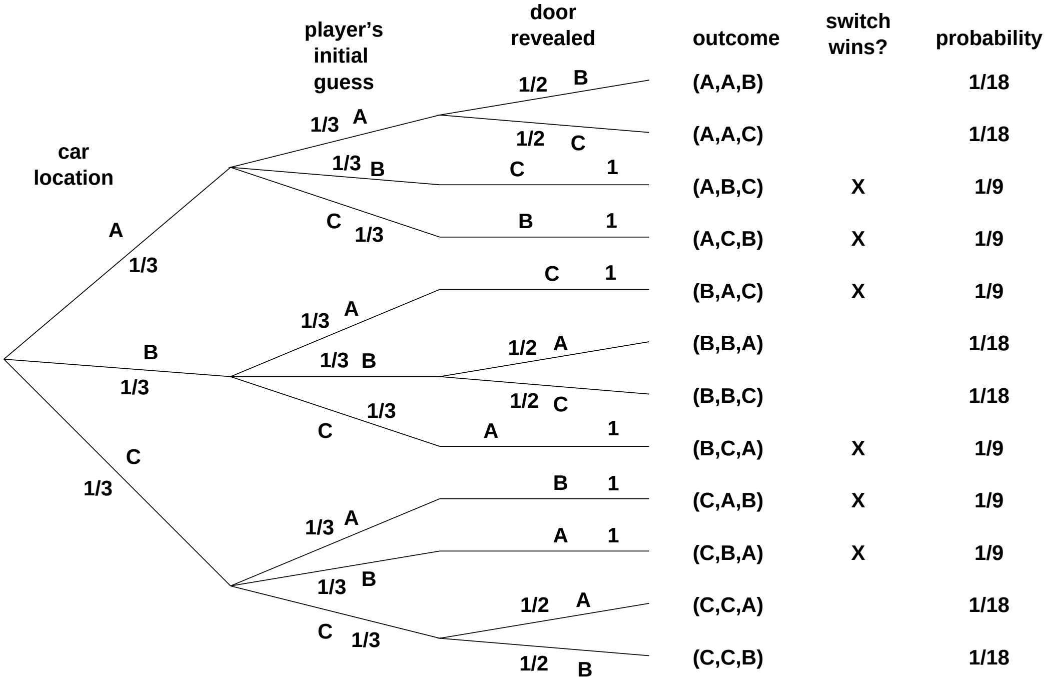
Our next job is to convert edge probabilities into outcome probabilities. This is a purely mechanical process: *the probability of an outcome is equal to the product of the edge-probabilities on the path from the root to that outcome*. For example, the probability of the topmost outcome, (A, A, B) is

$$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{18}.$$

We'll justify this process formally later on. In the meantime, here is a nice informal justification to tide you over. Remember that the whole experiment can be regarded as a walk from the root of the tree diagram down to a leaf, where the branch taken at each step is randomly determined. In particular, the probabilities on the edges indicate how likely the walk is to proceed along each path. For example, a walk starting at the root in our example is equally likely to go down each of the three top-level branches.

Now, how likely is such a walk to arrive at the topmost outcome, (A, A, B)? Well, there is a 1-in-3 chance that a walk would follow the A-branch at the top level, a 1-in-3 chance it would continue along the A-branch at the second level, and 1-in-2 chance it would follow the B-branch at the third level. Thus, it seems that about 1 walk in 18 should arrive at the (A, A, B) leaf, which is precisely the probability we assign it.

Anyway, let's record all the outcome probabilities in our tree diagram.



outcome	switch wins?	probability
(A,A,B)		1/18
(A,A,C)		1/18
(A,B,C)	X	1/9
(A,C,B)	X	1/9
(B,A,C)	X	1/9
(B,B,A)		1/18
(B,B,C)		1/18
(B,C,A)	X	1/9
(C,A,B)	X	1/9
(C,B,A)	X	1/9
(C,C,A)		1/18
(C,C,B)		1/18

Specifying the probability of each outcome amounts to defining a function that maps each outcome to a probability. This function is usually called **Pr**. In these terms, we've just determined that:

$$\begin{aligned}
\Pr\{(A,A,B)\} &= \frac{1}{18}\\
\Pr\{(A,A,C)\} &= \frac{1}{18}\\
\Pr\{(A,B,C)\} &= \frac{1}{9}\\
&\text{etc.}
\end{aligned}$$

Earlier, we noted that the sum of all outcome probabilties must be 1 since exactly one outcome must occur. We can now express this symbolically:

$$\sum_{x\in\mathcal{S}} \Pr\{x\} = 1$$

In this equation, $\mathcal{S}$ denotes the sample space.

Though $\Pr$ is an ordinary function, just like your old friends f and g from calculus, we will subject it to all sorts of horrible notational abuses that f and g were mercifully spared. Just for starters, all of the following are common notations for the probability of an outcome x:

$$\Pr\{x\} \qquad \Pr(x) \qquad \Pr[x] \qquad \Pr x \qquad p(x)$$

A sample space $\mathcal{S}$ and a probability function $\Pr : \mathcal{S} \to [0,1]$ together form a ***probability space***. Thus, a probability space describes all possible outcomes of an experiment *and* the probability of each outcome. A probability space is a complete mathematical model of an experiment.

1.6 Step 4: Compute Event Probabilities

We now have a probability for each *outcome*, but we want to determine the probability of an *event*. We can bridge this gap with a definition:

> The *probability of an event* is the sum of the probabilities of the outcomes it contains.

As a notational matter, the probability of an event $E \subseteq \mathcal{S}$ is written $\Pr\{E\}$. Thus, our definition of the probability of an event can be written:

$$\Pr\{E\} ::= \sum_{x\in E} \Pr\{x\}.$$

For example, the probability of the event that the player wins by switching is:

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

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

We're done with the problem! We didn't need any appeals to intuition or ingenious analogies. In fact, no mathematics more difficult than adding and multiplying fractions was required. The only hard part was resisting the temptation to leap to an "intuitively obvious" answer.

1.7 An Alternative Interpretation of the Monty Hall Problem

Was Marilyn really right? Our analysis suggests she was. But a more accurate conclusion is that her answer is correct *provided we accept her interpretation of the question*. There is an equally plausible interpretation in which Marilyn's answer is wrong. Notice that Craig Whitaker's original letter does not say that the host is *required* to reveal a goat and offer the player the option to switch, merely that he *did* these things. In fact, on the *Let's Make a Deal* show, Monty Hall sometimes simply opened the door that the contestant picked initially. Therefore, if he wanted to, Monty could give the option of switching only to contestants who picked the correct door initially. In this case, switching never works!

1.8 Probability Identities

The definitions we've introduced lead to some useful identities involving probabilities. Many probability problems can be solved quickly with such identities, once you're used to them. If E is an event, then the ***complement of*** E consists of all outcomes not in E and is denoted $\overline{E}$. The probabilities of complementary events sum to 1:

$$\Pr\{E\} + \Pr\left\{\overline{E}\right\} = 1.$$

About half of the time, the easiest way to compute the probability of an event is to compute the probability of its complement and then apply this formula.

Suppose that events $E_1, \ldots, E_n$ that are disjoint; that is, every outcome is in at most one event E_i. The ***sum rule*** says that the probability of the union of these events is equal to the sum of their probabilities:

$$\Pr\{E_1 \cup \cdots \cup E_n\} = \Pr\{E_1\} + \cdots + \Pr\{E_n\}.$$

The probability of the union of events that are not necessarily disjoint is given by an ***inclusion-exclusion formula*** analogous to the one for set sizes:

$$\Pr\{E_1 \cup \cdots \cup E_n\} = \sum_i \Pr\{E_i\} - \sum_{i,j} \Pr\{E_i \cap E_j\} + \sum_{i,j,k} \Pr\{E_i \cap E_j \cap E_k\} - \cdots .$$

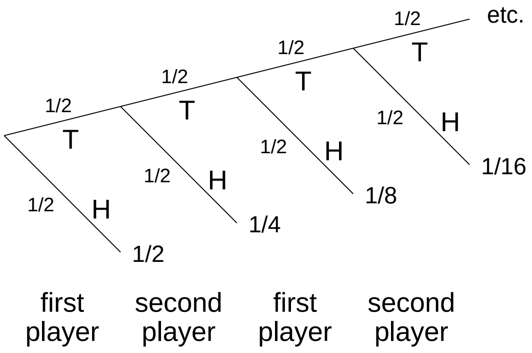
The following inequality, called the ***union bound***, also holds even if events $E_1, \ldots, E_n$ are not disjoint:

$$\Pr\{E_1 \cup \cdots \cup E_n\} \leq \Pr\{E_1\} + \cdots + \Pr\{E_n\}.$$

The union bound is simple and "good enough" for many probability calculations. For example, suppose that E_i is the event that the i-th critical component in a spacecraft fails. Then $E_1 \cup \cdots \cup E_n$ is the event that *some* critical component fails. The union bound gives an upper bound on this vital probability and does not require engineers to estimate all the terms in the gigantic inclusion-exclusion formula.

2 Infinite Sample Spaces

Suppose two players take turns flipping a fair coin. Whoever flips heads first is declared the winner. What is the probability that the first player wins? A tree diagram for this problem is shown below:



The event that the first player wins contains an infinite number of outcomes, but we can still sum their probabilities:

$$\begin{aligned}\Pr\{\text{first player wins}\} &= \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \cdots \\ &= \frac{2}{3}.\end{aligned}$$

The second step uses the formula for the sum of a geometric series. Similarly, we can compute the probability that the second player wins:

$$\begin{aligned}\Pr\{\text{second player wins}\} &= \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \cdots \\ &= \frac{1}{3}.\end{aligned}$$

In principle, the game could go on forever if both players kept flipping tails. In our tree diagram, this situation does not correspond to any leaf —rather, it corresponds to the infinite path. So, this is not an outcome. If we wanted to consider it as such, we could add an extra leaf as a child of the root. The probability on the edge to this leaf should then be the probability of the infinite path

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdots = \lim_{n \to \infty} \left(\frac{1}{2}\right)^n = 0.$$

Since this probability is 0, there would be no impact on our calculations.

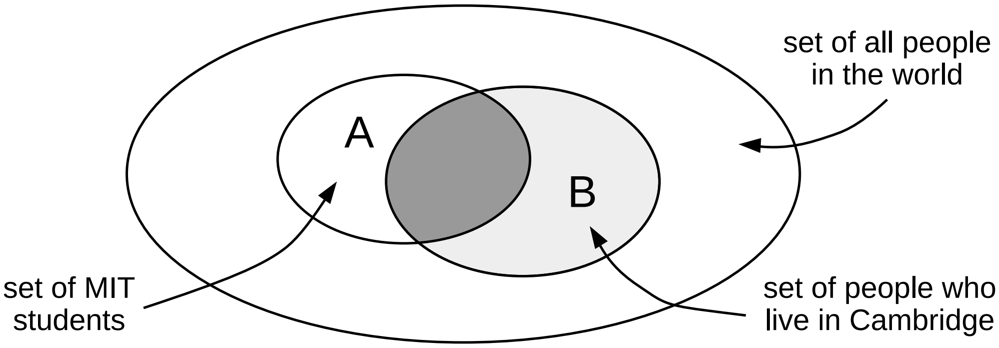
The mathematical machinery we've developed is adequate to model and analyze many interesting probability problems with infinite sample spaces. However, some intricate infinite processes require more powerful (and more complex) measure-theoretic notions of probability. For example, if we generate an infinite sequence of random bits $b_1, b_2, b_3, \ldots$, then what is the probability that

$$\frac{b_1}{2^1} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots$$

is a rational number? We won't take up such problems in this course.

3 Conditional Probability

Suppose that we pick a random person in the world. Everyone has an equal chance of being selected. Let A be the event that the person is an MIT student, and let B be the event that the person lives in Cambridge. What are the probabilities of these events? Intuitively, we're picking a random point in the big ellipse shown below and asking how likely that point is to fall into region A or B:


set of all people in the world
A
B
set of MIT students
set of people who live in Cambridge

The vast majority of people in the world neither live in Cambridge nor are MIT students, so events A and B both have low probability. But what is the probability that a person is an MIT student, *given* that the person lives in Cambridge? This should be much greater— but what it is exactly?

What we're asking for is called a ***conditional probability***; that is, the probability that one event happens, given that some other event definitely happens. Questions about conditional probabilities come up all the time:

- What is the probability that it will rain this afternoon, given that it is cloudy this morning?
- What is the probability that two rolled dice sum to 10, given that both are odd?
- What is the probability that I'll get four-of-a-kind in Texas No Limit Hold 'Em Poker, given that I'm initially dealt two queens?

There is a special notation for conditional probabilities. In general, $\Pr\{A \mid B\}$ denotes the probability of event A, given that event B happens. So, in our example, $\Pr\{A \mid B\}$ is the probability that a random person is an MIT student, given that he or she is a Cambridge resident.

How do we compute $\Pr\{A \mid B\}$? Since we are *given* that the person lives in Cambridge, we can forget about everyone in the world who does not. Thus, all outcomes outside event B are irrelevant. So, intuitively, $\Pr\{A \mid B\}$ should be the fraction of Cambridge residents that are also MIT students; that is, the answer should be the probability that the person is in set $A \cap B$ (darkly shaded) divided by the probability that the person is in set B (lightly shaded). This motivates the definition of conditional probability:

$$\Pr\{A \mid B\} ::= \frac{\Pr\{A \cap B\}}{\Pr\{B\}}$$

If $\Pr\{B\} = 0$, then the conditional probability $\Pr\{A \mid B\}$ is undefined.

Probability is generally counterintuitive, but conditional probability is the worst! Conditioning can subtly alter probabilities and produce unexpected results in randomized algorithms and computer systems as well as in betting games. Yet, the mathematical definition of conditional probability given above is very simple and should give you no trouble— provided you rely on formal reasoning and not intuition.

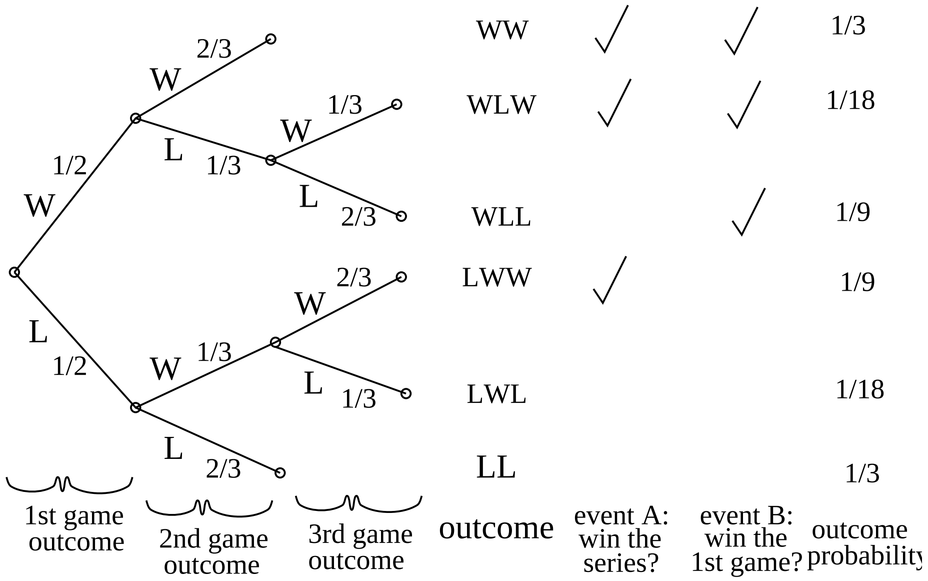
3.1 The Halting Problem

The *Halting Problem* is the canonical undecidable problem in computation theory that was first introduced by Alan Turing in his seminal 1936 paper. The problem is to determine whether a Turing machine halts on a given blah, blah, blah. Anyway, *much more importantly*, it is the name of the MIT EECS department's famed C-league hockey team.

In a best-of-three tournament, the Halting Problem wins the first game with probability $1/2$. In subsequent games, their probability of winning is determined by the outcome of the previous game. If the Halting Problem won the previous game, then they are invigorated by victory and win the current game with probability $2/3$. If they lost the previous game, then they are demoralized by defeat and win the current game with probablity only $1/3$. What is the probability that the Halting Problem wins the tournament, given that they win the first game?

This is a question about a conditional probability. Let A be the event that the Halting Problem wins the tournament, and let B be the event that they win the first game. Our goal is then to determine the conditional probability $\Pr\{A \mid B\}$.

We can tackle conditional probability questions just like ordinary probability problems: using a tree diagram and the four-step method. A complete tree diagram is shown below, followed by an explanation of its construction and use.



Step 1: Find the Sample Space

Each internal vertex in the tree diagram has two children, one corresponding to a win for the Halting Problem (labeled W) and one corresponding to a loss (labeled L). The complete sample space is:

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

Step 2: Define Events of Interest

The event that the Halting Problem wins the whole tournament is:

$$T = \{WW,\ WLW,\ LWW\}$$

And the event that the Halting Problem wins the first game is:

$$F = \{WW, WLW, WLL\}$$

The outcomes in these events are indicated with checkmarks in the tree diagram.

Step 3: Determine Outcome Probabilities

Next, we must assign a probability to each outcome. We begin by labeling edges as specified in the problem statement. Specifically, The Halting Problem has a $1/2$ chance of winning the first game, so the two edges leaving the root are each assigned probability $1/2$. Other edges are labeled $1/3$ or $2/3$ based on the outcome of the preceding game. We then find the probability of each outcome by multiplying all probabilities along the corresponding root-to-leaf path. For example, the probability of outcome WLL is:

$$\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{9}$$

Step 4: Compute Event Probabilities

We can now compute the probability that The Halting Problem wins the tournament, given that they win the first game:

$$\begin{aligned}
\Pr\{A \mid B\} &= \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \\
&= \frac{\Pr\{\{WW, WLW\}\}}{\Pr\{\{WW, WLW, WLL\}\}} \\
&= \frac{1/3 + 1/18}{1/3 + 1/18 + 1/9} \\
&= \frac{7}{9}
\end{aligned}$$

We're done! If the Halting Problem wins the first game, then they win the whole tournament with probability $7/9$.

3.2 Why Tree Diagrams Work

We've now settled into a routine of solving probability problems using tree diagrams. But we've left a big question unaddressed: what is the mathematical justification behind those funny little pictures? Why do they work?

The answer involves conditional probabilities. In fact, the probabilities that we've been recording on the edges of tree diagrams *are* conditional probabilities. For example, consider the uppermost path in the tree diagram for the Halting Problem, which corresponds to the outcome WW. The first edge is labeled $1/2$, which is the probability that the Halting Problem wins the first game. The second edge is labeled $2/3$, which is the probability that the Halting Problem wins the second game, *given* that they won the first— that's a conditional probability! More generally, on each edge of a tree diagram, we record the probability that the experiment proceeds along that path, given that it reaches the parent vertex.

So we've been using conditional probabilities all along. But why can we multiply edge probabilities to get outcome probabilities? For example, we concluded that:

$$\begin{aligned}\Pr\{WW\} &= \frac{1}{2}\cdot\frac{2}{3}\\ &= \frac{1}{3}\end{aligned}$$

Why is this correct?

The answer goes back to the definition of conditional probability. Rewriting this in a slightly different form gives the ***Product Rule*** for probabilities:

Definition 3.1 (Product Rule for 2 Events). If $\Pr\{A_2\} \neq 0$, then:

$$\Pr\{A_1 \cap A_2\} = \Pr\{A_1\} \cdot \Pr\{A_2 \mid A_1\}$$

Multiplying edge probabilities in a tree diagram amounts to evaluating the right side of this equation. For example:

$$\begin{aligned}&\Pr\{\text{win first game} \cap \text{win second game}\}\\ &\qquad = \Pr\{\text{win first game}\} \cdot \Pr\{\text{win second game} \mid \text{win first game}\}\\ &\qquad = \frac{1}{2}\cdot\frac{2}{3}\end{aligned}$$

So the Product Rule is the formal justification for multiplying edge probabilities to get outcome probabilities!

To justify multiplying edge probabilities along longer paths, we need a more general form the Product Rule:

Definition 3.2 (Product Rule for n Events). If $\Pr\{A_1 \cap \ldots \cap A_{n-1}\} \neq 0$, then:

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

Let's interpret this big formula in terms of tree diagrams. Suppose we want to compute the probability that an experiment traverses a particular root-to-leaf path of length n. Let A_i be the event that the experiment traverses the i-th edge of the path. Then $A_1 \cap \ldots \cap A_n$ is the event that the experiment traverse the whole path. The Product Rule says that the probability of this is the probability that the experiment takes the first edge times the probability that it takes the second, *given* it takes the first edge, times the probability it takes the third, *given* it takes the first two edges, and so forth. In other words, the probability of an outcome is the product of the edge probabilities along the corresponding root-to-leaf path.

3.3 The Law of Total Probability

The following identity

$$\Pr\{A\} = \Pr\{A \mid E\} \cdot \Pr\{E\} + \Pr\left\{A \mid \overline{E}\right\} \cdot \Pr\left\{\overline{E}\right\}.$$

is called the Law of Total Probability and lets you compute the probability of an event A using case analysis based on whether or not event E occurs. For example, suppose we conduct the following experiment. First, we flip a coin. If heads comes up, then we roll one die and take the result. If tails comes up, then we roll two dice and take the sum of the two results. What is the probability that this process yields a 2? Let E be the event that the coin comes up heads, and let A be the event that we get a 2 overall. Assuming that the coin is fair, $\Pr\{E\} = \Pr\left\{\overline{E}\right\} = 1/2$. There are now two cases. If we flip heads, then we roll a 2 on a single die with probabilty $\Pr\{A \mid E\} = 1/6$. On the other hand, if we flip tails, then we get a sum of 2 on two dice with probability $\Pr\left\{A \mid \overline{E}\right\} = 1/36$. Therefore, the probability that the whole process yields a 2 is

$$\Pr\{A\} = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{36} = \frac{7}{72}.$$

More generally, if $E_1, \ldots, E_n$ are disjoint events whose union is the whole sample space, then:

$$\Pr\{A\} = \sum_{i=1}^{n} \Pr\{A \mid E_i\} \cdot \Pr\{E_i\}.$$

3.4 *A Posteriori* Probabilities

Suppose that we turn the hockey question around: what is the probability that the Halting Problem won their first game, given that they won the series?

This seems like an absurd question! After all, if the Halting Problem won the series, then the winner of the first game has already been determined. Therefore, who won the first game is a question of fact, not a question of probability. However, our mathematical theory of probability contains no notion of one event preceding another— there is no notion of time at all. Therefore, from a mathematical perspective, this is a perfectly valid question. And this is also a meaningful question from a practical perspective. Suppose that you're told that the Halting Problem won the series, but not told the results of individual games. Then, from your perspective, it makes perfect sense to wonder how likely it is that The Halting Problem won the first game.

A conditional probability $\Pr\{B \mid A\}$ is called an ***a posteriori*** if event B precedes event A in time. Here are some other examples of a posteriori probabilities:

- The probability it was cloudy this morning, given that it rained in the afternoon.

- The probability that I was initially dealt two queens in Texas No Limit Hold 'Em poker, given that I eventually got four-of-a-kind.

Mathematically, a posteriori probabilities are *no different* from ordinary probabilities; the distinction is only at a higher, philosophical level. Our only reason for drawing attention to them is to say, "Don't let them rattle you."

Let's return to the original problem. The probability that the Halting Problem won their first game, given that they won the series is $\Pr\{B \mid A\}$. We can compute this using the definition of conditional probability and our earlier tree diagram:

$$\begin{aligned}\Pr\{B \mid A\} &= \frac{\Pr\{B \cap A\}}{\Pr\{A\}} \\ &= \frac{1/3 + 1/18}{1/3 + 1/18 + 1/9} \\ &= \frac{7}{9}\end{aligned}$$

This answer is suspicious! In the preceding section, we showed that $\Pr\{A \mid B\}$ was also $7/9$. Could it be true that $\Pr\{A \mid B\} = \Pr\{B \mid A\}$ in general? Some reflection suggests this is unlikely. For example, the probability that I feel uneasy, given that I was abducted by aliens, is pretty large. But the probability that I was abducted by aliens, given that I feel uneasy, is rather small.

Let's work out the general conditions under which $\Pr\{A \mid B\} = \Pr\{B \mid A\}$. By the definition of conditional probability, this equation holds if an only if:

$$\frac{\Pr\{A \cap B\}}{\Pr\{B\}} = \frac{\Pr\{A \cap B\}}{\Pr\{A\}}$$

This equation, in turn, holds only if the denominators are equal or the numerator is 0:

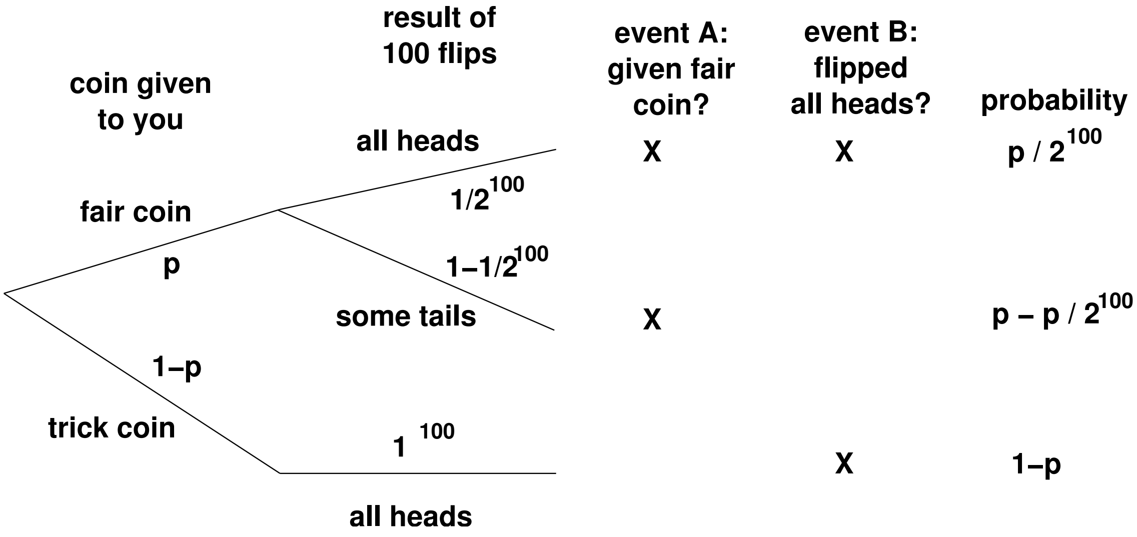
$$\Pr\{B\} = \Pr\{A\} \qquad \text{or} \qquad \Pr\{A \cap B\} = 0$$

The former condition holds in the hockey example; the probability that the Halting Problem wins the series (event A) is equal to the probability that it wins the first game (event B). In fact, both probabilities are $1/2$.

3.5 A Coin Problem

Someone hands you either a fair coin or a trick coin with heads on both sides. You flip the coin 100 times and see heads every time. What can you say about the probability that you flipped the fair coin? Remarkably— nothing!

In order to make sense out of this outrageous claim, let's formalize the problem. The sample space is worked out in the tree diagram below. We do not know the probability that you were handed the fair coin initially— you were just given one coin or the other— so let's call that p.



Let A be the event that you were handed the fair coin, and let B be the event that you flipped 100 heads. Now, we're looking for $\Pr\{A \mid B\}$, the probability that you were handed the fair coin, given that you flipped 100 heads. The outcome probabilities are worked out in the tree diagram. Plugging the results into the definition of conditional probability gives:

$$\begin{aligned}\Pr\{A \mid B\} &= \frac{\Pr\{A \cap B\}}{\Pr\{B\}}\\ &= \frac{p/2^{100}}{1-p+p/2^{100}}\\ &= \frac{p}{2^{100}(1-p)+p}\end{aligned}$$

This expression is very small for moderate values of p because of the 2^{100} term in the denominator. For example, if $p = 1/2$, then the probability that you were given the fair coin is essentially zero.

But we *do not know* the probability p that you were given the fair coin. And perhaps the value of p is *not* moderate; in fact, maybe $p = 1 - 2^{-100}$. Then there is nearly an even chance that you have the fair coin, given that you flipped 100 heads. In fact, maybe you were handed the fair coin with probability $p = 1$. Then the probability that you were given the fair coin is, well, 1!

A similar problem arises in polling before an election. A pollster picks a random American and asks his or her party affiliation. If this process is repeated many times, what can be said about the population as a whole? To clarify the analogy, suppose that the country contains only two people. There is either one Republican and one Democrat (like the fair coin), or there are two Republicans (like the trick coin). The pollster picks a random citizen 100 times, which is analogous to flipping the coin 100 times. Suppose that he picks a Republican every single time. However, even given this polling data, the probability that there is one citizen in each party could still be anywhere between 0 and 1!

What the pollster *can* say is that either:

1. Something earth-shatteringly unlikely happened during the poll.
2. There are two Republicans.

This is as far as probability theory can take us; from here, you must draw your own conclusions. Based on life experience, many people would consider the second possibility more plausible. However, if you are just *convinced* that the country isn't entirely Republican (say, because you're a citizen and a Democrat), then you might believe that the first possibility is actually more likely.

3.6 Medical Testing

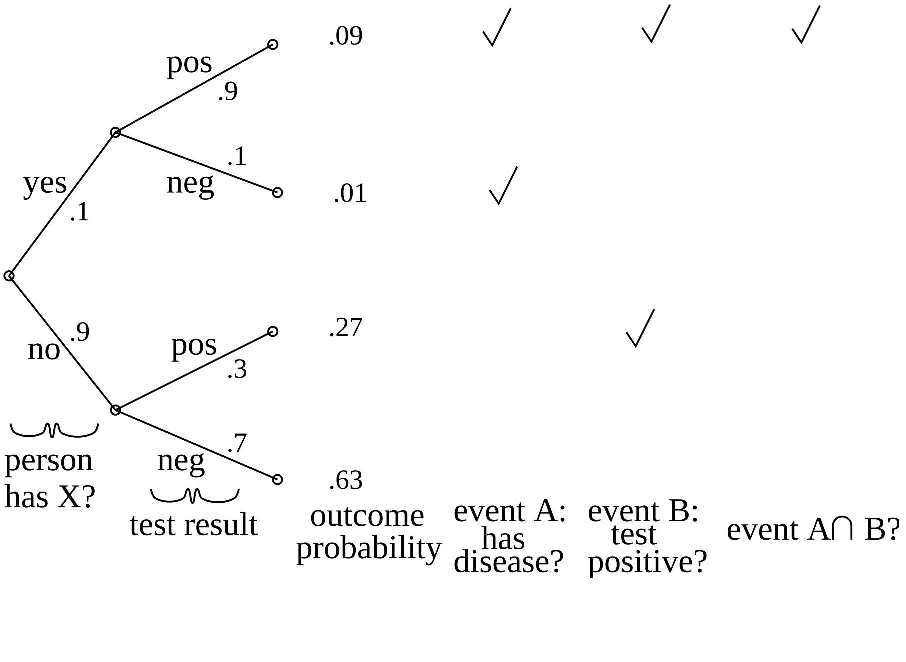
There is a deadly disease called X that has infected 10% of the population. There are no symptoms; victims just drop dead one day. Fortunately, there is a test for the disease. The test is not perfect, however:

- If you have the disease, there is a 10% chance that the test will say you do not. (These are called "false negatives".)
- If you do not have disease, there is a 30% chance that the test will say you do. (These are "false positives".)

A random person is tested for the disease. If the test is positive, then what is the probability that the person has the disease?

Step 1: Find the Sample Space

The sample space is found with the tree diagram below.



Step 2: Define Events of Interest

Let A be the event that the person has the disease. Let B be the event that the test was positive. The outcomes in each event are marked in the tree diagram. We want to find $\Pr\{A \mid B\}$, the probability that a person has disease X, given that the test was positive.

Step 3: Find Outcome Probabilities

First, we assign probabilities to edges. These probabilities are drawn directly from the problem statement. By the Product Rule, the probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path. All probabilities are shown in the figure.

Step 4: Compute Event Probabilities

$$\begin{aligned}\Pr\{A \mid B\} &= \frac{\Pr\{A \cap B\}}{\Pr\{B\}}\\ &= \frac{0.09}{0.09 + 0.27}\\ &= \frac{1}{4}\end{aligned}$$

If you test positive, then there is only a 25% chance that you have the disease!

This answer is initially surprising, but makes sense on reflection. There are two ways you could test positive. First, it could be that you are sick and the test is correct. Second, it could be that you are healthy and the test is incorrect. The problem is that almost everyone is healthy; therefore, most of the positive results arise from incorrect tests of healthy people!

We can also compute the probability that the test is correct for a random person. This event consists of two outcomes. The person could be sick and the test positive (probability 0.09), or the person could be healthy and the test negative (probability 0.63). Therefore, the test is correct with probability $0.09 + 0.63 = 0.72$. This is a relief; the test is correct almost three-quarters of the time.

But wait! There is a simple way to make the test correct 90% of the time: always return a negative result! This "test" gives the right answer for all healthy people and the wrong answer only for the 10% that actually have the disease. The best strategy is to completely ignore the test result!

There is a similar paradox in weather forecasting. During winter, almost all days in Boston are wet and overcast. Predicting miserable weather every day may be more accurate than really trying to get it right!

3.7 Other Identities

There is a close relationship between computing the size of a set and computing the probability of an event. The inclusion-exclusion formula is one such example; the probability of a union of events and the cardinality of a union of sets are computed using similar formulas.

In fact, all of the methods we developed for computing sizes of sets carry over to computing probabilities. This is because a probability space is just a weighted set; the sample space is the set and the probability function assigns a weight to each element. Earlier, we were counting the number of items in a set. Now, when we compute the probability of an event, we are just summing the weights of items. We'll see many examples of the close relationship between probability and counting over the next few weeks.

Many general probability identities still hold when all probabilities are conditioned on the same event. For example, the following identity is analogous to the Inclusion-Exclusion formula for two sets, except that all probabilities are conditioned on an event C.

$$\Pr\{A \cup B \mid C\} = \Pr\{A \mid C\} + \Pr\{B \mid C\} - \Pr\{A \cap B \mid C\}.$$

As a special case we have

$$\Pr\{A \cup B \mid C\} = \Pr\{A \mid C\} + \Pr\{B \mid C\} \qquad \text{when } A \cap B = \emptyset.$$

Be careful not to mix up events before and after the conditioning bar! For example, the following is *not* a valid identity:

False Claim.

$$\Pr\{A \mid B \cup C\} = \Pr\{A \mid B\} + \Pr\{A \mid C\} \qquad \textit{when } B \cap C = \emptyset. \tag{1}$$

4 Independence

Suppose that we flip two fair coins simultaneously on opposite sides of a room. Intuitively, the way one coin lands does not affect the way the other coin lands. The mathematical concept that captures this intuition is called ***independence***:

Definition. Events A and B are independent if and only if:

$$\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$$

Generally, independence is something you *assume* in modeling a phenomenon— or wish you could realistically assume. Many useful probability formulas only hold if certain events are independent, so a dash of independence can greatly simplify the analysis of a system.

4.1 Examples

Let's return to the experiment of flipping two fair coins. Let A be the event that the first coin comes up heads, and let B be the event that the second coin is heads. If we assume that A and B are independent, then the probability that both coins come up heads is:

$$\begin{aligned}\Pr\{A \cap B\} &= \Pr\{A\} \cdot \Pr\{B\} \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{4}\end{aligned}$$

On the other hand, let C be the event that tomorrow is cloudy and R be the event that tomorrow is rainy. Perhaps $\Pr\{C\} = 1/5$ and $\Pr\{R\} = 1/10$ around here. If these events were independent, then we could conclude that the probability of a rainy, cloudy day was quite small:

$$\begin{aligned}\Pr\{R \cap C\} &= \Pr\{R\} \cdot \Pr\{C\} \\ &= \frac{1}{5} \cdot \frac{1}{10} \\ &= \frac{1}{50}\end{aligned}$$

Unfortunately, these events are definitely not independent; in particular, every rainy day is cloudy. Thus, the probability of a rainy, cloudy day is actually $1/10$.

4.2 Working with Independence

There is another way to think about independence that you may find more intuitive. According to the definition, events A and B are independent if and only if $\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$. This equation holds even if $\Pr\{B\} = 0$, but assuming it is not, we can divide both sides by $\Pr\{B\}$ and use the definition of conditional probability to obtain an alternative formulation of independence:

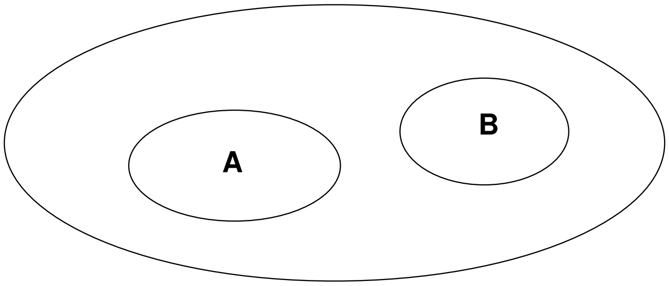
Proposition. *If* $\Pr\{B\} \neq 0$, *then events* A *and* B *are independent if and only if*

$$\Pr\{A \mid B\} = \Pr\{A\}. \tag{2}$$

Equation (2) says that events A and B are independent if the probability of A is unaffected by the fact that B happens. In these terms, the two coin tosses of the previous section were independent, because the probability that one coin comes up heads is unaffected by the fact that the other came up heads. Turning to our other example, the probability of clouds in the sky is strongly affected by the fact that it is raining. So, as we noted before, these events are not independent.

4.3 Some Intuition

Suppose that A and B are disjoint events, as shown in the figure below.



Are these events independent? Let's check. On one hand, we know

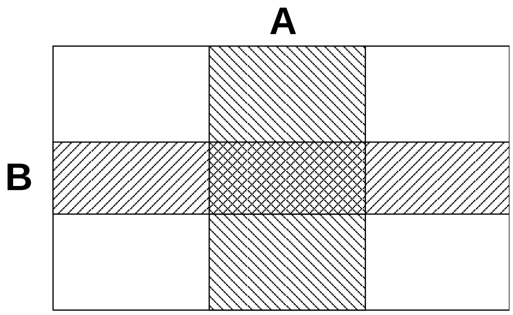
$$\Pr\{A \cap B\} = 0$$

because $A \cap B$ contains no outcomes. On the other hand, we have

$$\Pr\{A\} \cdot \Pr\{B\} > 0$$

except in degenerate cases where A or B has zero probability. Thus, *disjointness and independence are very different ideas*.

Here's a better mental picture of what independent events look like.



The sample space is the whole rectangle. Event A is a vertical stripe, and event B is a horizontal stripe. Assume that the probability of each event is proportional to its area in the diagram. Now if A covers an α-fraction of the sample space, and B covers a β-fraction, then the area of the intersection region is $\alpha \cdot \beta$. In terms of probability:

$$\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$$

5 Mutual Independence

We have defined what it means for two events to be independent. But how can we talk about independence when there are more than two events? For example, how can we say that the orientations of n coins are all independent of one another?

Events $E_1, \ldots, E_n$ are ***mutually independent*** if and only if *for every subset* of the events, the probability of the intersection is the product of the probabilities. In other words, all of the following equations must hold:

$$\begin{aligned}
\Pr\{E_i \cap E_j\} &= \Pr\{E_i\} \cdot \Pr\{E_j\} && \text{for all distinct } i, j \\
\Pr\{E_i \cap E_j \cap E_k\} &= \Pr\{E_i\} \cdot \Pr\{E_j\} \cdot \Pr\{E_k\} && \text{for all distinct } i, j, k \\
\Pr\{E_i \cap E_j \cap E_k \cap E_l\} &= \Pr\{E_i\} \cdot \Pr\{E_j\} \cdot \Pr\{E_k\} \cdot \Pr\{E_l\} && \text{for all distinct } i, j, k, l \\
&\ldots \\
\Pr\{E_1 \cap \cdots \cap E_n\} &= \Pr\{E_1\} \cdots \Pr\{E_n\}
\end{aligned}$$

As an example, if we toss 100 fair coins and let E_i be the event that the ith coin lands heads, then we might reasonably assume than $E_1, \ldots, E_{100}$ are mutually independent.

5.1 DNA Testing

This is testimony from the O. J. Simpson murder trial on May 15, 1995:

> **MR. CLARKE:** When you make these estimations of frequency— and I believe you touched a little bit on a concept called independence?
>
> **DR. COTTON:** Yes, I did.
>
> **MR. CLARKE:** And what is that again?
>
> **DR. COTTON:** It means whether or not you inherit one allele that you have is not— does not affect the second allele that you might get. That is, if you inherit a band at 5,000 base pairs, that doesn't mean you'll automatically or with some probability inherit one at 6,000. What you inherit from one parent is what you inherit from the other. *(Got that? – EAL)*
>
> **MR. CLARKE:** Why is that important?
>
> **DR. COTTON:** Mathematically that's important because if that were not the case, it would be improper to multiply the frequencies between the different genetic locations.
>
> **MR. CLARKE:** How do you— well, first of all, are these markers independent that you've described in your testing in this case?

The jury was told that genetic markers in blood found at the crime scene matched Simpson's. Furthermore, the probability that the markers would be found in a randomly-selected person was at most 1 in 170 million. This astronomical figure was derived from statistics such as:

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

Then these numbers were multiplied to give the probability that a randomly-selected person would have all five markers:

$$\begin{aligned}\Pr\{A \cap B \cap C \cap D \cap E\} &= \Pr\{A\} \cdot \Pr\{B\} \cdot \Pr\{C\} \cdot \Pr\{D\} \cdot \Pr\{E\} \\ &= \frac{1}{100} \cdot \frac{1}{50} \cdot \frac{1}{40} \cdot \frac{1}{5} \cdot \frac{1}{170} \\ &= \frac{1}{170,000,000}\end{aligned}$$

The defense pointed out that this assumes that the markers appear mutually independently. Furthermore, all the statistics were based on just a few hundred blood samples. The jury was widely mocked for failing to "understand" the DNA evidence. If you were a juror, would *you* accept the 1 in 170 million calculation?

5.2 Pairwise Independence

The definition of mutual independence seems awfully complicated— there are so many conditions! Here's an example that illustrates the subtlety of independence when more than two events are involved and the need for all those conditions. Suppose that we flip three fair, mutually-independent coins. Define the following events:

- A_1 is the event that coin 1 matches coin 2.
- A_2 is the event that coin 2 matches coin 3.
- A_3 is the event that coin 3 matches coin 1.

Are A_1, A_2, A_3 mutually independent?

The sample space for this experiment is:

$$\{HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT\}$$

Every outcome has probability $(1/2)^3 = 1/8$ by our assumption that the coins are mutually independent.

To see if events A_1, A_2, and A_3 are mutually independent, we must check a sequence of equalities. It will be helpful first to compute the probability of each event A_i:

$$\begin{aligned}
\Pr\{A_1\} &= \Pr\{HHH\} + \Pr\{HHT\} + \Pr\{TTH\} + \Pr\{TTT\} \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{2}
\end{aligned}$$

By symmetry, $\Pr\{A_2\} = \Pr\{A_3\} = 1/2$ as well. Now we can begin checking all the equalities required for mutual independence.

$$\begin{aligned}
\Pr\{A_1 \cap A_2\} &= \Pr\{HHH\} + \Pr\{TTT\} \\
&= \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{4} \\
&= \frac{1}{2} \cdot \frac{1}{2} \\
&= \Pr\{A_1\} \Pr\{A_2\}
\end{aligned}$$

By symmetry, $\Pr\{A_1 \cap A_3\} = \Pr\{A_1\} \cdot \Pr\{A_3\}$ and $\Pr\{A_2 \cap A_3\} = \Pr\{A_2\} \cdot \Pr\{A_3\}$ must hold also. Finally, we must check one last condition:

$$\begin{aligned}
\Pr\{A_1 \cap A_2 \cap A_3\} &= \Pr\{HHH\} + \Pr\{TTT\} \\
&= \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{4} \\
&\neq \Pr\{A_1\} \Pr\{A_2\} \Pr\{A_3\} = \frac{1}{8}
\end{aligned}$$

The three events A_1, A_2, and A_3 are not mutually independent, even though all *pairs* of events are independent!

A set of events in ***pairwise independent*** if every pair is independent. Pairwise independence is a much weaker property than mutual independence. For example, suppose that the prosecutors in the O. J. Simpson trial were wrong and markers A, B, C, D, and E appear only *pairwise* independently. Then the probability that a randomly-selected person has all five markers is no more than:

$$\begin{aligned}
\Pr\{A \cap B \cap C \cap D \cap E\} &\leq \Pr\{A \cap E\} \\
&= \Pr\{A\} \cdot \Pr\{E\} \\
&= \frac{1}{100} \cdot \frac{1}{170} \\
&= \frac{1}{17,000}
\end{aligned}$$

The first line uses the fact that $A \cap B \cap C \cap D \cap E$ is a subset of $A \cap E$. (We picked out the A and E markers because they're the rarest.) We use pairwise independence on the second line. Now the probability of a random match is 1 in 17,000— a far cry from 1 in 170 million! And this is the strongest conclusion we can reach assuming only pairwise independence.