Math 3332: Probability and Inference

Unit I: Random Experiments and Sampling

Mikhail Lavrov (mlavrov@kennesaw.edu)

January 11–February 17, 2021

Sets

In math, when we want to consider a bunch of things all at once, we call this a **set**. For example,

$$\{\text{A}\heartsuit, \ldots, \text{K}\heartsuit, \text{A}\clubsuit, \ldots, \text{K}\clubsuit, \text{A}\diamondsuit, \ldots, \text{K}\diamondsuit, \text{A}\spadesuit, \ldots, \text{K}\spadesuit\}$$

is the set of cards in a 52-card deck.

The individual things are called **elements** of the set. For example, Q$\spadesuit$ and 2$\clubsuit$ are elements of the set above. We write $x \in S$ to say that x is an element of S, and $x \notin S$ to say that it isn't.

We say that one set A is a subset of another set B if all elements of A are also elements of B. This is denoted $A \subseteq B$ or $A \subset B$. For example, $\{7\heartsuit, 7\clubsuit, 7\diamondsuit, 7\spadesuit\}$ is a subset of the set above.

Common sets

Some common sets that we will need to know about:

- The set of **natural numbers** $\mathbb{N} = \{1, 2, 3, \dots\}$.
- The **integers** $\mathbb{Z}$, **rational numbers** $\mathbb{Q}$, **real numbers** $\mathbb{R}$, and complex numbers $\mathbb{C}$.
- Intervals of real numbers:
  - $[0, 1]$, the set of all real numbers x such that $0 \leq x \leq 1$.
  - $(1, 2)$, the set of all real numbers x such that $1 < x < 2$.
  - $[2, \infty)$, the set of all real numbers x such that $2 \leq x$.
- The **empty set** $\varnothing = \{\}$ containing no elements at all.

Set-builder notation

To describe new sets, we often write

$$\{x : x \text{ has property } P\} \quad \text{or} \quad \{x \in S : x \text{ has property } P\}$$

for the set of everything that has property P, or the set of all elements of S which have property P.

Examples:

- $\mathbb{N} = \{x \in \mathbb{Z} : x > 0\}$.
- $(-1, 1] = \{x \in \mathbb{R} : -1 < x \leq 1\}$.
- $\mathbb{Q} = \{\frac{a}{b} : a, b \in \mathbb{Z} \text{ and } b \neq 0\}$.

The three sizes of sets

We distinguish between the following three kinds of sets:

1. Finite sets, such as

$$\{\text{A}\heartsuit, \ldots, \text{K}\heartsuit, \text{A}\clubsuit, \ldots, \text{K}\clubsuit, \text{A}\diamondsuit, \ldots, \text{K}\diamondsuit, \text{A}\spadesuit, \ldots, \text{K}\spadesuit\}.$$

2. Countably infinite sets, such as

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$$

3. Uncountably infinite sets, such as $\mathbb{R}$, the set of all real numbers.

Finite sets

A set is **finite** if we can number its elements $1, 2, \ldots, n$ for some natural number n.

For example,

$$\{\mathsf{A}\heartsuit, \ldots, \mathsf{K}\heartsuit, \mathsf{A}\clubsuit, \ldots, \mathsf{K}\clubsuit, \mathsf{A}\diamondsuit, \ldots, \mathsf{K}\diamondsuit, \mathsf{A}\spadesuit, \ldots, \mathsf{K}\spadesuit\}.$$

is finite because we can number $\mathsf{A}\heartsuit = 1$, $\ldots$, $\mathsf{K}\spadesuit = 52$.

Probability connection: We can pick from a finite set uniformly at random, choosing each element with probability $\frac{1}{n}$.

Countably infinite sets

A set is **countably infinite** if we can write its elements as an infinite list. For example:

- We can list the elements of $\mathbb{Z}$ one by one as

$$0, 1, -1, 2, -2, 3, -3, 4, -4, \ldots$$

- We can list the elements of $\mathbb{Q}$ one by one as

$$\frac{0}{1}, \frac{1}{1}, -\frac{1}{1}, \frac{2}{1}, -\frac{2}{1}, \frac{1}{2}, -\frac{1}{2}, \ldots$$

  going in order of $|\text{numerator}| + |\text{denominator}|$.

Probability connection: We cannot pick from an infinite set uniformly. But we can still assign each element a probability.

Uncountably infinite sets

A set is **uncountably infinite** if it's so big we can't even list out all the elements.

For example, $\mathbb{R}$ (the set of real numbers), or even an interval of real numbers such as $[0, 1]$, is an uncountably infinite set.

Probability connection: We can't pick a random element from an uncountable set by giving each element a positive probability of being chosen.

We will talk about choosing random elements from uncountable sets later in the class, but we'll need special tools to do this.

Unions, intersections, and differences

Given two sets A and B:

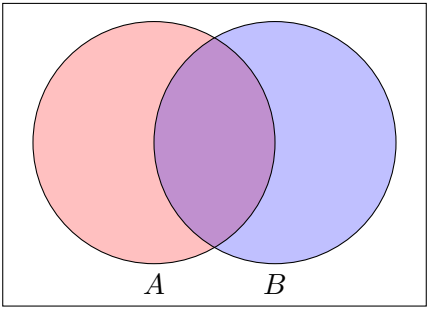
- The **union** $A \cup B$ is the set of all elements in A **or** B.
- The **intersection** $A \cap B$ is the set of all elements in A **and** B.
- The **difference** $A - B$ is the set of all elements in A **but not** B.

For example, if $A = [1, 3)$ and $B = [2, 4)$, then:

$$A \cup B = [1, 4) \quad A \cap B = [2, 3) \quad A - B = [1, 2)$$

Visualizing set operations

A handy way to visualize set operations with arbitrary sets is by using Venn diagrams.



The union $A \cup B$ is the entire shaded area. The intersection $A \cup B$ is purple. The red (but not purple) is $A - B$.

Complements

We often assume that all the elements of our sets come from some "universal set" S. For example, if we're solving a problem about real numbers, we might take the universal set to be $\mathbb{R}$.

In this case, we write A^c (or $\overline{A}$) as shorthand for the difference $S - A$. This is called the **complement** of A: the set of all elements **not** in A.

Example: if the universal set is $\mathbb{Z}$, then the complement of $\mathbb{N}$ is

$$\mathbb{N}^c = \{0, -1, -2, -3, \dots\} = \{x \in \mathbb{Z} : x \leq 0\}.$$

Example: if the universal set is $\{\mathsf{A}\heartsuit, \dots, \mathsf{K}\heartsuit\}$, then the complement of $\{\mathsf{A}\heartsuit, \mathsf{J}\heartsuit, \mathsf{Q}\heartsuit, \mathsf{K}\heartsuit\}$ is

$$\{2\heartsuit, 3\heartsuit, 4\heartsuit, \dots, 9\heartsuit, 10\heartsuit\}.$$

Cardinalities

We write $|A|$ for the number of elements in A.

(We'll mostly consider the case where A is finite; if A is infinite, we just write $|A| = \infty$, though sometimes people make further distinctions.)

This interacts with previous set operations via the **inclusion-exclusion principle**:

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

Intuition: if we count the elements in A, then count the elements in B, we've counted the elements in $A \cap B$ twice, and have to fix this by subtracting.

There are more complicated cases of the inclusion-exclusion principle. For example, we can write an expression for $|A \cup B \cup C|$ in terms of intersections.

Products

Given two sets A and B, the **Cartesian product** $A \times B$ is the set of all pairs (a, b) where $a \in A$ and $b \in B$. For example, we can think of the set of cards in a 52-card deck as the Cartesian product

$$\{\mathsf{A}, 2, 3, 4, 5, 6, 7, 8, 9, 10, \mathsf{J}, \mathsf{Q}, \mathsf{K}\} \times \{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\}.$$

Fundamental fact: we have $|A \times B| = |A| \cdot |B|$.

We write A^2 for $A \times A$ and in general $A^n = \underbrace{A \times \cdots \times A}_{n \text{ times}}$.

This is how we get $\mathbb{R}^2$ (the set of points (x, y) in the plane), $\mathbb{R}^3$ (the set of points (x, y, z) in 3-dimensional space) and so on.

What is a probability?

A "probability" is a number we attach to an "event".

We'll soon say some formal mathematical things about what "probability" and "event" are, but we'd like them to mean something in the real world.

What should they mean?

- **Frequentist** interpretation: the probability of an event measures how often it will happen over many repeated trials.
- **Bayesian** interpretation: the probability of an event measures our degree of belief in the event happening.

Frequentist interpretation of probability

Imagine tossing a coin many times.

- We expect that after N coinflips, close to $\frac{1}{2}N$ outcomes will be Heads, and close to $\frac{1}{2}N$ outcomes will be Tails.
- If we tossed the coin $N = 1\,000\,000$ times, and about $\frac{2}{3}N$ of the outcomes were Heads, we'd suspect that the coin is unfair.

The frequentist interpretation says that the probability of the coin landing heads **is** the limiting fraction it lands heads, after many tosses.

Advantage: this tells us what to do if we want to find out the probability. Disadvantage: it's not practical in some cases.

Bayesian interpretation of probability

The Bayesian interpretation says that the probability of the coin landing heads measures our confidence or subjective degree of belief in the coin landing heads.

- Big disadvantage: different people have different beliefs, sometimes for stupid reasons.
- Advantage: the event doesn't have to be repeatable for us to have beliefs about it.
- Advantage: different people have different beliefs, because they have different information.

  If I saw the coin land Heads $1\,000\,000$ times in a row, but you haven't, it makes sense for us to say different things about the probability of it landing Heads.

Limits on beliefs about probability

Are "beliefs about probability" completely arbitrary? Not quite!

Suppose I roll a 6-sided die, and decide to believe that there's a $\frac{1}{2}$ chance that it will come up 1, but **also** a $\frac{1}{2}$ chance that it will come up 2, **and**, a $\frac{1}{2}$ chance that it will come up 3, and so on.

Then I'll probably accept a double-or-nothing bet on any outcome: I give you $\$10$, and if the die comes up 1, you'll give me $\$20$. If you make this bet for all six outcomes, then I make a guaranteed loss!

Whether we're talking about frequentist or Bayesian probability, the probabilities of events have to satisfy some axioms to avoid being stupid.

What are events?

In a random experiment:

- We start with a **sample space**: the set of all possible **outcomes**.

  Example 1: $S = \{\text{Heads}, \text{Tails}\} = \{\text{H}, \text{T}\}$ is our sample space for a single coin toss.

  Example 2: $S = \{1, 2, 3, 4, 5, 6\}$ is our sample space for rolling a 6-sided die.

- Something random happens, and one of the outcomes is chosen.

An **event** is a subset of S. Each event $A \subseteq S$ has a **probability** $\Pr[A]$.

Technical note: when S is uncountably infinite, it's possible that not all subsets of S should be considered valid events. Uncountable sets are weird!

Events and set operations

Set operations on events have reasonable meanings:

- $A \cap B$ is an event which happens if both A and B happen.

  $\bigcap_{i=1}^{n} A_i = A_1 \cap \cdots \cap A_n$ happens if all of $A_1, \ldots, A_n$ happen.

- $A \cup B$ is an event which happens if either A or B happens.

  $\bigcup_{i=1}^{n} A_i$ happens if at least one of $A_1, \ldots, A_n$ happens.

- A^c is an event which happens if A does not happen.

We don't assign a probability to an outcome x, only to the event $\{x\}$.

Axioms of probability

Here are some basic facts which **must** be true of probabilities:

1. For any event A, $\Pr[A] \geq 0$.
2. The whole sample space S has probability $\Pr[S] = 1$.
3. If events A and B are **disjoint** (if $A \cap B = \varnothing$) then $\Pr[A \cup B] = \Pr[A] + \Pr[B]$.

   More generally, if events $A_1, A_2, A_3, \ldots$ are pairwise disjoint, then

$$\Pr\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \Pr[A_i].$$

A hint at more formalism

(Don't worry about this slide, unless you're curious.)

Formally, we say that a **probability space** is a triple $(\Omega, \mathcal{F}, \Pr)$, where:

- Ω is the sample space; it can be anything.
- $\mathcal{F}$ is a collection of subsets of Ω (a set whose elements are themselves sets); elements of $\mathcal{F}$ are events.

  It must satisfy certain conditions: if we take some events, and perform the basic operations we discussed, the result must also be an event.
- $\Pr$ is a function from $\mathcal{F}$ to $[0, 1]$ that assigns each event a probability. It must satisfy the axioms on our previous slide.

Some consequences

Claim 1. For any event A, $\Pr[A^c] = 1 - \Pr[A]$.

Proof. A and A^c are disjoint, so $\Pr[A \cup A^c] = \Pr[A] + \Pr[A^c]$. But $A \cup A^c = S$, so $\Pr[A \cup A^c] = \Pr[S] = 1$. Therefore $\Pr[A]$ and $\Pr[A^c]$ must add to 1, and $\Pr[A^c] = 1 - \Pr[A]$. □

Claim 2. If $A \subseteq B$, then $\Pr[A] \leq \Pr[B]$.

Proof. We can check (say, with Venn diagrams) that if $A \subseteq B$, then $B = A \cup (B - A)$, and A is disjoint to $B - A$. Therefore $\Pr[A] + \Pr[B - A] = \Pr[B]$. Since $\Pr[B - A] \geq 0$, we get the inequality we wanted. □

Claim 3. For any two events A and B, $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$.

(Try this one yourself!)

Uniform probability

A very common simple random experiment: the sample space S has n elements, and $\Pr[\{x\}] = \frac{1}{n}$ for all $x \in S$.

We say that we're choosing an element **uniformly at random** from S.

Claim. In this case, if $|A| = k$, then $\Pr[A] = \frac{k}{n}$.

Proof. Suppose that $A = \{x_1, x_2, \ldots, x_k\}$. Then

$$\begin{aligned}\Pr[A] &= \Pr[\{x_1\}] + \Pr[\{x_2\}] + \cdots + \Pr[\{x_k\}] \\ &= \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} \ (k \text{ times}) \\ &= \frac{k}{n}.\end{aligned}$$

□

How to find probabilities

We'll spend the entire semester talking about how to find probabilities, but some basic ideas will always stay the same.

- Almost every problem has **some** probabilities that are relatively easy to find, or maybe even already known.
- Other probabilities can be split up into cases. Sometimes, taking the complement simplifies the problem.
- We'll also look a bit at fancier counting strategies like the principle of inclusion-exclusion.

Today's example

Suppose we roll two fair 6-sided dice. What is the probability that we roll **at least** one 6?

What if we have three dice? What if we want a formula for n dice, in terms of n?

We'll look at several good and bad ways to solve this problem.

We can muddle through the two-dice problem even with a bad strategy. So the real test of a strategy for this problem will be how it generalizes to three or to n dice.

Uniform probability and counting

We saw yesterday that if we are sampling uniformly from a set S, then $\Pr[A] = \frac{|A|}{|S|}$ for all subsets A of S.

In today's example, we are sampling uniformly from the set

$$S = \{(1,1), (1,2), (1,3), \ldots, (6,5), (6,6)\}.$$

We want the probability of

$$A = \{(1,6), (2,6), \ldots, (6,6), (6,5), \ldots, (6,1)\}.$$

Just counting tells us that $|A| = 11$ and $|S| = 36$, so $\Pr[A] = \frac{11}{36}$.

Casework

We know that probabilities of **disjoint** events add. So if we split up our problem into several **disjoint** cases that are easy to handle, we can solve the whole problem.

Note: our event "at least one die lands 6" is **not** the disjoint union of "the first die lands 6" and "the second die lands 6". So $\Pr[A]$ is not $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. **Avoid this mistake!**

We could split up A into A_1 "the first die lands 6", and A_2 "the second die lands 6, but the first does not". Then

$$\Pr[A] = \Pr[A_1] + \Pr[A_2] = \frac{1}{6} + \frac{5}{36} = \frac{11}{36}$$

(though $\Pr[A_2]$ is not as obvious to find).

Inclusion-exclusion principle

Just as with sets, we have an inclusion-exclusion principle for events and probabilities:

1. For two events A and B, $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$.
2. For three events A, B, C,

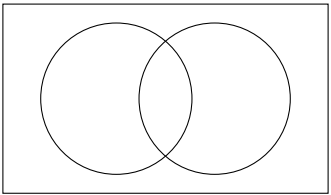
$$\begin{aligned}\Pr[A \cup B \cup C] = &\Pr[A] + \Pr[B] + \Pr[C] \\ &- \Pr[A \cap B] - \Pr[A \cap C] - \Pr[B \cap C] \\ &+ \Pr[A \cap B \cap C].\end{aligned}$$

3. For events $A_1, A_2, \ldots, A_n$, we include the k-fold intersection $\Pr[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}]$ with a coefficient of $(-1)^{k+1}$, for $k = 1, 2, \ldots, n$.

How do we know this?

Our reasoning with Venn diagrams carries over to probabilities.

We can **partition** the sample space S into four events: $A \cap B$, $A \cap B^c$, $A^c \cap B$, $A^c \cap B^c$. These are exactly the four regions in a Venn diagram!



We can write $\Pr[A]$, $\Pr[B]$, $\Pr[A \cup B]$, and $\Pr[A \cap B]$ in terms of these four events, then verify that $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$.

A Venn diagram proof without diagrams

We have:

$$\begin{aligned}
\Pr[A \cup B] &= \Pr[A \cap B] + \Pr[A \cap B^c] + \Pr[A^c \cap B] \\
\Pr[A] &= \Pr[A \cap B] + \Pr[A \cap B^c] \\
\Pr[B] &= \Pr[A \cap B] + \Pr[A^c \cap B] \\
\Pr[A \cap B] &= \Pr[A \cap B]
\end{aligned}$$

If we compute $\Pr[A] + \Pr[B] - \Pr[A \cap B]$ by adding up what we see on the left, we get all three terms of $\Pr[A \cup B]$ exactly once.

We can do this for $\Pr[A \cup B \cup C]$, but with eight regions in the Venn diagram: $\Pr[A \cap B \cap C]$, $\Pr[A \cap B \cap C^c]$, and so on.

For more than 3 events, we'd probably want to begin by finding a pattern in these expressions...

Three dice

Suppose we roll three fair dice. ($S = \{(1,1,1),(1,1,2),\ldots,(6,6,6)\}$.)

What's the probability that we see **at least one** 6?

Let A_i for $i = 1,2,3$ be the event that the i^{th} die comes up 6. Let's expand $\Pr[A_1 \cup A_2 \cup A_3]$ using the inclusion-exclusion principle.

- We have $\Pr[A_1] = \Pr[A_2] = \Pr[A_3] = \frac{1}{6}$.
- $\Pr[A_1 \cap A_2] = \Pr[A_1 \cap A_3] = \Pr[A_2 \cap A_3] = \frac{1}{36}$.
- $\Pr[A_1 \cap A_2 \cap A_3] = \frac{1}{216}$.

So the answer is $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \frac{1}{36} - \frac{1}{36} - \frac{1}{36} + \frac{1}{216} = \frac{91}{216}$.

Cartesian products and complements

When we roll two dice, and $S = \{(1,1),(1,2),\ldots,(5,6),(6,6)\}$, there is a shortcut to finding $|S|$. We have

$$S = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$$

so $|S| = 6 \cdot 6 = 36$.

The set A where at least one die lands 6 has no simple description. But A^c does!

$$A^c = \{1,2,3,4,5\} \times \{1,2,3,4,5\}$$

so $|A^c| = 5 \cdot 5 = 25$ and $\Pr[A] = 1 - \Pr[A^c] = 1 - \frac{25}{36} = \frac{11}{36}$.

Solving the n-dice problem

This last method is the only one that generalizes well to any number of dice.

Suppose we roll n dice. Then $S = \{1, 2, 3, 4, 5, 6\}^n$, so $|S| = 6^n$.

If A is the event "at least one 6", then A^c is the event "no sixes". So $A^c = \{1, 2, 3, 4, 5\}^n$, and $|A^c| = 5^n$.

We get $\Pr[A] = 1 - \Pr[A^c] = 1 - \frac{5^n}{6^n}$.

This gives $1 - \frac{5^3}{6^3} = 1 - \frac{125}{216} = \frac{91}{216}$ for three dice. This matches our answer via inclusion-exclusion, but it's easier to find.

Discrete and continuous models

Deciding on a sample space S and an assignment of probabilities $\Pr[A]$ to events A is what your textbook calls a "probability model".

(Sometimes, and with more formality, we talk about "probability spaces".)

This week we will discuss:

1. Discrete models (finite and infinite) which we can specify by giving $\Pr[\{x\}]$ for all $x \in S$.
2. Continuous models, whose sample space S is uncountable, and where $\Pr[\{x\}] = 0$ for all or most $x \in S$.

   These must be described by giving a rule for $\Pr[A]$ for a different class of events.

Rolling a die

If we roll a 6-sided die, our probability space is

$$S = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}.$$

Rolling a fair die means choosing uniformly at random from S: $\Pr[\{\boxdot\}] = \cdots = \Pr[\{\boxdot\}] = \frac{1}{6}$.

We can also consider weighted dice: for example, maybe $\Pr[\{\boxdot\}] = \frac{1}{2}$ and $\Pr[\{\boxdot\}] = \cdots = \Pr[\{\boxdot\}] = \frac{1}{10}$. However, we must **always** have

$$\Pr[\{\boxdot\}] + \Pr[\{\boxdot\}] + \Pr[\{\boxdot\}] + \Pr[\{\boxdot\}] + \Pr[\{\boxdot\}] + \Pr[\{\boxdot\}] = 1.$$

With this weighted die, we have

$$\Pr[\text{roll an even number}] = \Pr[\{\boxdot, \boxdot, \boxdot\}] = \frac{1}{10} + \frac{1}{10} + \frac{1}{2} = \frac{7}{10}.$$

Rolling many dice

What happens if we take the weighted die with $\Pr[\{\boxplus\}] = \frac{1}{2}$, and roll it twice?

- Our probability space is

$$S = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxplus\}^2 = \{\boxdot\boxdot, \boxdot\boxdot, \ldots, \boxplus\boxplus\}.$$

- For any $x, y \in \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxplus\}$, we can find $\Pr[\{(x, y)\}]$ by **multiplying** the probabilities for individual dice.

Example: $\Pr[\text{doubles}] = \Pr[\{\boxdot\boxdot, \boxdot\boxdot, \boxdot\boxdot, \boxdot\boxdot, \boxdot\boxdot, \boxplus\boxplus\}]$ is

$$(\tfrac{1}{10})^2 + (\tfrac{1}{10})^2 + (\tfrac{1}{10})^2 + (\tfrac{1}{10})^2 + (\tfrac{1}{10})^2 + (\tfrac{1}{2})^2 = \tfrac{3}{10}.$$

Generalization

In general, if we have a finite sample space $S = \{x_1, x_2, \ldots, x_n\}$, we can describe our probability model by giving $\Pr[\{x_i\}]$ for $i = 1, 2, \ldots, n$. For arbitrary sets A, we can then take the sum:

$$\Pr[A] = \sum_{x \in A} \Pr[\{x\}].$$

Suppose we have two experiments with sample spaces S_1, S_2 and probabilities $\Pr_1$, $\Pr_2$. Doing both experiments gives us a model with sample space $S = S_1 \times S_2$, where the probabilities of singleton events are

$$\Pr[\{(x, y)\}] = \Pr_1[\{x\}] \cdot \Pr_2[\{y\}].$$

Setting up your model

Roll two **fair** dice. What is the probability that their total is 8?

You have options for how to set this up:

- Make the sample space $S = \{2, 3, \ldots, 12\}$ for the various totals.

  Now, $\Pr[\{8\}]$ is the answer; but it's hard to find.

- Make the sample space consist of all 21 things you could see, such as "two ⚄'s" or "⚂ and ⚅".

  Still weird: "⚂ and ⚅" is twice as likely as "two ⚄'s".

- Make the sample space consist of all 36 pairs (first die, second die), distinguishing (⚂, ⚅) from (⚅, ⚂).

  Now the probability space is uniform, making things easy.

Specifying an infinite model

Suppose you pick a random positive integer. What is the probability that it is odd?

Trick question! We need to specify the distribution first. (It can't be uniform.)

If $S = \{1, 2, 3, \dots\}$, we specify the probability model in the same way as before: give a value $\Pr[\{k\}]$ for each $k \in S$. These must be nonnegative and satisfy

$$\sum_{k=1}^{\infty} \Pr[\{k\}] = 1.$$

Not much changes, except that we end up dealing with infinite sums.

A coin-flipping experiment

Suppose we flip a coin until we see Heads for the first time. Our sample space is $S = \{\mathsf{H}, \mathsf{TH}, \mathsf{TTH}, \mathsf{TTTH}, \mathsf{TTTTH}, \dots\}$ and the probabilities are given by

$$\Pr[\{\underbrace{\mathsf{TTT}\dots\mathsf{TH}}_{k \text{ flips}}\}] = \frac{1}{2^k}.$$

What is the probability that we see Heads in **at most five** flips?

$$\Pr[\{\mathsf{H}, \mathsf{TH}, \mathsf{TTH}, \mathsf{TTTH}, \mathsf{TTTTH}\}] = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} = \frac{31}{32}.$$

What is the probability that the number of coin flips is **odd**?

$$\Pr[\{\mathsf{H}, \mathsf{TTH}, \mathsf{TTTTH}, \dots\}] = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \cdots = ?$$

Geometric series

An **infinite geometric series** is a sum of the form

$$\sum_{k=0}^{\infty} a \cdot r^k.$$

When $|r| < 1$, this sum simplifies to $\frac{a}{1-r}$. This formula will come up often in this class, so keep it in mind!

In our current problem, $\Pr[\text{odd number of flips}]$ is

$$\frac{1}{2^1} + \frac{1}{2^3} + \frac{1}{2^5} + \frac{1}{2^7} + \cdots = \sum_{k=0}^{\infty} \frac{1}{2} \cdot \left(\frac{1}{4}\right)^k.$$

Here, $a = \frac{1}{2}$ and $r = \frac{1}{4}$, so the probability simplifies to $\frac{1/2}{1-1/4} = \frac{2}{3}$.

Finding the normalizing constant

There is a very common problem that arises in probability problems. You figure out a formula for probabilities in a sample space, but there's an unknown constant factor.

For example, maybe $S = \{0, 1, 2, 3, \dots\}$ and $\Pr[\{k\}] = \frac{C}{3^k}$ for some unknown C.

We can figure out the constant by knowing that $\Pr[S] = 1$. In this example:

$$\Pr[S] = \sum_{k=0}^{\infty} \frac{C}{3^k} = 1.$$

This is a geometric series with $a = \frac{C}{3^0} = C$ and $r = \frac{1}{3}$. So $\frac{C}{1-1/3} = 1$, and therefore $C = \frac{2}{3}$.

Motivation for infinite models

Infinite models often **don't quite** fit reality. In practice, if you're flipping a coin until it lands heads, there's a limiting factor: the time you have to spare for coin-flipping.

Or maybe Google finds out that the number of search requests per second is well-approximated by a random model which has $\Pr[\{k\}] = C \cdot \frac{40000^k}{k!}$. This is nonsense for very large k, such as k bigger than the number of computers on Earth.

Usually, we use the infinite model anyway. That's because:

- The true finite bound is often impossible to know;
- The infinite model is a good approximation;
- The infinite model is **simpler** than a finite one!

Random real numbers

We will begin by looking at models where our sample space is the real numbers $\mathbb{R}$, or a sub-interval of $\mathbb{R}$.

Example. During a thunderstorm, you measure the time (in seconds) until you hear thunder. The sample space is $S = [0, \infty)$.

- It's not practically possible to measure the exact time. Even if your stopwatch says "2.51 seconds", that just means that the time is in the interval $[2.51, 2.52)$.
- The more precision we ask for, the less likely the event is. Infinite precision means zero probability: $\Pr[\{t\}] = 0$ for any $t \in [0, \infty)$.
- We might describe the probability distribution by a formula such as $\Pr\left[[a, b)\right] = e^{0.1a} - e^{0.1b}$ (or whatever).

Uniform probability and buses

Paradoxically, by making our sample space much, much larger, we're back in a situation where we can (sometimes) have a uniform distribution!

Example. A bus scheduled for 1pm is actually equally likely to arrive at any time between 1pm and 2pm.

1. What is the probability it arrives between 1:20pm and 1:30pm?

   All 10-minute intervals are equally likely, so we should get $\frac{1}{6}$.

2. What is the probability it arrives between 1:17pm and 1:24pm?

   All 1-minute intervals are equally likely, so we should get $\frac{7}{60}$.

This generalizes: any t-minute interval has probability $\frac{t}{60}$.

Continuous uniform probability

When picking uniformly from a finite set S, we had the formula $\Pr[A] = \frac{|A|}{|S|}$.

In the world where we are picking uniformly from an interval $S = [a, b]$, something similar happens. We have

$$\Pr[A] = \frac{\text{length}(A)}{\text{length}(S)}$$

whenever A is a sub-interval of S.

Not all events are sub-intervals. But any event we'll care about can at least be broken up into intervals and has a well-defined "total length".

This does not allow us to pick uniformly from $S = \mathbb{R}$, or any subset of $\mathbb{R}$ with infinite length! Such sets are once again "too big".

Uniform probability in higher dimensions

We can also end up choosing uniformly from suitable subsets of $\mathbb{R}^2$. This can happen in two ways:

1. Points in the plane. We can imagine choosing a point from the unit square

$$[0,1]^2 = \{(x,y) \in \mathbb{R}^2 : 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1\}$$

uniformly at random.

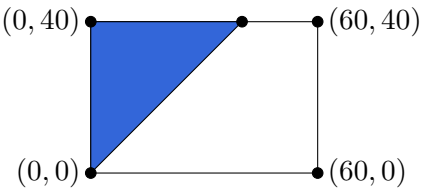
2. Iterated experiments. If we choose x uniformly from $[0,1]$, and separately choose y uniformly from $[0,1]$, then (x,y) is chosen uniformly from $[0,1]^2$ again.

Either way, we can deal with $\mathbb{R}^2$ in the same way as $\mathbb{R}$, except length is replaced by **area**.

Example: which bus arrives first?

You can take one of two buses home. Bus X arrives at a uniform time between 1pm and 2pm. Bus Y arrives at a uniform time between 1pm and 1:40pm. What is the probability that bus X will arrive first?

Our sample space is $S = [0, 60] \times [0, 40]$ and our event "X is first" is $A = \{(x, y) \in S : x < y\}$. Draw a picture:



S has area 2400 and A has area 800, so $\Pr[A] = \frac{800}{2400} = \frac{1}{3}$.

Why uncountability matters

Theorem. In any probability model, there are at most countably many outcomes x with $\Pr[\{x\}] > 0$.

Proof. We can list all such outcomes!

First list all outcomes x with $\Pr[\{x\}] \geq 0.1$. (There are at most 10.)

Then list all outcomes x with $\Pr[\{x\}] \geq 0.01$. (There are at most 100.)

Then repeat with 0.001, 0.0001, and so on. We will eventually get to any event with positive probability. □

This means that whenever the sample space S is uncountable, we have $\Pr[\{x\}] = 0$ for almost all x.

Another uncountable model

The most common uncountable sample spaces are subsets of $\mathbb{R}$ or $\mathbb{R}^2$. But some have nothing to do with the real numbers.

For example, suppose we flip a coin infinitely many times. Here, $S = \{\mathsf{H}, \mathsf{T}\}^\infty$ is the set of all infinite sequences of coinflips.

Pick any outcome, such as $x = \mathsf{HHHHH}\ldots$. What is $\Pr[\{x\}]$?

- We must at least get H on the first flip, which has probability $\frac{1}{2}$, so $\Pr[\{x\}] \leq \frac{1}{2}$.
- Actually, we must at least get H on the first ten flips, which has probability $\frac{1}{2^{10}}$, so $\Pr[\{x\}] \leq \frac{1}{2^{10}} = \frac{1}{1024}$.
- This continues, getting smaller and smaller bounds. We must have $\Pr[\{x\}] = 0$.

How to work in the infinite coin model

How can we find probabilities in this infinite-coinflips world?

With the real numbers, our basic objects to work with were intervals. Here, there is a different kind of set whose probability we understand.

Suppose an event can be described in terms of only the first k coinflips. Then, we can find its probability: we can ignore all other coinflips, and end up with a finite model.

We can deal with other events such as "the first H occurs on an odd-numbered flip" by summing over many events whose probabilities we can find.

(We've already done this!)

Probability 0 events

In all of these examples, there are some events (even ones containing infinitely many outcomes) with probability 0.

- Suppose we choose $(x, y) \in [0, 1]^2$ uniformly at random. Then the probability that $x = y$ is 0.
- Suppose we toss a coin infinitely many times. Then the probability that we see H fewer than 100 times is 0.

This does not mean that the outcomes contained in such an event **cannot** happen. It means that the event is a **vanishingly small** fraction of outcomes.

In real-life cases, often it would mean that if the event did happen, we wouldn't be able to confirm this!

An example with dice

Let's say we are rolling two fair dice: our sample space is

$$S = \{\boxdot\boxdot, \boxdot\boxdot, \ldots, \boxdot\boxdot\}$$

and we are sampling uniformly at random.

I look at the dice and give you partial information: I tell you that **the total is** 8. How should you think about your remaining uncertainty?

It makes sense to move to a smaller sample space: the sample space

$$S' = \{\boxdot\boxdot, \boxdot\boxdot, \boxdot\boxdot, \boxdot\boxdot, \boxdot\boxdot\}.$$

Probabilities change: before, $\Pr[\text{first die is } \boxdot]$ was $\frac{1}{6}$, now it's $\frac{1}{5}$.

Generalizing

In general, suppose we are sampling **uniformly** at random from a sample space S. We are told that a particular event B occurred. How do things change?

- The set B is our new sample space: only outcomes in B are possible now.
- No outcome in B is more likely than any other: we are still sampling uniformly, but now from B.
- If A is some other event, then $A \cap B$ is the new set of outcomes that can make A happen.

  The old probability of A was $\frac{|A|}{|S|}$. The new probability is $\frac{|A \cap B|}{|B|}$.

Conditional probability

The **conditional probability of A given B** is defined as

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

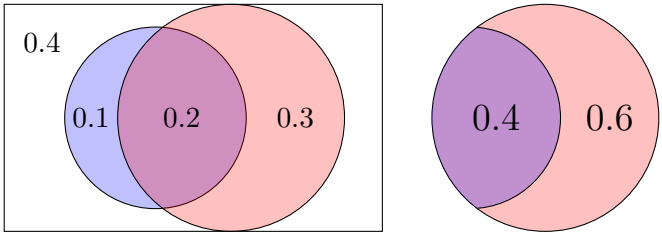
For sampling uniformly at random, this matches our "new probability of A, assuming B happened":

$$\frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|S|}{|B|/|S|} = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Even when we're not sampling uniformly, this definition turns out to behave correctly in situations where we learn that B happened, and want to know the probability that A also happened.

Conditional probability with Venn diagrams

Say $\Pr[\text{rain}] = 0.3$, $\Pr[80^\circ \text{ weather}] = 0.5$, $\Pr[\text{rain and } 80^\circ] = 0.2$.



$\Pr[\text{rain} \mid 80^\circ] = \frac{0.2}{0.2+0.3} = 0.4$ and $\Pr[\text{no rain} \mid 80^\circ] = \frac{0.3}{0.2+0.3} = 0.6$.

Proving things about conditional probability

Recall: a **probability measure** is a function $\Pr$ from events to $\mathbb{R}$ satisfying the axioms:

1. $\Pr[A] \geq 0$ for all events.
2. $\Pr[S] = 1$.
3. $\Pr[A_1 \cup A_2 \cup \dots] = \Pr[A_1] + \Pr[A_2] + \dots$ when $A_1, A_2, \dots$ are disjoint events.

Well, $\Pr[\cdot \mid B]$ is also a function from events to $\mathbb{R}$ satisfying these axioms!

1. $\Pr[A \mid B] \geq 0$ for all events.
2. $\Pr[S \mid B] = 1 = \Pr[B \mid B]$.
3. $\Pr[A_1 \cup A_2 \cup \dots \mid B] = \Pr[A_1 \mid B] + \Pr[A_2 \mid B] + \dots$ when $A_1, A_2, \dots$ are disjoint.

So everything we've proven for probabilities, we get for conditional probabilities for free! For example:

$$\Pr[A \cup B \mid C] = \Pr[A \mid C] + \Pr[B \mid C] - \Pr[A \cap B \mid C].$$

The chain rule for conditional probability

We can rewrite our definition as $\Pr[A \cap B] = \Pr[B] \cdot \Pr[A \mid B]$.

For three events:

$$\begin{aligned}\Pr[A \cap B \cap C] &= \Pr[B \cap C] \cdot \Pr[A \mid B \cap C] \\ &= \Pr[C] \cdot \Pr[B \mid C] \cdot \Pr[A \mid B \cap C].\end{aligned}$$

In general:

$$\Pr\left[\bigcap_{i=1}^{n} A_i\right] = \Pr[A_1] \cdot \Pr[A_2 \mid A_1] \cdot \Pr[A_3 \mid A_1 \cap A_2] \cdots \Pr\left[A_n \,\middle|\, \bigcap_{i=1}^{n-1} A_i\right].$$

"To find the probabilities of n events, multiply together the probability of each event **conditioned on the previous events having happened**."

The chain rule in action

We draw two cards from a 52-card deck, without replacement. What is the probability that they're both aces?

Let A_1 be "the first card is an ace" and let A_2 be "the second card is an ace". Then $\Pr[A_1] = \Pr[A_2] = \frac{1}{13}$, but we can't just multiply these.

We have

$$\Pr[A_1 \cap A_2] = \Pr[A_1] \cdot \Pr[A_2 \mid A_1].$$

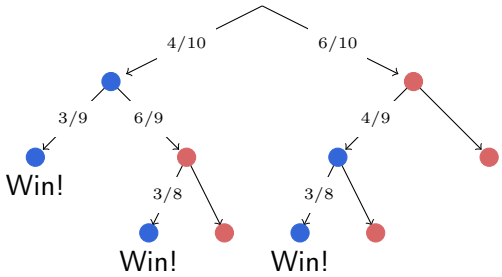
Here, $\Pr[A_1] = \frac{4}{52} = \frac{1}{13}\ldots$ and $\Pr[A_2 \mid A_1] = \frac{3}{51} = \frac{1}{17}\ldots$, so

$$\Pr[A_1 \cap A_2] = \frac{1}{13} \cdot \frac{1}{17} = \frac{1}{221}.$$

For three cards: $\frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{1}{5525}$.

Branching diagrams

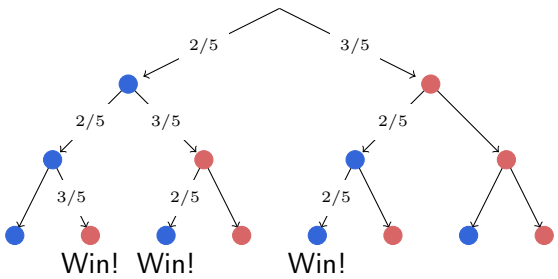
You draw 3 marbles (without replacement) from a bag containing $4 \times$ ● and $6 \times$ ●. What is the probability of drawing at least 2 ●?



Answer: $\frac{4}{10} \cdot \frac{6}{9} \cdot \frac{3}{8} + \frac{4}{10} \cdot \frac{3}{9} + \frac{6}{10} \cdot \frac{4}{9} \cdot \frac{3}{8} = \frac{1}{10} + \frac{2}{15} + \frac{1}{10} = \frac{1}{3}$.

Motivation: silly branching diagrams

You draw 3 marbles **with replacement** from a bag containing $4 \times$ ● and $6 \times$ ●. What is the probability of drawing **exactly** 2 ●?



We have three paths to victory, but they all have the same probability $\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{125}$. One draw does not affect the others!

Two independent events

Two events A and B are independent if $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$. Equivalently:

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[A] \cdot \Pr[B]}{\Pr[B]} = \Pr[A].$$

"Learning that B happened tells us nothing about A."

That this happens with repeated trials of the an experiment: If we roll a die twice, the events "first roll is ⚅" and "second roll is ⚅" are independent.

This also happens "spontaneously" in some cases: we can check that for a single roll, the events $\{⚀, ⚅\}$ and $\{⚀, ⚂, ⚃\}$ are independent.

Independent events are not disjoint events!

A common mistake is confusing **independent** events with **disjoint events**.

- Disjoint events: if one happens, it tells you a lot about the other event—it tells you that the other event can't happen!
  - "roll an odd #" and "roll ⚅" are disjoint.
  - when drawing cards **without** replacement, "first card is A♡" and "second card is A♡" are disjoint.
- Independent events: if one happens, it tells you **nothing** about the other event.
  - "roll an odd #" and "roll ⚀ or ⚂" are independent.
  - when drawing cards **with** replacement, "first card is A♡" and "second card is A♡" are independent.

Consequences of independence

If A and B are independent, so are: A and B^c; A^c and B; A^c and B^c. From $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$, we can prove:

$$\begin{aligned}\Pr[A \cap B^c] &= \Pr[A] - \Pr[A \cap B] \\ &= \Pr[A] - \Pr[A] \cdot \Pr[B] \\ &= \Pr[A] \cdot (1 - \Pr[B]) \\ &= \Pr[A] \cdot \Pr[B^c].\end{aligned}$$

and the same from the other cases.

This means that when A and B are independent, we don't have to do branching diagrams: knowing $\Pr[A]$ and $\Pr[B]$ tells everything about all the branches.

Independence of three events

Idea: want to say that A, B, C are independent if we don't have to do branching diagrams to understand combinations of A, B, C.

Two possible (equivalent) definitions of when A, B, C are independent:

When we have **all four of**

$$\begin{cases} \Pr[A \cap B] = \Pr[A] \cdot \Pr[B] \\ \Pr[A \cap C] = \Pr[A] \cdot \Pr[C] \\ \Pr[B \cap C] = \Pr[B] \cdot \Pr[C] \\ \Pr[A \cap B \cap C] = \Pr[A] \cdot \Pr[B] \cdot \Pr[C] \end{cases}$$

When we have **all eight of**

$$\begin{cases} \Pr[A \cap B \cap C] = \Pr[A] \cdot \Pr[B] \cdot \Pr[C] \\ \Pr[A \cap B \cap C^c] = \Pr[A] \cdot \Pr[B] \cdot \Pr[C^c] \\ \dots \\ \Pr[A^c \cap B^c \cap C^c] = \Pr[A^c] \cdot \Pr[B^c] \cdot \Pr[C^c] \end{cases}$$

Just $\Pr[A \cap B \cap C] = \Pr[A] \cdot \Pr[B] \cdot \Pr[C]$ isn't enough: that's only one of the branches!

Independence of any number of events

Events $A_1, A_2, \ldots, A_n$ are independent if:

- We have

$$\Pr[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = \Pr[A_{i_1}] \cdot \Pr[A_{i_2}] \cdots \Pr[A_{i_k}]$$

  whenever we intersect **all** or **some** of the events.

- (Equivalent definition) we have

$$\Pr[B_1 \cap B_2 \cap \cdots \cap B_n] = \Pr[B_1] \cdot \Pr[B_2] \cdots \Pr[B_n]$$

  where each B_i can be either A_i or A_i^c (in any combination).

Once again, the point is that when we have independence, we don't need branching diagrams.

1000 coinflips

Suppose that you flip 999 coins fairly, but instead of flipping the 1000^{th} coin, you pick a side for it (Heads or Tails) to make the total number of heads odd.

For $i = 1, 2, \ldots, 1000$, let A_i be the event "the i^{th} coin is H".

All 1000 events are not independent: their intersection has probability 0, not $(\frac{1}{2})^{1000}$.

But any 999 of these events are independent!

(This is easier to see for the 999 events that don't look at the last coin. But the coins are actually symmetric; we are picking uniformly from the 2^{999} outcomes with an odd number of heads.)

Divisibility: a number theory application

How many numbers from 1 to 100 have no common factors with 100?

The inclusion-exclusion solution: if $A_2 = \{\text{multiples of } 2\}$ and $A_5 = \{\text{multiples of } 5\}$, then the answer is

$$100 - |A_2| - |A_5| + |A_2 \cap A_5| = 100 - \frac{100}{2} - \frac{100}{5} + \frac{100}{10} = 40.$$

This factors as $100(1 - \frac{1}{2})(1 - \frac{1}{5})$, which tells us that we missed an easier approach. If we choose a random element of $S = \{1, \ldots, 100\}$, then A_2 and A_5 are independent!

How many numbers from 1 to n have no common factors with n? Independence tells us that the answer is

$$n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\ldots\left(1 - \frac{1}{p_k}\right).$$

Partitions and where to find them

Events $B_1, B_2, \ldots, B_n$ form a **partition** of the sample space if

1. For any $i \neq j$, $B_i \cap B_j = \varnothing$: any two of the events are disjoint.
2. $B_1 \cup B_2 \cup \cdots \cup B_n = S$: one of the events is guaranteed to happen.

Examples:

- For any event A, the two events $\{A, A^c\}$ are a partition.
- More generally, the 2^k regions in a k-event Venn diagram are a partition.
- The most interesting cases are where $B_1, B_2, \ldots, B_n$ describe competing hypotheses or competing models: instead of observing them directly, we observe their effects on other events.

Examples of partitions

Some examples:

- A six-sided die may be fair ($\frac{1}{6}$ chance of each side) or weighted ($\frac{1}{2}$ chance of a ⚅ and $\frac{1}{10}$ chance of other sides).
- A transmitter is trying to send a bit (0 or 1) or a message of several bits over a noisy channel that randomly changes part of the transmission.
- You bought a laptop that's either high-quality (and lasts for $[0, 5]$ years) or defective (and lasts for exactly one year).

In each of these cases, we can imagine observing consequences of these events (rolling the dice; receiving the message; using the laptop until it breaks) without knowing which event in the partition happened.

Law of total probability

The law of total probability says: suppose we have a partition $B_1, B_2, \ldots, B_n$ of the sample space. Then for any event A,

$$\Pr[A] = \sum_{i=1}^{n} \Pr[A \cap B_i] = \sum_{i=1}^{n} \Pr[A \mid B_i] \cdot \Pr[B_i].$$

Example: suppose you roll a die that's equally likely to be fair or weighted. Then

$$\Pr[\boxdot] = \Pr[\boxdot \mid \text{fair}] \Pr[\text{fair}] + \Pr[\boxdot \mid \text{weighted}] \Pr[\text{weighted}]$$

which evaluates to $\frac{1}{6} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{3}$.

The defective laptop

Suppose a laptop usually works for $[0, 5]$ years, but 10% of the time it's defective and breaks after exactly one year.

- What is the probability that it lasts exactly 2 years?

  Whether or not it's defective, the probability is 0.

- What is the probability that it lasts exactly 1 year?

  This is 0 for a good laptop, but 1 for a defective laptop, so the probability is $\frac{1}{10}$. (This is a rare example in which $\Pr[\{x\}] = 0$ for most outcomes x, but some have positive probability!)

- What is the probability it lasts less than 2 years?

  This is $\frac{2}{5}$ for a good laptop, and 1 for a defective laptop. We use the law of total probability and get $\frac{2}{5} \cdot \frac{9}{10} + 1 \cdot \frac{1}{10} = 0.46$.

Rolling a die twice

Suppose that we take a die which is equally likely to be fair or weighted, and roll it **twice**.

Let A_1 be the event that the first roll is ⚅. Let A_2 be the event that the second roll is ⚅.

- If the die is fair, then A_1 and A_2 are independent: $\Pr[A_1 \mid \text{fair}] = \Pr[A_2 \mid \text{fair}] = \frac{1}{6}$ and $\Pr[A_1 \cap A_2 \mid \text{fair}] = \frac{1}{36}$.
- If the die is weighted, then A_1 and A_2 are independent: $\Pr[A_1 \mid \text{weighted}] = \Pr[A_2 \mid \text{weighted}] = \frac{1}{2}$ and $\Pr[A_1 \cap A_2 \mid \text{weighted}] = \frac{1}{4}$.

But are A_1 and A_2 independent?

Conditional independence

We have already computed $\Pr[A_1] = \Pr[A_2] = \frac{1}{3}$.

By a similar process:

$$\Pr[A_1 \cap A_2] = \frac{1}{2}\Pr[A_1 \cap A_2 \mid \text{fair}] + \frac{1}{2}\Pr[A_1 \cap A_2 \mid \text{weighted}]$$

which simplifies to $\frac{1}{2}(\frac{1}{6})^2 + \frac{1}{2}(\frac{1}{2})^2 = \frac{5}{36}$.

This is not equal to $\Pr[A_1] \cdot \Pr[A_2] = \frac{1}{9} = \frac{4}{36}$, so A_1 and A_2 **are not independent**.

We call A_1 and A_2 **conditionally independent** given that the die is fair and also conditionally independent given that the die is weighted.

Deductions

From the calculations we've done so far, we can conclude

$$\Pr[A_2 \mid A_1] = \frac{\Pr[A_1 \cap A_2]}{\Pr[A_1]} = \frac{5/36}{1/3} = \frac{5}{12}.$$

If we see that the first roll is ⚅, this boosts the probability that the second roll is ⚅, from $\frac{1}{3} = \frac{4}{12}$ to $\frac{5}{12}$.

Intuition: of course, there's never a direct causal effect that causes the second roll to be more like the first roll! Seeing that the first roll is ⚅ makes it more likely that we're using the weighted die, which makes it more likely that the second roll will also be ⚅.

We'll discuss how to make such deductions simpler in the next lecture.

Bayes' rule

Bayes' rule itself is just a short formula: for any events A and B,

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

It's interesting in the same setting as the previous lecture.

- In this setting, we have a partition $B_1, B_2, \ldots, B_n$ of hypotheses, each of which says A will happen with some probability $\Pr[A \mid B_i]$.
- The law of total probability lets us combine these to find an overall probability $\Pr[A]$.
- Bayes' rule can be used to go the other way: if A happens, compute probabilities $\Pr[B_i \mid A]$ telling us how much more or less likely this makes hypothesis B_i.

Example: rolling dice

Suppose we have a die that's either fair (hypothesis B_1) or lands ⚃ half the time (hypothesis B_2) with equal probability. We roll it, and roll a ⚃ (observation A).

Based on this, how likely is the die to be fair? What is $\Pr[B_1 \mid A]$?

Step 1. Use the law of total probability to find $\Pr[A]$.

$$\Pr[A] = \Pr[A \mid B_1]\Pr[B_1] + \Pr[A \mid B_2]\Pr[B_2] = \frac{1}{6}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{3}.$$

Step 2. Use Bayes' rule to find $\Pr[B_1 \mid A]$.

$$\Pr[B_1 \mid A] = \frac{\Pr[A \mid B_1]\Pr[B_1]}{\Pr[A]} = \frac{\frac{1}{6}\cdot\frac{1}{2}}{\frac{1}{3}} = \frac{1}{4}.$$

Example: noisy transmission

I transmit a message: either 000 (hypothesis B_1) or 111 (hypothesis B_2). Each bit of my message gets corrupted and flipped with probability $\frac{1}{10}$. You receive message 010 (observation A).

How likely is it that I meant 000? What is $\Pr[B_1 \mid A]$?

Step 1. Use the law of total probability to find $\Pr[A]$.

$$\Pr[A] = \Pr[A \mid B_1]\Pr[B_1] + \Pr[A \mid B_2]\Pr[B_2] = \frac{81}{1000}\cdot\frac{1}{2} + \frac{9}{1000}\cdot\frac{1}{2} = \frac{9}{200}.$$

(Note: we've assumed B_1 and B_2 are equally likely!)

Step 2. Use Bayes' rule to find $\Pr[B_1 \mid A]$.

$$\Pr[B_1 \mid A] = \frac{\Pr[A \mid B_1]\Pr[B_1]}{\Pr[A]} = \frac{\frac{81}{1000}\cdot\frac{1}{2}}{\frac{9}{200}} = \frac{9}{10}.$$

Prior probability

A weird thing about Bayes' rule is that it uses a **prior probability** $\Pr[B]$ to compute the **posterior probability** $\Pr[B \mid A]$. To know how likely the hypothesis is, we need to know how likely it was!

Here's a classic example where this matters. A rare disease is present in 0.1% of the population. You have a test for it that is only wrong 1% of the time (for both healthy and sick people). If the test is positive, how likely is the patient to have the disease?

Bayes' rule for $\Pr[\text{disease} \mid \text{test says so}]$ gives us

$$\frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.01 \cdot 0.999} \approx 0.09.$$

Even with a positive test, the patient is probably fine. The test is wrong much more often than people actually have the disease!

From probability to odds

We say that quantities $x_1, x_2, \ldots, x_n$ are in a **ratio** of $r_1 : r_2 : \cdots : r_n$ if there is a constant C such that $x_1 = Cr_1, x_2 = Cr_2, \ldots, x_n = Cr_n$.

These are often used for probabilities, when they're called **odds**. If $\Pr[⚄] = \frac{1}{6}$ and $\Pr[\text{not } ⚄] = \frac{5}{6}$, we say that the **odds of rolling** ⚄ are

$$\Pr[⚄] : \Pr[\text{not } ⚄] = \frac{1}{6} : \frac{5}{6} = 1 : 5.$$

This is very mildly convenient for us: it can make numbers nicer. If the odds of an event A are $p : q$, we can go back to $\Pr[A]$ by finding $\frac{p}{p+q}$.

If we are given the ratio $\Pr[B_1] : \Pr[B_2] : \cdots : \Pr[B_n]$ **and** $B_1, \ldots, B_n$ **are a partition**, we can go back to probabilities in the same way: divide through by the total.

Odds and Bayes' rule

Odds are convenient for Bayes' rule, because $\Pr[B_1 \mid A], \ldots, \Pr[B_n \mid A]$ all have the same denominator $\Pr[A]$ that we can skip computing.

In odds form, Bayes' rule says that the **posterior odds**

$$\Pr[B_1 \mid A] : \Pr[B_2 \mid A] : \cdots : \Pr[B_n \mid A]$$

are equal to the **prior odds**

$$\Pr[B_1] : \Pr[B_2] : \cdots : \Pr[B_n]$$

multiplied by the **likelihood ratio**

$$\Pr[A \mid B_1] : \Pr[A \mid B_2] : \cdots : \Pr[A \mid B_n].$$

Rolling dice, with odds

Suppose we have a die that's either fair (hypothesis B_1) or lands 🎲 half the time (hypothesis B_2) with equal probability. We roll it, and roll a 🎲 (observation A).

Based on this, how likely is the die to be fair? What is $\Pr[B_1 \mid A]$?

Step 1. Our prior odds are $1:1$ (the two hypotheses are equally likely).

Step 2. Our likelihood ratio is $\frac{1}{6} : \frac{1}{2}$ or $1:3$.

Step 3. We get posterior odds of $1:3$, so $\Pr[B_1 \mid A] = \frac{1}{1+3} = \frac{1}{4}$.

What if we roll 🎲 three times? We multiply by $1:3$ three times, getting posterior odds of $1:27$. The die is fair with probability $\frac{1}{1+27} = \frac{1}{28}$.

Many hypotheses

Suppose I have dice with 4, 6, 8, and 20 sides. I pick a random die, roll it once, and roll a 1, then a 5. Which die was I rolling with what probability?

- We start with prior odds of $1:1:1:1$.
- Multiply by $\frac{1}{4}:\frac{1}{6}:\frac{1}{8}:\frac{1}{20} = 30:20:15:6$.
- Multiply by $0:\frac{1}{6}:\frac{1}{8}:\frac{1}{20} = 0:20:15:6$.
- Posterior odds are $0:400:225:36$ so for example the probability that I was rolling the 20-sided die is $\frac{36}{400+225+36} = \frac{36}{661}$.

Odds calculations are convenient when we have many hypotheses, or many observations to update on (or both). Also, the odds (but **not** the probabilities) will still be correct if there are some missing hypotheses!

Conditional independence

We've seen a bit of conditional independence already. We say:

- Events A_1, A_2 are independent if $\Pr[A_1 \cap A_2] = \Pr[A_1] \Pr[A_2]$.
- Events A_1, A_2 are conditionally independent given B if $\Pr[A_1 \cap A_2 \mid B] = \Pr[A_1 \mid B] \Pr[A_2 \mid B]$.

Events can be conditionally independent in some cases, yet not independent. If we roll a die of unknown fairness twice, "first roll ⚅" and "second roll ⚅" are conditionally independent given that the die is weighted (or given that it's fair). They're not independent.

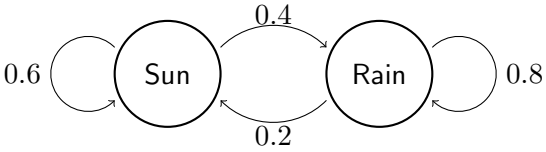
The reverse can also happen. If we roll a fair die twice, "first roll ⚅" and "second roll ⚅" are independent. They are not conditionally independent given that the total is 7.

Traditional example of a Markov chain

Suppose that the weather every day depends on the weather the previous day:

- If it was sunny today, there is a 60% chance it will be sunny tomorrow, and a 40% chance it will rain.
- If it rains today, there is a 20% chance it will be sunny tomorrow, and a 80% chance it will rain.

It's easier to think about this with a diagram:



Conditional independence in weather

- Weather two days apart is not independent: the weather on Monday affects the weather on Tuesday affects the weather on Wednesday.
- Even weather on days that are a year apart is not independent! (It is **very very close** to independent.)
- We have a conditional independence: the weather on Monday is independent from the weather on Wednesday **given that it rained on Tuesday**.

In general: given the weather on day t, an event about the future (days $t+1, t+2, \dots$) is independent from an event about the past (days $t-1, t-2, \dots$).

Markov chains

A **Markov chain** is a random process like this one, often described by a state diagram like the one we drew.

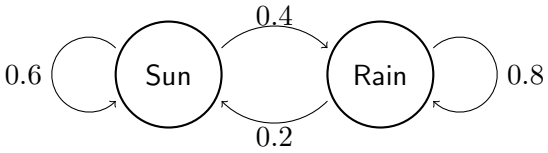
The property "given what happens at time t, events about the future are conditionally independent from events about the past" is the **Markov property**, and it is what defines a Markov chain.

In particular, we often have a set of states $\{1, 2, \ldots, n\}$, and a rule "if you're in state i at time t, then you go to state j at time $t+1$ with probability p_{ij}".

This rule can only exist because, knowing the state at time t, nothing about the past can influence the state at time $t+1$.

Predicting the future

Suppose that it's sunny on Monday. What is the probability it will rain on Friday?



A bad way to solve the problem is to think about all the possible cases for the whole week:

- $\Pr[\text{Sun} \to \text{Sun} \to \text{Sun} \to \text{Sun} \to \text{Rain}] = 0.6 \cdot 0.6 \cdot 0.6 \cdot 0.4.$
- $\Pr[\text{Sun} \to \text{Rain} \to \text{Rain} \to \text{Sun} \to \text{Rain}] = 0.4 \cdot 0.8 \cdot 0.2 \cdot 0.4.$
- There are six more cases...

Law of total probability

We can use the law of total probability to predict one day's weather in terms of the previous day's prediction.

Let $A_{\text{sun}}, A_{\text{rain}}$ be events for day t, and $B_{\text{sun}}, B_{\text{rain}}$ be events for day $t+1$. Then:

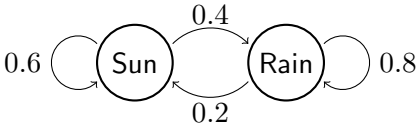
$$\begin{aligned}\Pr[B_{\text{sun}}] &= \Pr[B_{\text{sun}} \mid A_{\text{sun}}]\Pr[A_{\text{sun}}] + \Pr[B_{\text{sun}} \mid A_{\text{rain}}]\Pr[A_{\text{rain}}] \\ &= 0.6\Pr[A_{\text{sun}}] + 0.2\Pr[A_{\text{rain}}].\end{aligned}$$

Similarly,

$$\Pr[B_{\text{rain}}] = 0.4\Pr[A_{\text{sun}}] + 0.8\Pr[A_{\text{rain}}].$$

Predicting the future, take 2

Suppose that it's sunny on Monday. What happens on Friday?



If (s_n, r_n) are the probabilities of Sun and Rain on day n, then

$$(s_{n+1}, r_{n+1}) = (0.6s_n + 0.2r_n, 0.4s_n + 0.8r_n).$$

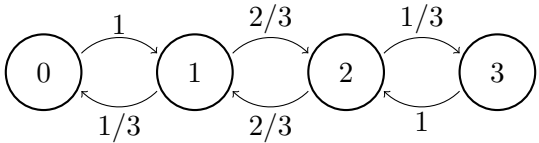
We use this recurrence to fill out a table:

	Monday	Tuesday	Wednesday	Thursday	Friday
Pr[sun]	1	0.6	0.44	0.376	0.3504
Pr[rain]	0	0.4	0.56	0.624	0.6496

Three cards

We start with three cards face down. Every minute, we take a random card and flip it. How does this system behave over time?

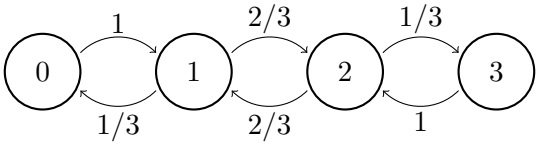
Here's a diagram (we represent each state by the number of face-up cards)



Suppose that at some point the four states have probability (a, b, c, d). The next minute the probabilities are:

$$(\tfrac{1}{3}b, a + \tfrac{2}{3}c, \tfrac{2}{3}b + d, \tfrac{1}{3}c).$$

Three cards: predicting the future



Let's make a table (this one is flipped around):

	$\Pr[0]$	$\Pr[1]$	$\Pr[2]$	$\Pr[3]$
$t = 0$	1	0	0	0
$t = 1$	0	1	0	0
$t = 2$	$1/3$	0	$2/3$	0
$t = 3$	0	$7/9$	0	$2/9$
$t = 4$	$7/27$	0	$20/27$	0
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

The Monty Hall problem

The problem:

- You're on a game show, and you win a prize. You're given a choice between three doors.

  Behind one door is a car. Behind the other two doors are goats.

- You pick door #1, but then the host (who knows what's behind each door) opens door #3, revealing one of the goats.

  The host asks, "Would you like to switch to door #2?"

What are your chances of winning the car if you open door #1? What are your chances if you switch to door #2?

Some background

We need to make some assumptions:

- The host always opens a door you didn't pick with a goat behind it, and offers a chance to switch.
- When the door you picked has a car behind it, the host chooses one of the other two doors with equal probability.

A famous columnist, Marilyn vos Savant, wrote about this puzzle in a newspaper, and said that you win the car with probability $\frac{2}{3}$ if you switch.

This seemed ridiculous to people, so many of them (including lots of mathematicians with Ph.D.'s) wrote letters arguing it was obviously $\frac{1}{2}$.

They were all wrong, and Marilyn was right.

Calculation with odds and Bayes' rule

- Initially, the three doors are equally likely to have a car. The prior odds are $1:1:1$.
- When you pick door #1, the host opens door #3 with a probability of

$$\begin{cases} \frac{1}{2} & \text{if the car is behind door \#1} \\ 1 & \text{if the car is behind door \#2} \\ 0 & \text{if the car is behind door \#3} \end{cases}$$

- So the posterior odds, given that the host opens door #3, are $\frac{1}{2}:1:0$ or $1:2:0$. Your chances with door #2 are $\frac{2}{3}$.

Objections and complications

- Wrong intuition: nothing has happened to distinguish door #1 and door #2, so the probability should be $\frac{1}{2}$.

  Door #2 has "survived a test" that door #1 hasn't: the host chose not to open it.

- Right intuition: if you play many times and never switch, you win whenever the car is behind door #1. If you always switch, you win whenever the car is behind door #2 **or** door #3.

- If the game show host does not always offer the chance to switch, or if the host has a preference between the doors, the probabilities may be different.

Two similar questions

Note: the standard treatment of this problem assumes children are boys with probability $\frac{1}{2}$ and girls with probability $\frac{1}{2}$ with no other possibilities.

Question 1. Mr. Jones has two children. You ask him, "Is the older child a girl?" He says yes.

What is the probability that the other is also a girl?

Answer 1. The probability is $\frac{1}{2}$: the two children are independent.

Question 2. Mr. Smith has two children. You ask him, "Is at least one a girl?" He says yes.

What is the probability that the other is also a girl?

Answer 2. Here, the probability changes to $\frac{1}{3}$. Why?

The work for question 2

I've asked the least confusing version of this second question.

Here, we can start with the sample space for a two-child family:

$$S = \{(\mathrm{B},\mathrm{B}), (\mathrm{B},\mathrm{G}), (\mathrm{G},\mathrm{B}), (\mathrm{G},\mathrm{G})\}.$$

We assume all 4 outcomes are equally likely.

When we ask "Is at least one child a girl?" and Mr. Smith says yes, all we learn is that (B,B) did not occur. This leaves us with

$$S' = \{\qquad\quad (\mathrm{B},\mathrm{G}), (\mathrm{G},\mathrm{B}), (\mathrm{G},\mathrm{G})\}.$$

Of these three outcomes, **one** is (G,G), so the probability that both children are girls is $\frac{1}{3}$.

Assumptions about the random experiment

As with Monty Hall, it is crucial to know how likely we are to make each observation, under each hypothesis. Thus:

Question 3. Mr. Smith tells you "I have two children, and at least one is a girl." What is the probability that the other is also a girl?

Answer 3. Hard to say without knowing Mr. Smith's psychology.

Question 4. Mr. Smith has two children. You see one of them, a girl, playing outside. What is the probability that the other is also a girl?

Answer 4. It's $\frac{1}{2}$, assuming you see a **randomly chosen child**.

The key is that if both children are girls, the probability you see a girl is 1. If the other child is a boy, the probability you see a girl is only $\frac{1}{2}$.

A final variant

Question 5. Mr. Smith has two children. At least one is a girl born on a Tuesday. (You asked Mr. Smith, "Is at least one of your children a girl born on a Tuesday?" and he said yes.)

What is the probability that the other child is also a girl?

Answer. Here, we need to invoke the power of Bayes' rule!

The prior odds of the four outcomes BB, BG, GB, GG are $1:1:1:1$.

The likelihood ratio is $0:\frac{1}{7}:\frac{1}{7}:\frac{13}{49}$ or $0:7:7:13$. (Check this!)

The posterior odds are also $0:7:7:13$.

Therefore $\Pr[\text{GG} \mid \text{Mr. Smith said yes}] = \frac{13}{7+7+13} = \frac{13}{27}$.

The plan for counting

Lots of probability problems are sampling uniformly from a sample space S. Remember that in this case, $\Pr[A] = \frac{|A|}{|S|}$. So being able to count the number of elements is a set is very important!

We will:

1. Go over some techniques we'll need for counting.
2. Talk about sampling **with** or **without** replacement...
3. ...and taking **ordered** or **unordered samples**.

Each of the four cases has its own formula!

Multiplication principle

If we can specify an element of S in k stages, and there are n_i options at the i^{th} stage, then $|S| = n_1 \cdot n_2 \cdots n_k$.

Examples:

- The number of cards in a standard deck is

$$(4 \text{ suits}) \cdot (13 \text{ ranks}) = 52.$$

- The number of ways to deal one card to Alice and one to Bob is

$$(52 \text{ options for Alice}) \cdot (51 \text{ options for Bob}) = 2652.$$

Related to independent events, where $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$, but more general.

Example: counting subsets

How many subsets does the set $\{1, 2, \ldots, n\}$ have?

We can specify a subset $A \subseteq \{1, 2, \ldots, n\}$ in n stages:

1. First, specify whether $1 \in A$ or $1 \notin A$ (2 options).
2. Second, specify whether $2 \in A$ or $2 \notin A$ (2 options).
3. And so on. For each $k \leq n$, specify whether $k \in A$ or $k \notin A$ (2 options).

At each stage we have 2 options, so the number of subsets is

$$\underbrace{2 \cdot 2 \cdot 2 \cdots 2}_{n \text{ factors}} = 2^n.$$

Example: counting 4-digit numbers

How many **odd** 4-digit integers are there? We can solve this problem by counting the digits one at a time.

1. There are 9 options for the first digit: $1, 2, 3, 4, 5, 6, 7, 8, 9$.
2. The second digit has 10 options; it can also be 0.
3. The third digit also has 10 options.
4. The fourth digit can be one of $1, 3, 5, 7, 9$: only 5 options.

The overall number is $9 \cdot 10 \cdot 10 \cdot 5 = 4500$.

We could also do this by taking $\frac{9999-1000+1}{2} = 4500$, to get half of the range $\{1000, 1001, 1002, \ldots, 9999\}$. But watch out: a range of the form $\{a, a+1, \ldots, b\}$ contains $b - a + 1$ numbers!

The handshake problem

There are 100 people at a party. Each of them shakes hands once with every other person. How many handshakes is that?

We could try to solve this using the multiplication principle:

1. There are 100 ways to pick a person at the party.
2. There are 99 ways to pick someone they shake hands with.

This gives 9900.

That's not the right answer! We counted each handshake twice. The handshake between Alice and Bob was counted once by picking Alice first, Bob second, and again by picking Bob first, Alice second.

We can fix this. Since each handshake is counted twice, dividing by 2 gives the right number of handshakes: $\frac{99 \cdot 100}{2} = 4950$.

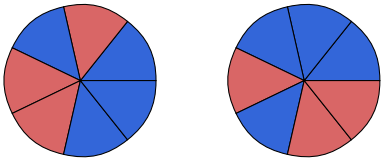
Ways to count

The basic way to count things in combinatorics is the multiplication principle. We have several fancier ways to count, which boil down to "count the wrong thing, then fix it".

- **Casework:** split the problem into several disjoint cases, and add up the number for each case.
- **Complementary counting:** To count the "good things", count everything, then subtract the number of "bad things".
- **Inclusion-exclusion:** split the problem into two cases, add up the number for each case, then subtract the overlap.
- **Overcounting:** Use an approach that counts everything k times, then divide by k.

Coloring a disk

We divide a disk into 7 equal parts, and want to color them red or blue. How many ways are there to do it?



As stated, there are just $2^7 = 128$ ways: for each part, we pick either red or blue.

What if we consider the two colorings above to be the same, because we can rotate one to get the other?

Coloring a disk: careful overcounting

Imagine that there's an unknown pile of all the **things we want to count**, and a known list of all the **things we did count**. Here, that list has 128 pictures on it; earlier, it had 9900 handshakes on it.

How many times does a coloring appear in the list?

- For most colorings, that number is 7: there are 7 pictures of that coloring that are only different by rotation.
- Exception: the all-blue and all-red coloring only have 1 picture!

Set those 2 pictures aside. Divide the remaining 126 pictures on the list by 7 to get $\frac{126}{2} = 18$. Then add the 2 pictures back to get 20 total colorings.

Sampling with replacement

Ordered sampling with replacement means we choose a sequence of several elements uniformly at random from S, and repeats are allowed.

Standard example: flipping coins, or rolling dice.

Let's look at a familiar problem: if we roll n dice, what is the probability of at least one ⚅?

Easier to count the reverse probability: that we roll n dice with no ⚅.

- Total number of outcomes: 6^n.
- Number of outcomes with no ⚅: 5^n.

So the answer is $1 - \frac{5^n}{6^n}$.

Hash functions and hackers

Secure messages are digitally signed by a **hash** of that message: a number calculated from that message, but intended to be as random as possible.

A hacker that wants to forge a message with your signature would need to generate a forged message that has the same hash.

Say that there are $n = 1\,000\,000$ possible hashes (Not very secure, but whatever) and a hacker has time to try $k = 1\,000$ forgeries. What is the hacker's probability of success?

Same problem! There are n^k total outcomes, and $(n-1)^k$ of them don't work, so

$$\Pr[\text{success}] = 1 - \frac{(n-1)^k}{n^k} = 1 - 0.999999^{1000} \approx 0.001.$$

Counting permutations

A **permutation** of a set S is a sequence that puts the elements of S in some order, with no repeats. The 6 permutations of $\{1, 2, 3\}$ are:

$$(1, 2, 3) \quad (1, 3, 2) \quad (2, 1, 3) \quad (2, 3, 1) \quad (3, 1, 2) \quad (3, 2, 1)$$

How many permutations does $\{1, 2, \ldots, n\}$ have?

- Choosing the elements one at a time, we have n ways to choose the first element.
- But there are only $n - 1$ options for the second element: no repeats!
- There are $n - 2$ options for the third, and so on. For the last element, only one option is left.

Factorials

By multiplying, we see that the number of permutations of $\{1, 2, \ldots, n\}$ (or of any n-element set) is

$$n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1.$$

This is called "n factorial" and written $n!$ for short.

About how big is n factorial? Here are some easy upper and lower bounds:

- We multiply together n numbers that are at most n, so $n! \leq n^n$.
- The largest $\frac{n}{2}$ numbers are at least $\frac{n}{2}$, so $n! \geq (\frac{n}{2})^{n/2}$.

It turns out, but it's harder to prove, that $n!$ is closer to $(\frac{n}{e})^n$, where $e \approx 2.718$ is Euler's number.

Ordered sampling without replacement

Ordered sampling without replacement means that we choose a sequence of elements from S, but repeats are forbidden.

Standard example: drawing cards from a deck. (Or objects from a bag.)

In a 4-player card game, how many ways are there to deal one card to each player?

By the same argument as for factorial:

$$52 \cdot 51 \cdot 50 \cdot 49 = 5\,197\,920.$$

This formula generalizes to choosing an **ordered sequence** of k elements from an n-element set.

k-permutations

The number of ways to choose a sequence of k elements from an n-element set (sometimes called the number of "k-permutations" of that set) is

$$P_k^n = \underbrace{n \cdot (n-1) \cdots (n-k+2) \cdot (n-k+1)}_{k \text{ factors}}.$$

We can write a formula for this using factorial notation:

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

When k is small compared to n, **don't use this!**

The birthday paradox

Suppose that 18 students come to class. What is the probability that two of them have the same birthday? (Assume 365 equally likely days.)

It's easier to count the reverse probability: the number of ways to choose 18 **different** birthdays is P_{18}^{365}. Therefore

$$\Pr[\text{birthday collision}] = 1 - \Pr[\text{all distinct}] = 1 - \frac{P_{18}^{365}}{365^{18}} \approx 0.347$$

This is called the "birthday paradox" because the probability is surprisingly high. It takes only 23 students before the probability is more than 50%, even though 23 is much smaller than 365.

The birthday attack

Secure messages are digitally signed by a **hash** of that message: a number calculated from that message, but intended to be as random as possible.

In the **birthday attack**, a hacker generates and signs many messages until two of them have the same hash, then passes one of these messages off for the other.

Say there are $n = 1\,000\,000$ possible hashes and the hacker has time to sign $1\,000$ messages. What's the probability of a hash collision?

This is the same as the birthday paradox, but with different numbers.

$$\Pr[\text{hash collision}] = 1 - \Pr[\text{all distinct}] = 1 - \frac{P_k^n}{n^k} \approx 0.393.$$

Warm-up: subsets of $\{1, 2, 3, 4, 5\}$

How many subsets of size 3 does the set $S = \{1, 2, 3, 4, 5\}$ have?

Let's solve this by the overcounting technique.

1. We know how to count the "3-permutations": sequences of 3 distinct elements from S.

   There are $P_3^5 = 5 \cdot 4 \cdot 3 = 60$ of them.

2. Each subset of size 3 can be ordered in $3! = 6$ ways to get such a sequence: for example, $\{1, 3, 5\}$ can be ordered as

$$(1, 3, 5),\ (1, 5, 3),\ (3, 1, 5),\ (3, 5, 1),\ (5, 1, 3),\ (5, 3, 1).$$

3. So there are $\frac{60}{6} = 10$ subsets of size 3.

Unordered sampling without replacement

In general, number of ways to take k elements from a set of size n, **in order**, is

$$P_k^n = \underbrace{n(n-1)\cdots(n-k+1)}_{k \text{ factors}} = \frac{n!}{(n-k)!}.$$

If we just care about which things we take, and **not** about the order, then this counts every outcome $k!$ times, because every outcome can be ordered in $k!$ ways.

Therefore the number of unordered samples—the number of k-element subsets of an n-element set—is

$$\binom{n}{k} = \frac{P_k^n}{k!} = \frac{n!}{k!\,(n-k)!} \quad \text{read } \text{“}n \text{ choose } k\text{”}.$$

Typical probability problem

You draw a 5-card poker hand. What is the probability that all 5 cards are spades?

1. Let S be the set of all 5-card hands. Let A be the set of all 5-card hands that consist only of spades. We want to know $\Pr[A] = \frac{|A|}{|S|}$.
2. We can think of S as the set of all 5-element subsets of the 52-card deck. There are $\binom{52}{5} = 2\,598\,960$ subsets.
3. Similarly, A is also the set of all 5-element subsets, but of a smaller set: the set $\{\text{A}\spadesuit, 2\spadesuit, 3\spadesuit, \ldots, \text{Q}\spadesuit, \text{K}\spadesuit\}$. There are $\binom{13}{5} = 1\,287$ subsets.
4. We get $\Pr[A] = \frac{1287}{2598960} = \frac{33}{66640}$.

Ordered or unordered?

In many probability problems, it doesn't matter if we use $\binom{n}{k}$ or P_k^n, as long as we're consistent!

For example, we could solve the same problem about 5-spade hands using k-permutations and ordered samples instead.

1. Let S be the set of all **ordered** 5-card hands. Now $|S| = P_5^{52}$ instead of $\binom{52}{5}$: larger by a factor of $5! = 120$.
2. Let A be the set of all **ordered** 5-card hands that are all spades. Now $|A| = P_5^{13}$ instead of $\binom{13}{5}$: also larger by a factor of $5! = 120$.
3. We get $\Pr[A] = \frac{P_5^{13}}{P_5^{52}} = \frac{33}{66640}$, as before.

Important: we cannot mix and match! If our sample space is ordered, all our events must be ordered.

Ordered or unordered?

Sometimes problems are easier to set up in one case compared to the other case. Example: what is the probability that you draw a 5-card poker hand with $3 \times \spadesuit$ and $2 \times \heartsuit$?

- With unordered sampling, the answer is

$$\frac{\binom{13}{3} \cdot \binom{13}{2}}{\binom{52}{5}}.$$

- With ordered sampling, the answer is

$$\frac{P_3^{13} \cdot P_2^{13} \cdot \binom{5}{2}}{P_5^{52}}.$$

  We have to add an extra factor to account for the order of $\spadesuit$ and $\heartsuit$ within your hand.

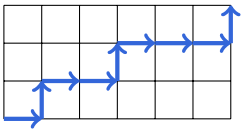
Binomial coefficient interpretations

We defined $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ to count the number of k-element subsets of $\{1, 2, \ldots, n\}$.

If we have two letters A and B, $\binom{n}{k}$ **also** counts the number of n-letter words made up of k letters A and $n-k$ letters B.

(The argument: there are $\binom{n}{k}$ ways to choose the positions of the A's.)

This is very flexible. For example, if our two letters are $\uparrow$ and $\rightarrow$, then a word with a $\rightarrow$'s and b $\uparrow$'s is a "lattice path" from $(0, 0)$ to (a, b).



Bernoulli trials

A **Bernoulli trial** is a biased coinflip: there is a probability p of heads and $1 - p$ of tails. (Or "success" and "failure", or whatever.)

Consider a random experiment where we flip such a biased coin n times.

There are two good models of this:

1. The sample space consists of all n-tuples of heads and tails.

   For an outcome x with k heads and $n - k$ tails, $\Pr[\{x\}] = p^k(1-p)^{n-k}$.

2. The sample space is $\{0, 1, 2, \ldots, n\}$: the possible values of the **number of heads** flipped.

   What is the probability $\Pr[\{k\}]$?

The binomial probability formula

A biased coin lands heads with probability p and tails with probability $1-p$. If we flip the coin n times, what is the probability that it lands heads k times?

1. We work in the first model, where each outcome is a sequence such as HTTHTHHHTHT.
2. The event $A =$ "flip H k times" contains $\binom{n}{k}$ outcomes. (The words with k H's and $n-k$ T's.)
3. Each outcome in A has probability $p^k(1-p)^{n-k}$, giving us

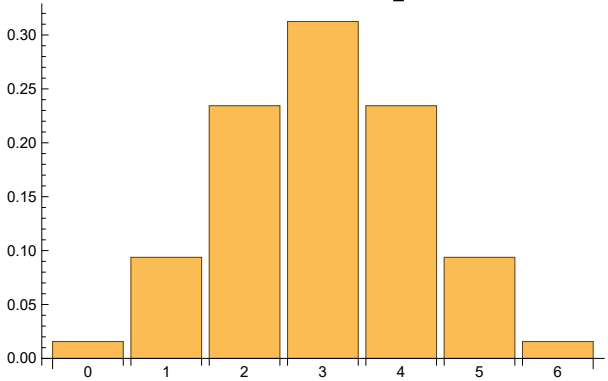
$$\Pr[A] = \binom{n}{k} p^k (1-p)^{n-k}.$$

This is the binomial probability formula.

A discrete bell curve

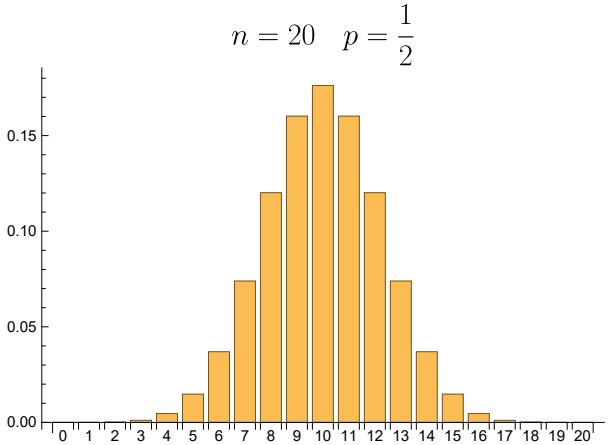
Here are some examples of what these probabilities look like.

$$n = 6 \quad p = \frac{1}{2}$$



A discrete bell curve

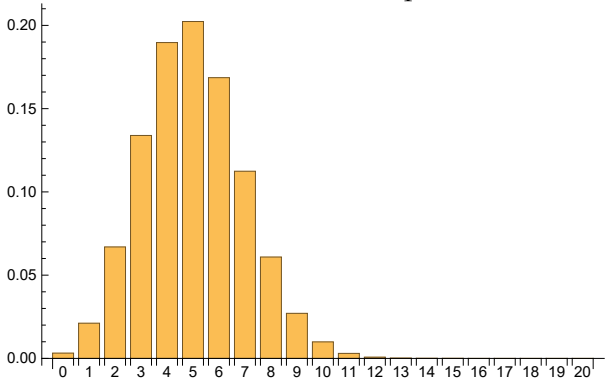
Here are some examples of what these probabilities look like.



A discrete bell curve

Here are some examples of what these probabilities look like.

$$n = 20 \quad p = \frac{1}{4}$$



Examples

- I flip a fair coin 5 times. What is the probability of seeing exactly 2 heads?

$$\Pr[\text{2 heads}] = \binom{5}{2}\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^3 = \frac{5!}{2!\,3!}\cdot\frac{1}{2^5} = \frac{5}{16}.$$

- I roll n fair dice. What is the probability of seeing **exactly one** ⚅?

$$\Pr[\text{one ⚅}] = \binom{n}{1}\left(\frac{1}{6}\right)^1\left(1-\frac{1}{6}\right)^{n-1} = \frac{n\cdot 5^{n-1}}{6^n}.$$

The anagram problem

What is the number of distinct anagrams of **ALFALFA**?

- There are $7! = 5040$ ways to rearrange the letters, if we ignore some of them being identical (or if we color the letters **ALFALFA**, so that they're all different).
- Each anagram is counted $3! \cdot 2! \cdot 2!$ times: there are $3!$ ways to color the **A**'s, $2!$ ways to color the **L**'s, and $2!$ ways to color the **F**'s.

So the total number of anagrams is

$$\frac{7!}{3! \cdot 2! \cdot 2!} = \frac{5040}{24} = 210.$$

Multinomial coefficients

In general, if there are k distinguishable types of objects, and n_i objects of the i^{th} type, the number of ways to order them is

$$\frac{(n_1 + n_2 + \cdots + n_k)!}{n_1!\, n_2! \cdots n_k!}.$$

This is called a **multinomial coefficient** and sometimes written

$$\binom{n_1 + n_2 + \cdots + n_k}{n_1, n_2, \ldots, n_k}.$$

For example, the number of anagrams of **ALFALFA** is $\binom{7}{3,2,2}$.

When $k = 2$, we just get back the ordinary binomial coefficient.

Generating random letters

A random letter generator gives the letter **A** with some probability p_{A}, **B** with some probability p_{B}, and so on for every letter.

If you ask it for 7 letters, what is the probability it will give you 3 **A**'s, 2 **L**'s, and 2 **F**'s in some order?

There are $\binom{7}{3,2,2}$ outcomes (all the anagrams of **ALFALFA**) and each of them happens with probability $(p_{\mathsf{A}})^3 \cdot (p_{\mathsf{L}})^2 \cdot (p_{\mathsf{F}})^2$, so

$$\Pr[3\ \textbf{A}\text{'s},\ 2\ \textbf{L}\text{'s},\ 2\ \textbf{F}\text{'s}] = \binom{7}{3,2,2}(p_{\mathsf{A}})^3(p_{\mathsf{L}})^2(p_{\mathsf{F}})^2.$$

This generalizes! See next slide.

Multinomial formula

Suppose we are sampling from $S = \{1, 2, \ldots, k\}$ (or any other k-element set) and $\Pr[\{i\}] = p_i$.

What is the probability of seeing n_1 1's, n_2 2's, $\ldots$, n_k k's after $n = n_1 + n_2 + \cdots + n_k$ trials?

The multinomial formula says that the probability is

$$\binom{n}{n_1, n_2, \ldots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} = \frac{n!}{n_1!\, n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}.$$

In particular, if we're sampling uniformly, then $p_1 = p_2 = \cdots = p_k = \frac{1}{k}$, and the probability is:

$$\binom{n}{n_1, n_2, \ldots, n_k} \frac{1}{k^n}.$$

Unordered sampling with replacement

In **unordered sampling with replacement**, we take many samples from a fixed set $S = \{x_1, x_2, \ldots, x_k\}$, but all we're interested in is the **number of times** we see each outcome.

For example: imagine rolling 100 dice, and counting the ⚀'s, ⚁'s, ⚂'s, ⚃'s, ⚄'s, ⚅'s.

- The multinomial formula tells us the probability of each outcome. For example, when rolling 100 dice, the probability of ⚀ $\times 20$, ⚁ $\times 20$, ⚂ $\times 15$, ⚃ $\times 15$, ⚄ $\times 18$, and ⚅ $\times 12$ is:

$$\binom{100}{20, 20, 15, 15, 18, 12} \frac{1}{6^{100}}.$$

- What if we want to know how many outcomes there are?

Counting multisets

The problem: in how many ways can we split n indistinguishable objects into k groups?

- Last slide: counting the possible outcomes when we roll $n = 100$ dice. (Here, $k = 6$, the number of outcomes.)
- Number of ways k pirates can split a pile of n pieces of gold.
- Number of solutions to $x_1 + x_2 + \cdots + x_k = n$ for nonnegative integers $x_1, x_2, \ldots, x_k$.
- Number of terms in the expansion of $(x_1 + x_2 + \cdots + x_k)^n$.

These are sometimes called **multisets** of n elements from $\{1, 2, \ldots, k\}$; these are not sets, because the same element can appear many times.

The stars and bars method

Our strategy:

1. Represent each multiset by a configuration of two symbols.
2. Count the number of possible configurations.

For example, if we roll 10 dice, one possible outcome is:

$$2 \times ⚀ \quad 1 \times ⚁ \quad 0 \times ⚂ \quad 1 \times ⚃ \quad 4 \times ⚄ \quad 2 \times ⚅$$

We represent that as

$$★★|★|\ |★|★★★★|★★$$

In general, we separate 10 ★'s into 6 groups by adding 5 |'s.

Counting the configurations

Each configuration such as

$$\bigstar\,\bigstar\,|\,\bigstar\,|\,\,|\,\bigstar\,|\,\bigstar\,\bigstar\,\bigstar\,\bigstar\,|\,\bigstar\,\bigstar$$

represents a different outcome when we roll 10 dice. How many configurations are there?

We have 15 total symbols: 10 $\bigstar$'s, and 5 $|$'s. The number of ways to rearrange them is $\binom{15}{5}$ or $\binom{15}{10}$.

In general, if we want to split n objects into k groups, we will have n $\bigstar$'s and $k-1$ $|$'s, for $n+k-1$ total symbols. The number of ways to rearrange them is

$$\binom{n+k-1}{k-1} = \binom{n+k-1}{n}.$$

Warning: do not use

Warning: Usually, the formula $\binom{n+k-1}{k-1}$ should **not** be used for probabilities, because usually, the distribution is **not uniform**!

Example: if we roll 100 dice, what is the probability of no ⚅'s?

- We know the probability is $(\frac{5}{6})^{100}$, because we've already done this several times.
- There are $\binom{105}{5}$ total outcomes, and $\binom{104}{4}$ outcomes with no ⚅'s.
  It would be wrong to say that the probability is $\frac{\binom{104}{4}}{\binom{105}{5}}$, because the distribution is not uniform.

In fact, assuming uniformity badly overestimates the probability. Outcomes with no ⚅'s are very unbalanced and therefore very unlikely.

Summary of sampling methods

Suppose we are taking s samples from a set with N elements. The distribution depends on whether the samples are ordered or unordered, and whether we are sampling with or without replacement:

Sampling type	**# outcomes**	**Pr**
Ordered without replacement	$P_s^N = \frac{N!}{(N-s)!}$	uniform
Ordered with replacement	N^s	uniform
Unordered without replacement	$\binom{N}{s} = \frac{N!}{s!(N-s)!}$	uniform
Unordered with replacement	$\binom{N+s-1}{N-1}$ or $\binom{N+s-1}{s}$	**multinomial**

A familiar problem

You draw 3 marbles (without replacement) from a bag containing $4 \times$ 🔵 and $6 \times$ 🔴. What is the probability of drawing at least 2 🔵?

We can solve this without branching diagrams, using binomial coefficients instead!

$$\Pr[3 \times \text{🔵}] = \frac{\binom{4}{3}}{\binom{10}{3}} = \frac{4}{120} = \frac{1}{30}.$$

$$\Pr[2 \times \text{🔵}, 1 \times \text{🔴}] = \frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}} = \frac{6 \cdot 6}{120} = \frac{3}{10}.$$

$$\Pr[\text{at least } 2 \times \text{🔵}] = \frac{1}{30} + \frac{9}{30} = \frac{1}{3}.$$

Marbles and conditional probability

Starting from the same bag of $4 \times$ ● and $6 \times$ ●, we scoop out a smaller bag of 3 marbles without looking.

By what we've done so far (and two more similar cases), we know that

$$\Pr[3 \times \bullet] = \frac{1}{30} \qquad \Pr[1 \times \bullet, 2 \times \bullet] = \frac{1}{2}$$
$$\Pr[2 \times \bullet, 1 \times \bullet] = \frac{3}{10} \qquad \Pr[3 \times \bullet] = \frac{1}{6}$$

Questions:

1. What's the probability that we draw ● from the smaller bag?
2. Given that we draw ●, what are the conditional probabilities of the bag's contents?

Law of total probability

By the law of total probability:

$$\Pr[\text{draw } \bullet] = \sum_{k=0}^{3} \Pr[\text{draw } \bullet \mid \text{bag has } k \times \bullet] \cdot \Pr[\text{bag has } k \times \bullet].$$

In our case:

$$\Pr[\text{draw } \bullet] = 0 \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{3}{10} + 1 \cdot \frac{1}{30} = \frac{2}{5}.$$

There was a shortcut to get here!

Scooping out 3 marbles from the large bag, then drawing a marble from the small bag should be the same as just drawing from the large bag.

The large bag has $4 \times$ $\bullet$ and $6 \times$ $\bullet$ for a probability of $\frac{4}{4+6} = \frac{2}{5}$.

Bayes' rule

Now, let's answer the harder question, which requires Bayes' rule.

There are four possibilities for the bag's contents: $0 \times$ ● through $3 \times$ ● and the rest ●. We'll write all odds in that order.

1. Prior odds: $\frac{1}{6} : \frac{1}{2} : \frac{3}{10} : \frac{1}{30}$ or $5 : 15 : 9 : 1$.
2. Bayes factor: $0 : \frac{1}{3} : \frac{2}{3} : 1$ or $0 : 1 : 2 : 3$.
3. Posterior odds: $0 : \frac{1}{6} : \frac{1}{5} : \frac{1}{30}$ or $0 : 15 : 18 : 3$ or $0 : 5 : 6 : 1$.
4. Conditional probabilities: 0, $\frac{5}{12}$, $\frac{1}{2}$, and $\frac{1}{12}$.

Without the odds version (let's just do one case):

$$\Pr[3 \times \bullet \mid \text{draw } \bullet] = \frac{\Pr[\text{draw } \bullet \mid 3 \times \bullet] \cdot \Pr[3 \times \bullet]}{\Pr[\text{draw } \bullet]} = \frac{1 \cdot \frac{1}{30}}{\frac{2}{5}} = \frac{1}{12}.$$

Disjoint subsets

We know there are 2^n subsets of $\{1, 2, \ldots, n\}$. Let's look at extensions.

1. How many ways are there to choose two subsets $A \subseteq \{1, 2, \ldots, 10\}$ and $B \subseteq \{1, 2, \ldots, 10\}$ such that $A \cap B = \varnothing$?

   For every element $k \in \{1, 2, \ldots, 10\}$, we have three choices. Either $k \in A$, or $k \in B$, or neither is true. (Both can't be true!)

   Multiplying, we get $3 \cdot 3 \cdot 3 \cdots 3 = 3^{10}$ choices.

2. How many of these pairs (A, B) have $|A| = 3$?

   We can go in stages: first choose A, then choose B. There are $\binom{10}{3}$ ways to choose A.

   B is a subset of the remaining 7 elements. There are 2^7 ways to pick B, giving us $\binom{10}{3} \cdot 2^7$ ways to choose A and B.

Sets and multinomials

How many ways are there to choose two subsets $A \subseteq \{1, 2, \ldots, 10\}$ and $B \subseteq \{1, 2, \ldots, 10\}$ such that $A \cap B = \varnothing$, $|A| = 3$, and $|B| = 5$?

For each element $k \in \{1, 2, \ldots, 10\}$, there are 3 options. However, the choices affect each other. We are supposed to choose $k \in A$ 3 times, choose $k \in B$ 5 times, and choose neither $10 - 3 - 5 = 2$ times.

This is counted by the multinomial coefficient $\binom{10}{3,5,2}$.

To see this, imagine writing down a 10-letter "word" whose k^{th} position is **A** if $k \in A$, **B** if $k \in B$, and **X** otherwise. For example:

$$A = \{1, 3, 5\}, B = \{2, 7, 8, 9, 10\} \rightsquigarrow \textbf{ABAXAXBBBB}.$$

This is an anagram of **AAABBBBBXX** encoding our choice of A and B, and there are $\binom{10}{3,5,2}$ anagrams.

Inclusion-exclusion principle

How many ways are there to choose subsets $A \subseteq \{1, 2, \ldots, 10\}$ and $B \subseteq \{1, 2, \ldots, 10\}$ such that $A \cap B = \varnothing$, **but neither A nor B is empty**?

We can use the inclusion-exclusion principle:

1. Total number of ways: 3^{10}, as seen earlier.
2. Number of ways where $A = \varnothing$: $\binom{10}{0}2^{10}$ or just 2^{10}.
3. Number of ways where $B = \varnothing$: also 2^{10}, by symmetry.
4. Number of ways where $A = B = \varnothing$: just 1.
5. Answer: $(1) - (2) - (3) + (4) = 3^{10} - 2 \cdot 2^{10} + 1 = 57\,002$.

What we can and can't simplify

Suppose we toss a fair coin 100 times.

- What is the probability that the coin lands heads **exactly** 10 times?

  This is $\binom{100}{10}(\frac{1}{2})^{100}$, by the binomial probability formula.

- What is the probability that the coin lands heads **at most** 10 times?

  This is the sum

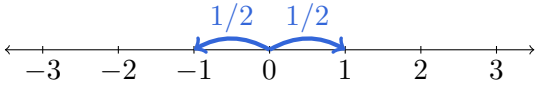
$$\frac{\binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{10}}{2^{100}}.$$

  In almost every case, a sum like this doesn't simplify any further.

Random walk on the number line

Two similar questions:

- You stand at 0 on the number line. Every second, you go left or right with equal probability. What is the probability that after 100 seconds, you'll be at 10?



- You are betting on a coin toss, winning or losing $\$1$ with equal probability. What is the probability that after 100 bets, you have exactly $\$10$ more than you started with?

In both cases, to get from 0 to 10 in 100 steps, we want 55 successes and 45 failures. This has probability $\binom{100}{55}(\frac{1}{2})^{100}$.

Binomial coefficients and conditional probability

We roll a fair die 10 times. Given that we rolled 3 ⚃'s, what is the probability that the first roll was a ⚃?

Intuitively: $\frac{3}{10}$. Now let's make sure complicated math gives us the common-sense answer.

Using Bayes' rule:

$$\Pr[\text{first } ⚃ \mid \text{three } ⚃\text{'s}] = \frac{\Pr[\text{three } ⚃\text{'s} \mid \text{first } ⚃] \cdot \Pr[\text{first } ⚃]}{\Pr[\text{three } ⚃\text{'s}]}.$$

This simplifies to

$$\Pr[\text{first } ⚃ \mid \text{three } ⚃\text{'s}] = \frac{\binom{9}{2}(\frac{1}{6})^2(\frac{5}{6})^7 \cdot \frac{1}{6}}{\binom{10}{3}(\frac{1}{6})^3(\frac{5}{6})^7} = \frac{\binom{9}{2}}{\binom{10}{3}} = \frac{36}{120} = \frac{3}{10}.$$