Probability

Many events can't be predicted with certainty. We use the tools of probability to quantify how likely they are to happen. To reason about probabilities, we'll use a data type called a *probability space*. We'll build up this data type over the course of this set of lecture notes.

Atomic events

We start from a *universe* or *sample space*, a set U of mutually exclusive and exhaustive possible outcomes. The elements of U are called *simple* or *atomic* events. Often these simple events will be interpretable: e.g., which face of a die lands on top when we roll it (like this: ⚄), or which card we draw from a deck (like this: 7♠). We assign probabilities to atomic events: $P(a)$ is the probability that $a \in U$ will occur.

> *For our purposes, it's best to think of U as finite. If we want to deal with continuous outcomes, a good way to think about them is to discretize: pretend that there are a very large but finite number of events, spaced out so that there is always one very close to any possible continuous outcome. Since this viewpoint is for understanding rather than computation, there's no limit on how big we can make U: e.g., we could have one event for every 64-bit floating point number, or even every tuple of 100 64-bit numbers. Discretization is actually better in many cases than trying to deal directly with continuous outcomes: probability on continuous spaces holds a lot of mathematical traps, such as hidden infinities and counterintuitive behaviors.*

We'll take the viewpoint that probabilities are *subjective*: I can assign any probabilities I want, so long as I follow the rules described here for manipulating them. And I can assign probabilities to any outcomes that I can imagine measuring, no matter whether I think of the process that determines the outcomes as being random.

> *What does the statement $P(a) = \frac{1}{3}$ mean? One common interpretation is that, if I reset the world and ran it forward many times, measuring whether event a happened in each trial, I'd see it happen in a third of them. This interpretation makes sense for something like a biased coin that we can flip many times. But it is not necessarily helpful in general: first, it may not really be possible to reset the world, even approximately. Second, we might want to assign a probability to an event that is deterministic but unknown to us, so that repeated trials don't yield independent measurements. Instead, in such cases we interpret the statement as a measure of my subjective confidence in a: I'd be willing to take $2:1$ odds in a bet on a, which is the same odds that would have me break even if betting on a biased coin with $P(H) = \frac{1}{3}$.*

Probabilities are always nonnegative; an impossible event has probability 0. Probabilities always sum to 1 over the universe:

$$\sum_{e \in U} P(e) = 1$$

That means that each individual event has probability at most 1. An event with probability 1 is certain to happen, since the sum-to-1 rule means that all other events must have probability zero.

> *In continuous spaces, we measure probability density instead of probability. Probability density differs from probability in a number of important ways; see below for more information.*

Compound events

A compound event is a subset of the universe, $E \subseteq U$. The probability of E is the sum of the probabilities of atomic events in E:

$$P(E) = \sum_{a \in E} P(a)$$

The sum-to-1 rule means that we are certain to get some event in the universe:

$$P(U) = \sum_{a \in U} P(a) = 1$$

We'll sometimes abuse notation and conflate a with $\{a\}$: that is, we'll interchange an atomic event and a compound event with only one element.

Since events are sets, we can combine them using set operations: given events A and B ,

- $A \cup B$ means that we get some atomic event in the union of A and B; that is, $A \cup B$ is the event that *either* A or B happens
- $A \cap B$ is the event that *both* A and B happen
- $A \setminus B$ is the event that A happens but B does not happen
- $A^C = U \setminus A$ is the event that A does not happen

As you can see from the examples above, set operations correspond to logical combinations of events. So we'll sometimes use logical notation as well:

- $A \vee B$ is the same as $A \cup B$

- $A \wedge B$ is the same as $A \cap B$
- $\neg A$ or $\bar{A}$ is the same as A^C

We can make the correspondence with logical notation precise:

- Logical predicates correspond to events. For example, in our die-roll example, the predicate even corresponds to the set of even-valued outcomes.
- Logical functions can be used to build predicates. For example, if we have a function suit that extracts the suit of a card, then the expression $\text{suit} = \spadesuit$ corresponds to the set $\{a \mid \text{suit}(a) = \spadesuit\}$, or

  $$\{A\spadesuit, 2\spadesuit, 3\spadesuit, 4\spadesuit, 5\spadesuit, 6\spadesuit, 7\spadesuit, 8\spadesuit, 9\spadesuit, 10\spadesuit, J\spadesuit, Q\spadesuit, K\spadesuit\}$$

  Note that we have omitted the argument a to the function suit: by convention, the atomic event is implicitly an argument of every function or predicate.
- General expressions correspond to the set of atomic events that satisfy them. For example, the expression $\text{suit} = \spadesuit \wedge \text{even}$ corresponds to the set

  $$\{2\spadesuit, 4\spadesuit, 6\spadesuit, 8\spadesuit, 10\spadesuit, Q\spadesuit\}$$

  which is $\{a \mid \text{suit}(a) = \spadesuit \wedge \text{even}(a)\}$.

Exercise: what is the set notation for the logical event $\neg(A \to B)$?

Probability density

If we want to write a probability distribution over a continuous set such as $\mathbb{R}$, we will have infinitely many atomic events: one for each real number. We can't assign positive probability to this many events, since the total probability would be ∞ instead of 1. So instead we define a new function $p \in \mathbb{R} \to \mathbb{R}$, and say that the probability of any event $A \subseteq \mathbb{R}$ is the integral

$$P(A) = \int_A p(x)dx$$

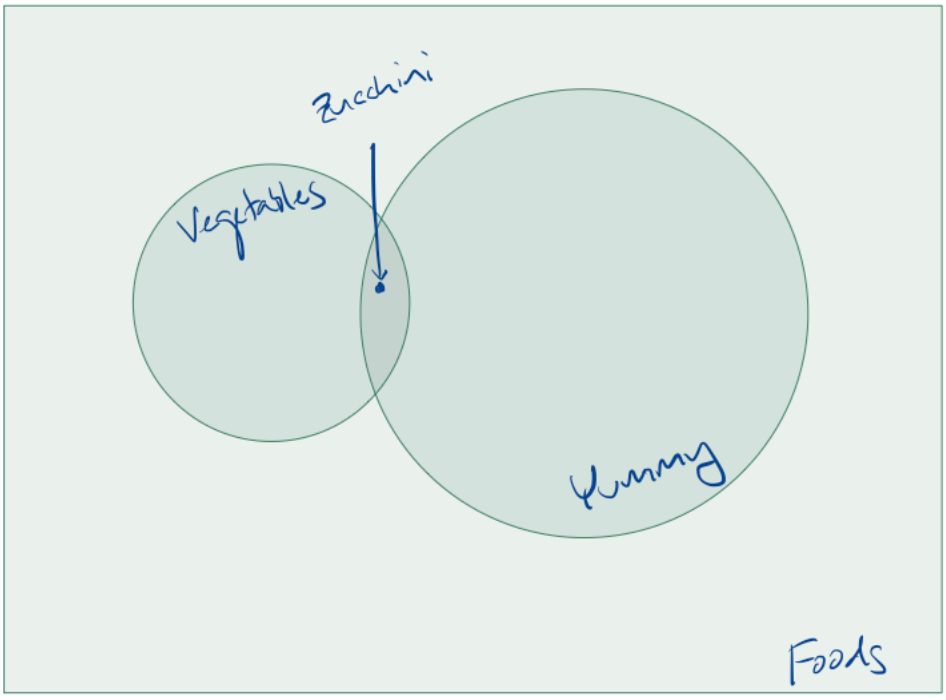
The function p is called a *probability density*, and it obeys many (but not all) of the same rules that a probability does. For example, it must be nonnegative, and the integral over the entire universe must be 1. It also follows the rule that we'll define below for computing conditional distributions (i.e., Bayes' rule).

But it's important to remember that a probability density is *not* a probability. For example, a density can be greater than 1, so long as there's no set A such that $\int_A p(x)dx > 1$. And, the probability of an atomic event x is *not* equal to $p(x)$; as noted above, it's typically zero, so that the probability of x is usually not what we want to ask for. Instead, we might ask for the probability that our random variable will fall in a small interval of length ϵ centered at x; this is approximately $\epsilon\, p(x)$.

We'll skip over most of the math behind continuous probability distributions. If you're interested, this math is interesting and sometimes surprising; a good text is Billingsley.

Venn diagrams

We can visualize our universe and the probabilities of various events using a *Venn diagram*: we picture the universe as a large rectangle, and events as shapes that are subsets of the rectangle. Every point in the rectangle corresponds to a different atomic event. The overlap between shapes tells us which compound events intersect, that is, which can occur simultaneously.



For example, if we are picking a random food to eat for dinner, the universe is the set of all possible foods we might pick, corresponding to the whole rectangle. The event that we pick something yummy is the larger circle, and the event that we pick a vegetable is the smaller circle. The overlap between the two circles is the set of yummy vegetables, which includes the atomic event zucchini — at least, according to your professor. The

area outside the two circles corresponds to the set of outcomes which are neither yummy nor vegetables.

Often we try to make the areas of shapes within the Venn diagram proportional to their probabilities. But sometimes we can't or don't want to — for example, if we don't know those probabilities yet.

> Exercise: where in the above diagram does the event that we pick pizza fit? What about Brussels sprouts? What about escargot? (We realize that the answers are subjective.)

Uniform distributions

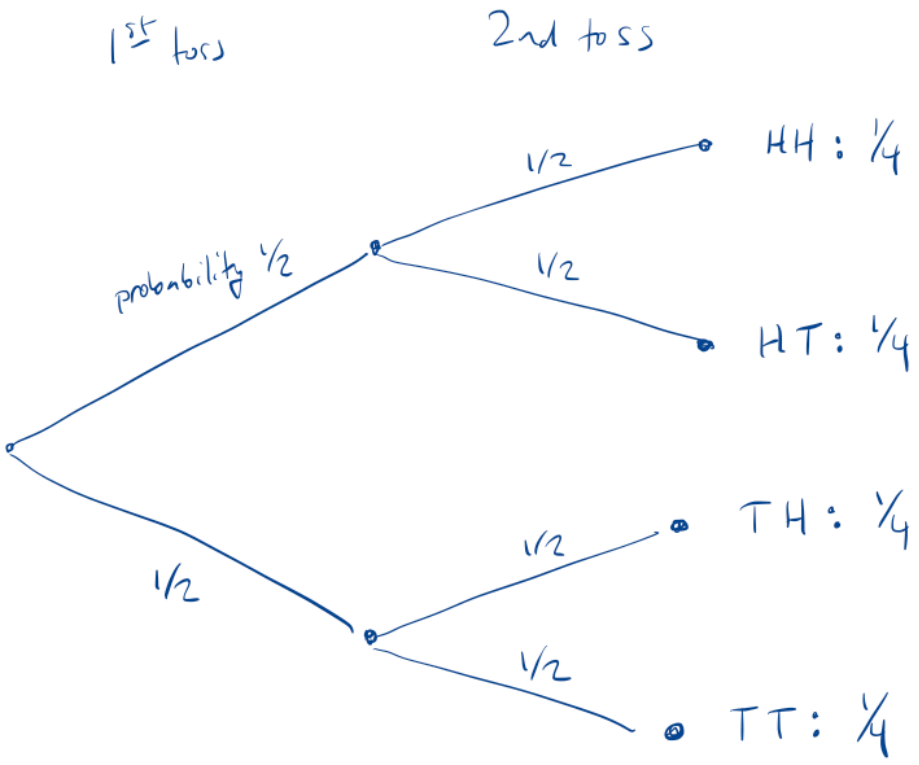
If all atomic events are equally likely, our probability space represents a *uniform distribution*. This distribution is often a default choice for an initial model. But we also often need *nonuniform distributions*, those in which different atomic events have different probabilities.

For example, a fair die is modeled well by a uniform distribution: $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$. But a weighted die needs a non-uniform distribution. If rolling a 6 is four times as likely as any other outcome, we have $P(1) = P(2) = P(3) = P(4) = P(5) = \frac{1}{9}$ and $P(6) = \frac{4}{9}$.

Experiments

We can use probabilities to describe *experiments* that we are thinking of conducting. We make a probability space that describes what we think might happen; atomic events represent possible outcomes of the experiment, and the universe contains all possible outcomes.

For example, suppose we flip a coin twice. We can make a probability space that says that each sequence of heads and tails is equally likely:



Here the universe is $\{HH, HT, TH, TT\}$, and we've assigned probability $\frac{1}{4}$ to each of these atomic events.

In a realistic experiment, we might not observe the distinctions between all atomic events. For example, above we might just observe the total number of heads instead of the exact sequence of coin flips. We can describe what we observe using compound events: in this case, $\{HH\}$, $\{HT, TH\}$, and $\{TT\}$.

Samples

A typical experiment is *repeatable*: we can imagine doing the same experiment more than once, with the distribution over outcomes $P(X)$ being the same each time. If we repeat T times, we can collect the outcomes into a *sample* or *data set* $X_1, X_2, \ldots, X_T$.

By assumption, order doesn't matter: the sample is *exchangeable*. That is, we can permute the indices 1:T without changing the information contained in the sample.

> *Despite the name, a data set is not a set: like a set, order is unimportant, but unlike a set, it matters if we see the same outcome more than once.*

For example, if we flip a single coin seven times, we might see the sample H, H, T, H, T, T, T. (That is, $X_1 = H$, $X_2 = H$, $X_3 = T$, and so forth.) This outcome is

equivalent to T, H, T, H, T, H, T or T, T, T, T, H, H, H: all that matters is that we saw three heads and four tails.

For another example, we might flip a pair of coins three times, recording the number of heads each time. Then we might see the samples $0, 0, 2$ or $1, 1, 1$. What matters here is the number of times we saw 0, the number of times we saw 1, and the number of times we saw 2.

Working with probabilities

In small probability spaces we can just list out all the atomic events, and calculate probabilities of compound events by the sum rule. But it's common to have so many atomic events that we have no hope of listing them all: for example, there are more possible shuffles of a 52-card deck than there are atoms in the Earth. So, we need tools to help us do the calculations without explicitly enumerating everything. There are *lots* of such tools; but in the rest of these notes we'll cover a few of the most useful.

Disjoint union

If A and B are disjoint events, meaning that $A \cap B = \emptyset$, then

$$P(A \cup B) = P(A) + P(B)$$

This follows directly from our definitions:

$$P(A \cup B) = \sum_{e \in A \cup B} P(e) = \sum_{e \in A} P(e) + \sum_{e \in B} P(e) = P(A) + P(B)$$

Disjoint events are also called *mutually exclusive*.

The disjoint union rule also works for multiple sets: if the sets A_i are mutually exclusive, meaning that they are pairwise disjoint, then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

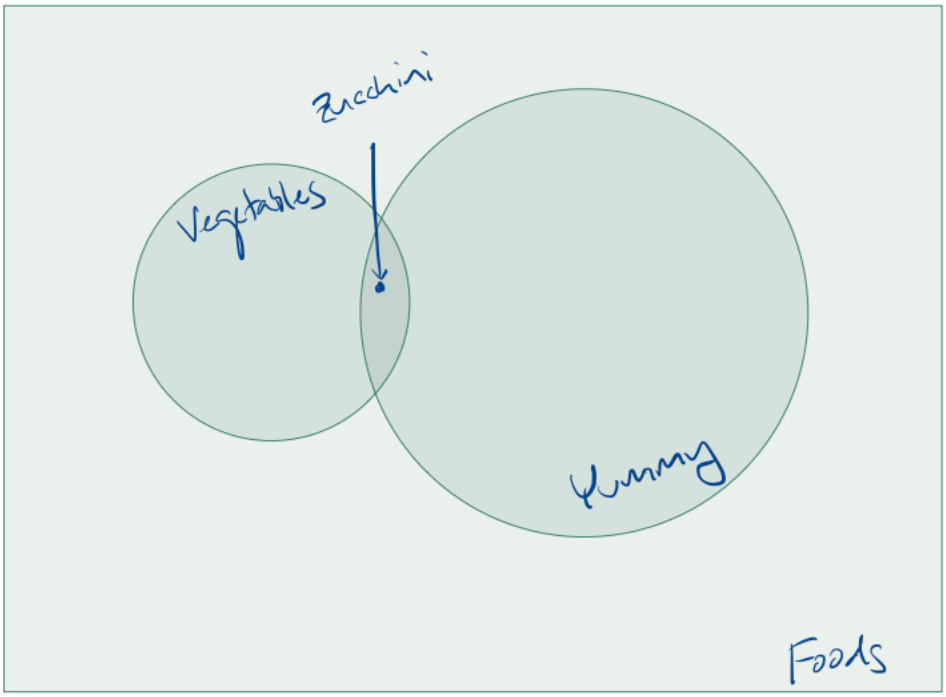
A useful generalization of the disjoint union rule is the *union bound*: for any A and B, disjoint or not, we have

$$P(A \cup B) \leq P(A) + P(B)$$

We'll prove this in the next section.

Non-disjoint union

We can use the tools above to figure out what happens when we take a union of sets that overlap, like in our food example from before.



Let A be the event that we pick a vegetable, and B be the event that we pick a yummy food. How can we calculate $P(A \cup B)$?

We can split the set $A \cup B$ into three disjoint parts: $A \setminus B$, $B \setminus A$, and $A \cap B$. So,

$$P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B)$$

We can also split A and B into two pieces each:

$$P(A) = P(A \setminus B) + P(A \cap B)$$

$$P(B) = P(B \setminus A) + P(A \cap B)$$

Adding these together,

$$P(A) + P(B) = P(A \setminus B) + P(B \setminus A) + 2P(A \cap B)$$

Since $P(A \cap B) \geq 0$, we see that $P(A) + P(B) \geq P(A \cup B)$; this is the union bound we described above. By subtracting the above expressions for $P(A) + P(B)$ and $P(A \cup B)$ we can figure out the exact difference:

$$P(A) + P(B) - P(A \cup B) = P(A \cap B)$$

Exercise: Suppose we stand on a street corner in Pittsburgh and survey people. We ask two questions: do you like computer science, and do you watch the Steelers. You find that 67% of passers-by respond yes to the first question, and 90% respond yes to the second. Furthermore, 65% respond yes to both questions. What is the probability of finding a person who responds no to both questions?

Logical rules

We've already talked about the relationship between logical and set-based descriptions of events. This relationship means that there is a probability inference rule for every logical inference rule. For example, De Morgan's laws:

$$P(\neg(A \wedge B)) = P(\neg A \vee \neg B)$$

$$P(\neg(A \vee B)) = P(\neg A \vee \neg B)$$

Perhaps more subtly, if one logical statement implies another, $\psi \to \phi$, then $P(\phi) \geq P(\psi)$. For example, if being human implies being mortal, then $P(\text{mortal}) \geq P(\text{human})$. This follows since the event ψ (here, being human) has to be a subset of the event ϕ (here, being mortal).

Combinatorics

If we start from a uniform distribution over our entire probability space, then calculating the probabilities of events boils down to counting: how many atomic events satisfy a given property? For this purpose, combinatoric tools can be very useful. The general strategy is to describe a compound event using a sequence of simpler choices, then count the ways to make these choices, so that we can get an exhaustive description of all the atomic events that make up a compound event. The key difficulty is to make sure to count each atomic event exactly once.

For example, what is the probability of drawing three of a kind in a five-card hand? The atomic events here are the possible unordered five-card hands: $\frac{52!}{47!5!}$ of them. To get three of a kind, we must pick what rank to have three of (13 choices), which card of that rank to exclude (4 choices), and what two other cards to include in the hand ($\frac{48 \cdot 47}{2}$ choices). Every way of making these choices describes a different three-of-a-kind hand, and all such hands can be constructed with some sequence of these choices. So the total number of atomic events that correspond to three-of-a-kind hands is

$13 \cdot 4 \cdot 48 \cdot 47/2$, and the probability of getting one is

$$\frac{13 \cdot 4 \cdot 48 \cdot 47 \cdot 5!}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 2}$$

or about 2.26%.

> *The above calculation assumes that we care about getting exactly 3 of one rank, and that the other two cards can be anything. We could include 4-of-a-kind hands (since they contain three of a kind); or we could exclude hands with a full house (where the remaining two cards are a pair). Either of these would change the result. For example, to exclude a full house, we replace* $48 \cdot 47$ *with* $48 \cdot 44$*: we choose the fourth card to have a different rank from the previous three, then we choose the fifth card to have a different rank from the previous four. This changes the probability to about 2.11%.*

The basic building block in the above calculations is the number of ways to choose k distinct things out of a menu of n. This is written

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

and pronounced "n choose k". For example, choosing two distinct cards out of 48 yields $\left(\begin{array}{c} 48 \\ 2 \end{array} \right)$ options, and choosing 5 out of 52 yields $\left(\begin{array}{c} 52 \\ 5 \end{array} \right)$ options; we used both of these values above. See Wikipedia's article on poker hand probabilities for lots more examples.

Extensions

Probability spaces are extremely expressive, but they don't cover every possible situation. Here are a few examples of situations that might need additional tools.

First, what if I think that I'm not the only agent in the universe? That is, I believe there are other intelligent beings who predict my actions and try to optimize their own actions in response. It seems hard to assign probabilities to the actions of these agents.

For another example, suppose two people disagree on the probabilities of some outcomes. They can each build a separate probability space; but how should they talk with each other about the probabilities of events that they both can observe? How should they interpret each other's behavior?

For a third example, exact probability calculations can be very expensive. What should I do if I can't afford the necessary computation? How should I predict the behavior of a

computationally limited agent?

For a final example, the conclusions I reach by probabilistic reasoning can be very sensitive to my starting model. Can I recognize when one of my assumptions causes a large change in my conclusions? Can I design a more stable reasoning process?