Chapter 12

Probability theory

Here are two examples of questions we might ask about the likelihood of some event:

- Gambling: I throw two six-sided dice, what are my chances of seeing a 7?
- Insurance: I insure a typical resident of Smurfington-upon-Tyne against premature baldness. How likely is it that I have to pay a claim?

Answers to these questions are summarized by a **probability**, a number in the range 0 to 1 that represents the likelihood that some event occurs. There are two dominant interpretations of this likelihood:

- The **frequentist interpretation** says that if an event occurs with probability p, then in the limit as I accumulate many examples of similar events, I will see the number of occurrences divided by the number of samples converging to p. For example, if I flip a fair coin over and over again many times, I expect that heads will come up roughly half of the times I flip it, because the probability of coming up heads is $1/2$.
- The **Bayesian interpretation** says that when I say that an event occurs with probability p, that means my subjective beliefs about the event would lead me to take a bet that would be profitable on average if this were the real probability. So a Bayesian would take a double-or-nothing bet on a coin coming up heads if they believed that the probability it came up heads was at least $1/2$.

Frequentists and Bayesians have historically spent a lot of time arguing with each other over which interpretation makes sense. The usual argument against frequentist probability is that it only works for repeatable experiments, and doesn't allow for statements like "the probability that it will rain tomorrow is 50%" or the even more problematic "based on what I know, there is a 50% probability that it rained yesterday." The usual argument against Bayesian probability is that it's hopelessly subjective—it's possible (even likely) that my subjective guesses about the probability that it will rain tomorrow are not the same as yours.[1]

As mathematicians, we can ignore such arguments, and treat probability axiomatically as just another form of counting, where we normalize everything so that we always end up counting to exactly 1. It happens to be the case that this approach to probability works for both frequentist interpretations (assuming that the probability of an event measures the proportion of outcomes that cause the event to occur) and Bayesian interpretations (assuming our subjective beliefs are consistent).

12.1 Events and probabilities

We'll start by describing the basic ideas of probability in terms of probabilities of events, which either occur or don't. Later we will generalize these ideas and talk about random variables, which may take on many different values in different outcomes.

12.1.1 Probability axioms

Coming up with axioms for probabilities that work in all the cases we want to consider took much longer than anybody expected, and the current set in common use only go back to the 1930's. Before presenting these, let's talk a bit about the basic ideas of probability.

An **event** A is something that might happen, or might not; it acts like a predicate over possible outcomes. The **probability** $\Pr[A]$ of an event A is a real number in the range 0 to 1, that must satisfy certain consistency rules like $\Pr[\neg A] = 1 - \Pr[A]$.

In **discrete probability**, there is a finite set of **atoms**, each with an assigned probability, and every event is a union of atoms. The probability assigned to an event is the sum of the probabilities assigned to the atoms it contains. For example, we could consider rolling two six-sided dice. The

[1] This caricature of the debate over interpreting probability is thoroughly incomplete. For a thoroughly complete discussion, including many other interpretations, see http://plato.stanford.edu/entries/probability-interpret/.

atoms are the pairs (i, j) that give the value on the first and second die, and we assign a probability of 1/36 to each pair. The probability that we roll a 7 is the sum of the cases $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$, and $(6, 1)$, or $6/36 = 1/6$.

Discrete probability doesn't work if we have infinitely many atoms. Suppose we roll a pair of dice infinitely many times (e.g., because we want to know the probability that we never accumulate more 6's than 7's in this infinite sequence). Now there are infinitely many possible outcomes: all the sequences of pairs (i, j). If we make all these outcomes equally likely, we have to assign each a probability of zero. But then how do we get back to a probability of 1/6 that the first roll comes up 7?

12.1.1.1 The Kolmogorov axioms

A triple $(\Omega, \mathcal{F}, P)$ is a **probability space** if Ω is a set of **outcomes** (where each outcome specifies everything that ever happens, in complete detail); $\mathcal{F}$ is a **sigma-algebra**, which is a family of subsets of Ω, called **measurable sets**, that is closed under complement (i.e., if A is in $\mathcal{F}$ then $\Omega \setminus A$ is in $\mathcal{F}$) and countable union (union of $A_1, A_2, \dots$ is in $\mathcal{F}$ if each set A_i is); and P is a **probability measure** that assigns a number in $[0, 1]$ to each set in $\mathcal{F}$. The measure P must satisfy three axioms, due to Kolmogorov [Kol33]:

1. $P(A) \geq 0$ for all $A \in \mathcal{F}$.
2. $P(\Omega) = 1$.
3. For any sequence of pairwise disjoint events $A_1, A_2, A_3, \dots$, $P(\cup A_i) = \sum P(A_i)$.

From these one can derive rules like $P(\Omega \setminus A) = 1 - P(A)$ etc.

Most of the time, Ω is finite, and we can just make $\mathcal{F}$ include all subsets of Ω, and define $P(A)$ to be the sum of $P(\{x\})$ over all x in A. This gets us back to the discrete probability model we had before.

Unless we are looking at multiple probability spaces or have some particular need to examine Ω, $\mathcal{F}$, or P closely, we usually won't bother specifying the details of the probability space we are working in. So most of the time we will just refer to "the" probability $\Pr[A]$ of an event A, bearing in mind that we are implicitly treating A as a subset of some implicit Ω that is measurable with respect to an implicit $\mathcal{F}$ and whose probability is really $P(A)$ for some implicit measure P.

12.1.1.2 Examples of probability spaces

- $\Omega = \{\mathsf{H}, \mathsf{T}\}$, $\mathcal{F} = \mathcal{P}(\Omega) = \{\{,\} \{\mathsf{H}\}, \{\mathsf{T}\}, \{\mathsf{H}, \mathsf{T}\}\}$, $\Pr[A] = |A|/2$. This represents a fair coin with two outcomes H and T that each occur with probability 1/2.

- $\Omega = \{\mathsf{H}, \mathsf{T}\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\Pr[\{\mathsf{H}\}] = p$, $\Pr[\{\mathsf{T}\}] = 1 - p$. This represents a biased coin, where H comes up with probability p.

- $\Omega = \{(i, j) \mid i, j \in \{1, 2, 3, 4, 5, 6\}\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\Pr[A] = |A|/36$. Roll of two fair dice. A typical event might be "the total roll is 4", which is the set $\{(1, 3), (2, 2), (3, 1)\}$ with probability $3/36 = 1/12$.

- $\Omega = \mathbb{N}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\Pr[A] = \sum_{n \in A} 2^{-n-1}$. This is an infinite probability space; a real-world process that might generate it is to flip a fair coin repeatedly and count how many times it comes up tails before the first time it comes up heads. Note that even though it is infinite, we can still define all probabilities by summing over atoms: $\Pr[\{0\}] = 1/2$, $\Pr[\{1\}] = 1/4$, $\Pr[\{0, 2, 4, \dots\}] = 1/2 + 1/8 + 1/32 + \cdots = 2/3$, etc.

It's unusual for anybody doing probability to actually write out the details of the probability space like this. Much more often, a writer will just assert the probabilities of a few basic events (e.g. $\Pr[\{\mathsf{H}\}] = 1/2$), and claim that any other probability that can be deduced from these initial probabilities from the axioms also holds (e.g. $\Pr[\{\mathsf{T}\}] = 1 - \Pr[\{\mathsf{H}\}] = 1/2$). The main reason Kolmogorov gets his name attached to the axioms is that he was responsible for **Kolmogorov's extension theorem**, which says (speaking very informally) that as long as your initial assertions are consistent, there exists a probability space that makes them and all their consequences true.

12.1.2 Probability as counting

The easiest probability space to work with is a **uniform discrete probability space**, which has N outcomes each of which occurs with probability $1/N$. If someone announces that some quantity is "random" without specifying probabilities (especially if that someone is a computer scientist), the odds are that what they mean is that each possible value of the quantity is equally likely. If that someone is being more careful, they would say that the quantity is "drawn uniformly at random" from a particular set.

Such spaces are among the oldest studied in probability, and go back to the very early days of probability theory where randomness was almost always expressed in terms of pulling tokens out of well-mixed urns, because such "urn models" where one of the few situations where everybody agreed on what the probabilities should be.

12.1.2.1 Examples

- A **random bit** has two outcomes, 0 and 1. Each occurs with probability 1/2.
- A **die roll** has six outcomes, 1 through 6. Each occurs with probability 1/6.
- A roll of two dice has 36 outcomes (order of the dice matters). Each occurs with probability 1/36.
- A random n-bit string has 2^n outcomes. Each occurs with probability 2^{-n}. The probability that exactly one bit is a 1 is obtained by counting all strings with a single 1 and dividing by 2^n. This gives $n2^{-n}$.
- A **poker hand** consists of a subset of 5 cards drawn uniformly at random from a deck of 52 cards. Depending on whether the order of the 5 cards is considered important (usually it isn't), there are either $\binom{52}{5}$ or $(52)_5$ possible hands. The probability of getting a flush (all five cards in the hand drawn from the same suit of 13 cards) is $4\binom{13}{5}/\binom{52}{5}$; there are 4 choices of suits, and $\binom{13}{5}$ ways to draw 5 cards from each suit.
- A **random permutation** on n items has $n!$ outcomes, one for each possible permutation. A typical event might be that the first element of a random permutation of $1 \dots n$ is 1; this occurs with probability $(n-1)!/n! = 1/n$. Another example of a random permutation might be a uniform shuffling of a 52-card deck (difficult to achieve in practice!). Here, the probability that we get a particular set of 5 cards as the first 5 in the deck is obtained by counting all the permutations that have those 5 cards in the first 5 positions (there are $5! \cdot 47!$ of them) divided by 52!. The result is the same $1/\binom{52}{5}$ that we get from the uniform poker hands.

12.1.3 Independence and the intersection of two events

Events A and B are **independent** if $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$. In general, a set of events $\{A_i\}$ is independent if each A_i is independent of any event defined only in terms of the other events.

It can be dangerous to assume that events are independent when they aren't, but quite often when describing a probability space we will explicitly state that certain events are independent. For example, one typically describes the space of random n-bit strings (or n coin flips) by saying that one has n independent random bits and then deriving that each particular sequence occurs with probability 2^{-n} rather than starting with each sequence occurring with probability 2^{-n} and then calculating that each particular bit is 1 with independent probability 1/2. The first description makes much more of the structure of the probability space explicit, and so is more directly useful in calculation.

12.1.3.1 Examples

- What is the probability of getting two heads on independent fair coin flips? Calculate it directly from the definition of independence: $\Pr[H_1 \cap H_2] = (1/2)(1/2) = 1/4$.
- Suppose the coin-flips are *not* independent (maybe the two coins are glued together). What is the probability of getting two heads? This can range anywhere from zero (coin 2 always comes up the opposite of coin 1) to 1/2 (if coin 1 comes up heads, so does coin 2).
- What is the probability that both you and I draw a flush (all 5 cards the same suit) from the same poker deck? Since we are fighting over the same collection of same-suit subsets, we'd expect $\Pr[A \cap B] \neq \Pr[A] \cdot \Pr[B]$—the event that you get a flush (A) is not independent of the event that I get a flush (B), and we'd have to calculate the probability of both by counting all ways to draw two hands that are both flushes. But if we put your cards back and then shuffle the deck again, the events in this new case *are* independent, and we can just square the $\Pr[\text{flush}]$ that we calculated before.
- Suppose the Red Sox play the Yankees. What is the probability that the final score is exactly 4–4? Amazingly, it appears that it is equal to[2]

  $$\begin{aligned} &\Pr[\text{Red Sox score 4 runs against the Yankees}] \\ &\cdot \Pr[\text{Yankees score 4 runs against the Red Sox}]. \end{aligned}$$

  To the extent we can measure the underlying probability distribution, the score of each team in a professional baseball game appears to be independent of the score of the other team.

[2] See http://arXiv.org/abs/math/0509698.

12.1.4 Union of events

What is the probability of $A \cup B$? If A and B are disjoint, then the axioms give $\Pr[A \cup B] = \Pr[A] + \Pr[B]$. But what if A and B are not disjoint?

By analogy to inclusion-exclusion in counting we would expect that

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

Intuitively, when we sum the probabilities of A and B, we double-count the event that both occur, and must subtract it off to compensate. To prove this formally, consider the events $A \cap B$, $A \cap \neg B$, and $\neg A \cap B$. These are disjoint, so the probability of the union of any subset of this set of events is equal to the sum of its components. So in particular we have

$$\begin{aligned}
&\Pr[A] + \Pr[B] - \Pr[A \cap B] \\
&\quad = (\Pr[A \cap B] + \Pr[A \cap \neg B]) + (\Pr[A \cap B] + \Pr[\neg A \cap B]) - \Pr[A \cap B] \\
&\quad = \Pr[A \cap B] + \Pr[A \cap \neg B] + \Pr[\neg A \cap B] \\
&\quad = \Pr[A \cup B].
\end{aligned}$$

12.1.4.1 Examples

- What is the probability of getting at least one head out of two independent coin-flips? Compute $\Pr[H_1 \cup H_2] = 1/2 + 1/2 - (1/2)(1/2) = 3/4$.
- What is the probability of getting at least one head out of two coin-flips, when the coin-flips are not independent? Here again we can get any probability from 0 to 1, because the probability of getting at least one head is just $1 - \Pr[T_1 \cap T_2]$.

For more events, we can use a probabilistic version of the inclusion-exclusion formula (Theorem 11.2.2). The new version looks like this:

Theorem 12.1.1. *Let $A_1 \dots A_n$ be events on some probability space. Then*

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{S \subseteq \{1 \dots n\}, S \neq \emptyset} (-1)^{|S|+1} \Pr\left[\bigcap_{j \in S} A_j\right]. \tag{12.1.1}$$

For discrete probability, the proof is essentially the same as for Theorem 11.2.2; the difference is that instead of showing that we add 1 for each possible element of $\bigcap A_i$, we show that we add the probability of each outcome in $\bigcap A_i$. The result continues to hold for more general spaces, but requires a little more work.[3]

[3] The basic idea is to chop $\bigcap A_i$ into all sets of the form $\bigcup B_i$ where each B_i is either A_i or $\neg A_i$; this reduces to the discrete case.

12.1.5 Conditional probability

Suppose I want to answer the question "What is the probability that my dice add up to 6 if I know that the first one is an odd number?" This question involves **conditional probability**, where we calculate a probability subject to some conditions. The probability of an event A conditioned on an event B, written $\Pr[A \mid B]$, is defined by the formula

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

One way to think about this is that when we assert that B occurs we are in effect replacing the entire probability space with just the part that sits in B. So we have to divide all of our probabilities by $\Pr[B]$ in order to make $\Pr[B \mid B] = 1$, and we have to replace A with $A \cap B$ to exclude the part of A that can't happen any more.

Note also that conditioning on B only makes sense if $\Pr[B] > 0$. If $\Pr[B] = 0$, $\Pr[A \mid B]$ is undefined.

12.1.5.1 Conditional probabilities and intersections of non-independent events

Simple algebraic manipulation gives

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

So one of the ways to compute the probability of two events occurring is to compute the probability of one of them, and the multiply by the probability that the second occurs conditioned on the first. For example, if my attempt to reach the summit of Mount Everest requires that I first learn how to climb mountains ($\Pr[B] = 0.1$) and then make it to the top safely ($\Pr[A \mid B] = 0.9$), then my chances of getting to the top are $\Pr[A \cap B] = \Pr[A \mid B] \cdot \Pr[B] = (0.9)(0.1) = 0.09$.

We can do this for sequences of events as well. Suppose that I have an urn that starts with k black balls and 1 red ball. In each of n trials I draw one ball uniformly at random from the urn. If it is red, I give up. If it is black, I put the ball back and add another black ball, thus increasing the number of balls by 1. What is the probability that on every trial I get a black ball?

Let A_i be the event that I get a black ball in each of the first i trials. Then $\Pr[A_0] = 1$, and for larger i we have $\Pr[A_i] = \Pr[A_i \mid A_{i-1}] \Pr[A_{i-1}]$. If A_{i-1} holds, then at the time of the i-th trial we have $k + i$ total balls in the urn, of which one is red. So the probability that we draw a black ball is $1 - \frac{1}{k+i} = \frac{k+i-1}{k+i}$. By induction we can then show that

$$\Pr\left[A_i\right] = \prod_{j=1}^{i} \frac{k+j-1}{k+j}.$$

This is an example of a collapsing product, where the denominator of each fraction cancels out the numerator of the next; we are left only with the denominator $k+i$ of the last term and the numerator k of the first, giving $\Pr\left[A_i\right] = \frac{k}{k+i}$. It follows that we make it through all n trials with probability $\Pr\left[A_n\right] = \frac{k}{k+n}$.

12.1.5.2 The law of total probability

We can use the fact that A is the disjoint union of $A \cap B$ and $A \cap \overline{B}$ to get $\Pr\left[A\right]$ by case analysis:

$$\begin{aligned}\Pr\left[A\right] &= \Pr\left[A \cap B\right] + \Pr\left[A \cap \overline{B}\right] \\ &= \Pr\left[A \mid B\right]\Pr\left[B\right] + \Pr\left[A \mid \overline{B}\right]\Pr\left[\overline{B}\right].\end{aligned}$$

For example, if there is a 0.2 chance I can make it to the top of Mt Everest safely without learning how to climb first, my chances of getting there go up to $(0.9)(0.1) + (0.2)(0.9) = 0.27$.

This method is sometimes given the rather grandiose name of the **law of total probability**. The most general version is that if $B_1 \dots B_n$ are all disjoint events and the sum of their probabilities is 1, then

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

12.1.5.3 Bayes's formula

If one knows $\Pr\left[A \mid B\right]$, $\Pr\left[A \mid \neg B\right]$, and $\Pr\left[B\right]$, it's possible to compute $\Pr\left[B \mid A\right]$:

$$\begin{aligned}\Pr\left[B \mid A\right] &= \frac{\Pr\left[A \cap B\right]}{\Pr\left[A\right]} \\ &= \frac{\Pr\left[A \mid B\right]\Pr\left[B\right]}{\Pr\left[A\right]} \\ &= \frac{\Pr\left[A \mid B\right]\Pr\left[B\right]}{\Pr\left[A \mid B\right]\Pr\left[B\right] + \Pr\left[A \mid \overline{B}\right]\Pr\left[\overline{B}\right]}.\end{aligned}$$

This formula is used heavily in statistics, where it goes by the name of **Bayes's formula**. Say that you have an Airport Terrorist Detector that lights up with probability 0.75 when inserted into the nostrils of a Terrorist, but lights up with probability 0.001 when inserted into the nostrils of a non-Terrorist. Suppose that for other reasons you know that Granny has only a 0.0001 chance of being a Terrorist. What is the probability that Granny is a Terrorist if the detector lights up?

Let B be the event "Granny is a terrorist" and A the event "Detector lights up." Then $\Pr[B \mid A] = (0.75{\times}0.0001)/(0.75{\times}0.0001{+}0.001{\times}0.9999) \approx 0.0007495$. This example shows how even a small **false positive** rate can make it difficult to interpret the results of tests for rare conditions.

12.2 Random variables

A **random variable** X is a variable that takes on particular values randomly. This means that for each possible value x, there is an event $[X = x]$ with some probability of occurring that corresponds to X (the random variable, usually written as an upper-case letter) taking on the value x (some fixed value). Formally, a random variable X is really a function $X(\omega)$ of the outcome ω that occurs, but we save a lot of ink by leaving out ω.[4]

12.2.1 Examples of random variables

- Indicator variables: The **indicator variable** for an event A is a variable X that is 1 if A occurs and 0 if it doesn't (i.e., $X(\omega) = 1$ if $\omega \in A$ and 0 otherwise). There are many conventions out there for writing indicator variables. I am partial to 1_A, but you may also see them written using the Greek letter chi (e.g. χ_A) or by abusing the bracket notation for events (e.g., $[A]$, $[Y^2 > 3]$, [all six coins come up heads]).

- Functions of random variables: Any function you are likely to run across of a random variable or random variables is a random variable. If X and Y are random variables, $X+Y$, XY, and $\log X$ are all random variables.

- Counts of events: Flip a fair coin n times and let X be the number of times it comes up heads. Then X is an integer-valued random variable.

[4] For some spaces, not all functions $X(\omega)$ work as random variables, because the events $[X = x]$ might not be measurable with respect to $\mathcal{F}$. We will generally not run into these issues.

- Random sets and structures: Suppose that we have a set T of n elements, and we pick out a subset U by flipping an independent fair coin for each element to decide whether to include it. Then U is a set-valued random variable. Or we could consider the infinite sequence $X_0, X_1, X_2, \dots$, where $X_0 = 0$ and X_{n+1} is either $X_n + 1$ or $X_n - 1$, depending on the result of independent fair coin flip. Then we can think of the entire sequence X as a sequence-valued random variable.

12.2.2 The distribution of a random variable

The **distribution** of a random variable describes the probability that it takes on various values. For real-valued random variables, the **distribution function** or **cumulative distribution function** is a function $F(x) = \Pr[X \leq x]$. This allows for very general distributions—for example, a variable that is uniform on $[0, 1]$ can be specified by $F(x) = x$ when $0 \leq x \leq 1$, and 0 or 1 as appropriate outside this interval—but for **discrete random variables** that take on only countably many possible values, this is usually more power than we need.

For discrete variables, the distribution is most easily described by just giving the **probability mass function** $\Pr[X = x]$ for each possible value x. If we need to, it's not too hard to recover the distribution function from the mass function (or vice versa). So we will often cheat a bit and treat a mass function as specifying a distribution even if it isn't technically a distribution function.

Typically, if we know the distribution of a random variable, we don't bother worrying about what the underlying probability space is. The reason for this is we can just take Ω to be the range of the random variable, and define $\Pr[\omega]$ for each ω in Ω to be $\Pr[X = \omega]$. For example, a six-sided die corresponds to taking $\Omega = \{1, 2, 3, 4, 5, 6\}$, assigning $\Pr[\omega] = 1/6$ for all ω, and letting $X(\omega) = \omega$. This will give the probabilities for any events involving X that we would have gotten on whatever original probability space X might have been defined on.

The same thing works if we have multiple random variables, but now we let each point in the space be a tuple that gives the values of all of the variables. Specifying the probability in this case is done using a **joint distribution** (see below).

12.2.2.1 Some standard distributions

Here are some common distributions for a random variable X:

- **Bernoulli distribution**: $\Pr[X=1]=p$, $\Pr[X=0]=q$, where p is a parameter of the distribution and $q=1-p$. This corresponds to a single biased coin-flip.
- **Binomial distribution**: $\Pr[X=k]=\binom{n}{k}p^k q^{(n-k)}$, where n and p are parameters of the distribution and $q=1-p$. This corresponds to the sum of n biased coin-flips.
- **Geometric distribution**: $\Pr[X=k]=q^k p$, where p is a parameter of the distribution and q is again equal to $1-p$. This corresponds to number of tails we flip before we get the first head in a sequence of biased coin-flips.
- **Poisson distribution**: $\Pr[X=k]=e^{-\lambda}\lambda^k/k!$. This is what happens to a binomial distribution when we make $p=\lambda/n$ and then take the limit as n goes to infinity. We can think of it as counting the number of events that occur in one time unit if the events occur at a constant continuous rate that averages λ events per time unit. The canonical example is radioactive decay.
- **Uniform distribution**: For the uniform distribution on $[a,b]$, the distribution function F of X is given by $F(x)=0$ when $x \le a, (x-a)/(b-a)$ when $a \le x \le b$, and 1 when $b \le x$, where a and b are parameters of the distribution. This is a continuous random variable that has equal probability of landing anywhere in the $[a,b]$ interval.

  The term *uniform distribution* may also refer to a uniform distribution on a finite set S; this assigns $\Pr[X=x]=\frac{1}{|S|}$ when x is in S and 0 otherwise. As a distribution function, $F(x)$ is the rather discontinuous function $|\{y \in S \mid y \le x\}|/|S|$.
- **Normal distribution**: The normal distribution function is given by

  $$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-x^2/2}\,dx.$$

  This corresponds to another limit of the binomial distribution, where now we fix $p=1/2$ but compute $\frac{X-n/2}{\sqrt{n}}$ to converge to a single fixed distribution as n goes to infinity. The normal distribution shows up (possibly scaled and shifted) whenever we have a sum of many independent, identically distributed random variables: this is the **Central Limit Theorem**, and is the reason why much of statistics works, and why we can represent 0 and 1 bits using buckets of jumpy randomly-positioned electrons.

12.2.2.2 Joint distributions

Two or more random variables can be described using a **joint distribution**. For discrete random variables, we often represent this as a joint probability mass function $\Pr[X = x \wedge Y = y]$ for all fixed values x and y, or more generally $\Pr[\forall i : X_i = x_i]$. For continuous random variables, we may instead need to use a joint distribution function $F(x_1, \dots, x_n) = \Pr[\forall i : X_i \leq x_i]$.

Given a joint distribution on X and Y, we can recover the distribution on X or Y individually by summing up cases: $\Pr[X = x] = \sum_y \Pr[X = x \wedge Y = y]$ (for discrete variables), or $\Pr[X \leq x] = \lim_{y \to \infty} \Pr[X \leq x \wedge Y \leq y]$ (for more general variables). The distribution of X obtained in this way is called a **marginal distribution** of the original joint distribution. In general, we can't go in the other direction, because just knowing the marginal distributions doesn't tell us how the random variables might be dependent on each other.

Examples

- Let X and Y be six-sided dice. Then $\Pr[X = x \wedge Y = y] = 1/36$ for all values of x and y in $\{1, 2, 3, 4, 5, 6\}$. The underlying probability space consists of all pairs (x, y) in $\{1, 2, 3, 4, 6\} \times \{1, 2, 3, 4, 5, 6\}$.

- Let X be a six-sided die and let $Y = 7-X$. Then $\Pr[X = x \wedge Y = y] = 1/6$ if $1 \leq x \leq 6$ and $y = 7 - x$, and 0 otherwise. The underlying probability space is most easily described by including just six points for the X values, although we could also do $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ as in the previous case, just assigning probability 0 to most of the points. However, even though the joint distribution is very different from the previous case, the marginal distributions of X and Y are exactly the same as before: each of X and Y takes on all values in $\{1, 2, 3, 4, 5, 6\}$ with equal probability.

12.2.3 Independence of random variables

The difference between the two preceding examples is that in the first case, X and Y are independent, and in the second case, they aren't.

Two random variables X and Y are **independent** if any pair of events of the form $X \in A$, $Y \in B$ are independent. For discrete random variables, it is enough to show that $\Pr[X = x \wedge Y = y] = \Pr[X = x] \cdot \Pr[Y = y]$, or in other words that the events $[X = x]$ and $[Y = y]$ are independent for all values x and y. For continuous random variables, the corresponding equation is $\Pr[X \le x \wedge Y \le y] = \Pr[X \le x] \cdot \Pr[Y \le y]$. In practice, we will typically either be told that two random variables are independent or deduce it from the fact that they arise from separated physical processes.

12.2.3.1 Examples

- Roll two six-sided dice, and let X and Y be the values of the dice. By convention we assume that these values are independent. This means for example that $\Pr[X \in \{1,2,3\} \wedge Y \in \{1,2,3\}] = \Pr[X \in \{1,2,3\}] \cdot \Pr[Y \in \{1,2,3\}] = (1/2)(1/2) = 1/4$, which is a slightly easier computation than counting up the 9 cases (and then arguing that each occurs with probability $(1/6)^2$, which requires knowing that X and Y are independent).

- Take the same X and Y, and let $Z = X + Y$. Now Z and X are not independent, because $\Pr[X = 1 \wedge Z = 12] = 0$, which is not equal to $\Pr[X = 1] \cdot \Pr[Z = 12] = (1/6)(1/36) = 1/216$.

- Place two radioactive sources on opposite sides of the Earth, and let X and Y be the number of radioactive decay events in each source during some 10 millisecond interval. Since the sources are 42 milliseconds away from each other at the speed of light, we can assert that either X and Y are independent, or the world doesn't behave the way the physicists think it does. This is an example of variables being independent because they are physically independent.

- Roll one six-sided die X, and let $Y = \lceil X/2 \rceil$ and $Z = X \bmod 2$. Then Y and Z are independent, even though they are generated using the same physical process.

12.2.3.2 Independence of many random variables

In general, if we have a collection of random variables X_i, we say that they are all independent if the joint distribution is the product of the marginal distributions, i.e., if $\Pr[\forall i : X_i \le x_i] = \prod_i \Pr[X_i \le x_i]$. It may be that a collection of random variables is not independent even though all subcollections are.

For example, let X and Y be fair coin-flips, and let $Z = X \oplus Y$. Then any two of X, Y, and Z are independent, but the three variables X, Y, and Z are not independent, because $\Pr[X = 0 \wedge Y = 0 \wedge Z = 0] = 1/4$ instead of $1/8$ as one would get by taking the product of the marginal probabilities.

Since we can compute the joint distribution from the marginal distributions for independent variables, we will often just specify the marginal distributions and declare that a collection of random variables are independent. This implicitly gives us an underlying probability space consisting of all sequences of values for the variables.

12.2.4 The expectation of a random variable

For a real-valued random variable X, its **expectation** $\mathrm{E}\left[X\right]$ (sometimes just $\mathrm{E}\, X$) is its average value, weighted by probability.[5] For discrete random variables, the expectation is defined by

$$\mathrm{E}\left[X\right] = \sum_x x \Pr\left[X = x\right].$$

For a continuous random variable with distribution function $F(x)$, the expectation is defined by

$$\mathrm{E}\left[X\right] = \int_{-\infty}^{\infty} x\, dF(x).$$

The integral here is a **Lebesgue-Stieltjes integral**, which generalizes the usual integral for continuous $F(x)$ by doing the right thing if $F(x)$ jumps due to some x that occurs with nonzero probability. We will avoid thinking about this by mostly worrying about expectations for discrete random variables.

Example (discrete variable) Let X be the number rolled with a fair six-sided die. Then $\mathrm{E}\left[X\right] = (1/6)(1+2+3+4+5+6) = 3\frac{1}{2}$.

Example (unbounded discrete variable) Let X be a geometric random variable with parameter p. This means that $\Pr\left[X = k\right] = q^k p$, where as usual $q = 1 - p$. Then $\mathrm{E}[X] = \sum_{k=0}^{\infty} kq^k p = p \sum_{k=0}^{\infty} kq^k = p \cdot \frac{q}{(1-q)^2} = \frac{pq}{p^2} = \frac{q}{p} = \frac{1-p}{p} = \frac{1}{p} - 1$.

Expectation is a way to summarize the distribution of a random variable without giving all the details. If you take the average of many independent copies of a random variable, you will be likely to get a value close to the expectation. Expectations are also used in **decision theory** to compare different choices. For example, given a choice between a 50% chance of

[5] Technically, this will work for any values we can add and multiply by probabilities. So if X is actually a vector in $\mathbb{R}^3$ (for example), we can talk about the expectation of X, which in some sense will be the average position of the location given by X.

winning \$100 (expected value: \$50) and a 20% chance of winning \$1000 (expected value: \$200), a **rational decision maker** would take the second option. Whether ordinary human beings correspond to an economist's notion of a rational decision maker often depends on other details of the situation.

Terminology note: If you hear somebody say that some random variable X takes on the value z **on average**, this usually means that $\mathrm{E}[X] = z$.

12.2.4.1 Variables without expectations

If a random variable has a particularly annoying distribution, it may not have a finite expectation, even thought the variable itself takes on only finite values. This happens if the sum for the expectation diverges.

For example, suppose I start with a dollar, and double my money every time a fair coin-flip comes up heads. If the coin comes up tails, I keep whatever I have at that point. What is my expected wealth at the end of this process?

Let X be the number of times I get heads. Then X is just a geometric random variable with $p = 1/2$, so $\Pr[X = k] = (1 - (1/2))^k (1/2)^k = 2^{-k-1}$. My wealth is also a random variable: 2^X. If we try to compute $\mathrm{E}\left[2^X\right]$, we get

$$\begin{aligned}\mathrm{E}[2^X] &= \sum_{k=0}^{\infty} 2^k \Pr[X = k] \\ &= \sum_{k=0}^{\infty} 2^k \cdot 2^{-k-1} \\ &= \sum_{k=0}^{\infty} \frac{1}{2},\end{aligned}$$

which diverges. Typically we say that a random variable like this has no expected value, although sometimes you will see people writing $\mathrm{E}\left[2^X\right] = \infty$.

(For an even nastier case, consider what happens with $\mathrm{E}\left[(-2)^X\right]$.)

12.2.4.2 Expectation of a sum

The expectation operator is **linear**: this means that $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ and $\mathrm{E}[aX] = a\,\mathrm{E}[X]$ when a is a constant. This fact holds for all random variables X and Y, whether they are independent or not, and is not hard to prove for discrete probability spaces:

$$\begin{aligned}
\mathrm{E}\left[aX+Y\right] &= \sum_{x,y}(ax+y)\Pr\left[X=x\wedge Y=x\right] \\
&= a\sum_{x,y}x\Pr\left[X=x\wedge Y=x\right]+\sum_{x,y}y\Pr\left[X=x\wedge Y=x\right] \\
&= a\sum_{x}x\sum_{y}\Pr\left[X=x\wedge Y=x\right]+\sum_{y}y\sum_{x}\Pr\left[X=x\wedge Y=x\right] \\
&= a\sum_{x}x\Pr\left[X=x\right]+\sum_{y}y\Pr\left[Y=y\right] \\
&= a\,\mathrm{E}\left[X\right]+\mathrm{E}\left[Y\right].
\end{aligned}$$

Linearity of expectation makes computing many expectations easy. Example: Flip a fair coin n times, and let X be the number of heads. What is $\mathrm{E}\left[X\right]$? We can solve this problem by letting X_i be the indicator variable for the event "coin i came up heads." Then $X = \sum_{i=1}^{n} X_i$ and $\mathrm{E}\left[X\right] = \mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}\left[X_i\right] = \sum_{i=1}^{n} \frac{1}{2} = \frac{n}{2}$. In principle it is possible to calculate the same value from the distribution of X (this involves a lot of binomial coefficients), but linearity of expectation is much easier.

Example Choose a random permutation π, i.e., a random bijection from $\{1\dots n\}$ to itself. What is the expected number of values i for which $\pi(i) = i$?

Let X_i be the indicator variable for the event that $\pi(i) = i$. Then we are looking for $\mathrm{E}\left[X_1 + X_2 + \dots X_n\right] = \mathrm{E}\left[X_1\right] + \mathrm{E}\left[X_2\right] + \dots \mathrm{E}\left[X_n\right]$. But $\mathrm{E}\left[X_i\right]$ is just $1/n$ for each i, so the sum is $n(1/n) = 1$. Calculating this by computing $\Pr\left[\sum_{i=1}^{n} X_i = x\right]$ first would be very painful.

12.2.4.3 Expectation of a product

For products of random variables, the situation is more complicated. Here the rule is that $\mathrm{E}\left[XY\right] = \mathrm{E}\left[X\right] \cdot \mathrm{E}\left[Y\right]$ if X and Y are independent. But if X and Y are not independent, the expectation of their product can't be computed without considering their joint distribution.

For example: Roll two dice and take their product. What value do we get on average? The product formula gives $\mathrm{E}\left[XY\right] = \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] = (7/2)^2 = (49/4) = 12\frac{1}{4}$. We could also calculate this directly by summing over all 36 cases, but it would take a while.

Alternatively, roll one die and multiply it by itself. Now what value do we get on average? Here we are no longer dealing with independent random variables, so we have to do it the hard way: $\mathrm{E}\left[X^2\right] = (1^2 + 2^2 + 3^2 + 4^2 +$

$5^2 + 6^2)/6 = 91/6 = 15\frac{1}{6}$. This is substantially higher than when the dice are uncorrelated. (Exercise: How can you rig the second die so it still comes up with each value $\frac{1}{6}$ of the time but *minimizes* $\mathrm{E}[XY]$?)

We can prove the product rule without too much trouble for discrete random variables. The easiest way is to start from the right-hand side.

$$\begin{aligned}
\mathrm{E}[X] \cdot \mathrm{E}[Y] &= \left(\sum_x x \Pr[X = x]\right)\left(\sum_y y \Pr[Y = y]\right) \\
&= \sum_{x,y} xy \Pr[X = x] \Pr[Y = y] \\
&= \sum_z z \left(\sum_{x,y,xy=z} \Pr[X = x] \Pr[Y = y]\right) \\
&= \sum_z z \left(\sum_{x,y,xy=z} \Pr[X = x \wedge Y = y]\right) \\
&= \sum_z z \Pr[XY = z] \\
&= \mathrm{E}[XY].
\end{aligned}$$

Here we use independence in going from $\Pr[X = x] \Pr[Y = y]$ to $\Pr[X = x \wedge Y = y]$ and use the union rule to convert the x, y sum into $\Pr[XY = z]$.

12.2.4.4 Conditional expectation

Like conditional probability, there is also a notion of **conditional expectation**. The simplest version of conditional expectation conditions on a single event A, is written $\mathrm{E}[X \mid A]$, and is defined for discrete random variables by

$$\mathrm{E}[X \mid A] = \sum_x x \Pr[X = x \mid A].$$

This is exactly the same as ordinary expectation except that the probabilities are now all conditioned on A.

To take a simple example, consider the expected value of a six-sided die conditioned on not rolling a 1. The conditional probability of getting 1 is now 0, and the conditional probability of each of the remaining 5 values is 1/5, so we get $(1/5)(2 + 3 + 4 + 5 + 6) = 4$.

Conditional expectation acts very much like regular expectation, so for example we have $\mathrm{E}[aX + bY \mid A] = a\,\mathrm{E}[X \mid A] + b\,\mathrm{E}[Y \mid A]$.

One of the most useful applications of conditional expectation is that it allows computing (unconditional) expectations by case analysis, using the fact that

$$\mathrm{E}\left[X\right] = \mathrm{E}\left[X \mid A\right]\Pr\left[A\right] + \mathrm{E}\left[X \mid \neg A\right]\Pr\left[\neg A\right].$$

or, more generally,

$$\mathrm{E}\left[X\right] = \sum_i \mathrm{E}\left[X \mid A_i\right]\Pr\left[A_i\right]$$

when $A_1, A_2, \dots$ are disjoint events whose union is the entire probability space Ω. This is the expectation analog of the law of total probability.

Examples

- I have a 50% chance of reaching the top of Mt Everest, where Sir Edmund Hilary and Tenzing Norgay hid somewhere between 0 and 10 kilograms of gold (a random variable with uniform distribution). How much gold do I expect to bring home? Compute

  $$\begin{aligned}\mathrm{E}\left[X\right] &= \mathrm{E}\left[X \mid \text{reached the top}\right]\Pr\left[\text{reached the top}\right] + \mathrm{E}\left[X \mid \text{didn't}\right]\Pr\left[\text{didn't}\right] \\ &= 5 \cdot 0.5 + 0 \cdot 0.5 = 2.5.\end{aligned}$$

- Suppose I flip a coin that comes up heads with probability p until I get heads. How many times on average do I flip the coin?

  We'll let X be the number of coin flips. Conditioning on whether the coin comes up heads on the first flip gives $\mathrm{E}\left[X\right] = 1 \cdot p + (1 + \mathrm{E}\left[X'\right]) \cdot (1-p)$, where X' is random variable counting the number of coin-flips needed to get heads ignoring the first coin-flip. But since X' has the same distribution as X, we get $\mathrm{E}\left[X\right] = p + (1-p)(1+\mathrm{E}\left[X\right])$ or $\mathrm{E}\left[X\right] = \frac{p+(1-p)}{p} = 1/p$. So a fair coin must be flipped twice on average to get a head, which is about what we'd expect if we hadn't thought about it much.

- Suppose I have my experimental test subjects complete a task that gets scored on a scale of 0 to 100. I decide to test whether rewarding success is a better strategy for improving outcomes than punishing failure. So for any subject that scores high than 50, I give them a chocolate bar. For any subject that scores lower than 50, I give them an electric shock. (Students who score exactly 50 get nothing.) I then have them each perform the task a second time and measure the average change in their scores. What happens?

Let's suppose that there is no effect whatsoever of my rewards and punishments, and that each test subject obtains each possible score with equal probability 1/101. Now let's calculate the average improvement for test subjects who initially score less than 50 or greater than 50. Call the outcome on the first test X and the outcome on the second test Y. The change in the score is then $Y - X$.

In the first case, we are computing $\mathrm{E}\left[Y - X \mid X < 50\right]$. This is the same as $\mathrm{E}\left[Y \mid X < 50\right] - \mathrm{E}\left[X \mid X < 50\right] = \mathrm{E}\left[Y\right] - \mathrm{E}\left[X \mid X < 50\right] = 50 - 24.5 = +25.5$. So punishing failure produces a 25.5 point improvement on average.

In the second case, we are computing $\mathrm{E}\left[Y - X \mid X > 50\right]$. This is the same as $\mathrm{E}\left[Y \mid X > 50\right] - \mathrm{E}\left[X \mid X > 50\right] = \mathrm{E}\left[Y\right] - \mathrm{E}\left[X \mid X > 50\right] = 50 - 75.5 = -25.5$. So rewarding success produces a 25.5 point decline on average.

Clearly this suggests that we punish failure if we want improvements and reward success if we want backsliding. This is intuitively correct: punishing failure encourages our slacker test subjects to do better next time, while rewarding success just makes them lazy and complacent. But since the test outcomes don't depend on anything we are doing, we get exactly the same answer if we reward failure and punish success: in the former case, a +25.5 point average change, in the later a −25.5 point average change. This is also intuitively correct: rewarding failure makes our subjects like the test so that they will try to do better next time, while punishing success makes them feel that it isn't worth it. From this we learn that our intuitions[6] provide powerful tools for rationalizing almost any outcome in terms of the good or bad behavior of our test subjects. A more careful analysis shows that we performed the wrong comparison, and we are the victim of **regression to the mean**. This phenomenon was one of several now-notorious cognitive biases described in a famous paper by Tversky and Kahneman [TK74].

For a real-world example of how similar problems can arise in processing data, the United States Bureau of Labor Statistics defines a small business as any company with 500 or fewer employees. So if a company has 400 employees in 2007, 600 in 2008, and 400 in 2009, then we just saw a net creation of 200 new jobs by a small business in 2007, followed by the destruction of 200 jobs by a large business in 2008. It has been argued that this effect accounts for much of the observed fact

[6] OK, my intuitions.

that small businesses generate proportionally more new jobs than large ones, although the details are tricky [NWZ11].

12.2.4.5 Conditioning on a random variable

There is a more general notion of conditional expectation for random variables, where the conditioning is done on some other random variable Y. Unlike $\mathrm{E}[X \mid A]$, which is a constant, the expected value of X conditioned on Y, written $\mathrm{E}[X \mid Y]$, is itself a random variable: when $Y = y$, it takes on the value $\mathrm{E}[X \mid Y = y]$.

Here's a simple example. Let's compute $\mathrm{E}[X + Y \mid X]$ where X and Y are the values of independent six-sided dice. When $X = x$, $\mathrm{E}[\mathrm{E}[X + Y \mid X] \mid X = x] = \mathrm{E}[X + Y \mid X = x] = x + \mathrm{E}[Y] = x + 7/2$. For the full random variable we can write $\mathrm{E}[X + Y \mid X] = X + 7/2$.

Another way to get the result in the preceding example is to use some general facts about conditional expectation:

- $\mathrm{E}[aX + bY \mid Z] = a\,\mathrm{E}[X \mid Z] + b\,\mathrm{E}[Y \mid Z]$. This is the conditional-expectation version of linearity of expectation.

- $\mathrm{E}[X \mid X] = X$. This is immediate from the definition, since $\mathrm{E}[X \mid X = x] = x$.

- If X and Y are independent, then $\mathrm{E}[Y \mid X] = \mathrm{E}[Y]$. The intuition is that knowing the value of X gives no information about Y, so $\mathrm{E}[Y]\, X = x = \mathrm{E}[Y]$ for any x in the range of X. (To do this formally requires using the fact that $\Pr[Y = y \mid X = x] = \frac{\Pr[Y=y \wedge X=x]}{\Pr[X=x]} = \frac{\Pr[Y=y]\Pr[X=x]}{\Pr[X=x]} = \Pr[Y = y]$, provided X and Y are independent and $\Pr[X = x] \neq 0$.)

- Also useful: $\mathrm{E}[\mathrm{E}[X \mid Y]] = \mathrm{E}[X]$. Averaging a second time removes all dependence on Y.

These in principle allow us to do very complicated calculations involving conditional expectation.

Some examples:

- Let X and Y be the values of independent six-sided dice. What is $\mathrm{E}[X \mid X + Y]$? Here we observe that $X + Y = \mathrm{E}[X + Y \mid X + Y] = \mathrm{E}[X \mid X + Y] + \mathrm{E}[Y \mid X + Y] = 2\,\mathrm{E}[X \mid X + Y]$ by symmetry. So $\mathrm{E}[X \mid X + Y] = (X + Y)/2$. This is pretty much what we'd expect: on average, half the total value is supplied by one of the dice. (It also works well for extreme cases like $X+Y=12$ or $X+Y=2$, giving a quick check on the formula.)

- What is $\mathrm{E}\left[(X+Y)^2 \mid X\right]$ when X and Y are independent? Compute $\mathrm{E}\left[(X+Y)^2 \mid X\right] = \mathrm{E}\left[X^2 \mid X\right] + 2\,\mathrm{E}\left[XY \mid X\right] + \mathrm{E}\left[Y^2 \mid X\right] = X^2 + 2X\,\mathrm{E}\left[Y\right] + \mathrm{E}\left[Y^2\right]$. For example, if X and Y are independent six-sided dice we have $\mathrm{E}\left[(X+Y)^2 \mid X\right] = X^2 + 7X + 91/6$, so if you are rolling the dice one at a time and the first one comes up 5, you can expect on average to get a squared total of $25 + 35 + 91/6 = 75\frac{1}{6}$. But if the first one comes up 1, you only get $1 + 7 + 91/6 = 23\frac{1}{6}$ on average.

12.2.5 Markov's inequality

Knowing the expectation of a random variable gives you some information about it, but different random variables may have the same expectation but very different behavior: consider, for example, the random variable X that is 0 with probability 1/2 and 1 with probability 1/2 and the random variable Y that is 1/2 with probability 1. In some cases we don't care about the average value of a variable so much as its likelihood of reaching some extreme value: for example, if my feet are encased in cement blocks at the beach, knowing that the average high tide is only 1 meter is not as important as knowing whether it ever gets above 2 meters. **Markov's inequality** lets us bound the probability of unusually high values of *non-negative* random variables as a function of their expectation. It says that, for any $a > 0$,

$$\Pr\left[X > a\,\mathrm{E}\left[X\right]\right] < 1/a.$$

This can be proved easily using conditional expectations. We have:

$$\mathrm{E}\left[X\right] = \mathrm{E}\left[X \mid X > a\,\mathrm{E}\left[X\right]\right] \Pr\left[X > a\,\mathrm{E}\left[X\right]\right] + \mathrm{E}\left[X\right] X \leq a\,\mathrm{E}\left[X\right] \Pr\left[X \leq a\,\mathrm{E}\left[X\right]\right].$$

Since X is non-negative, $\mathrm{E}\left[X \mid X \leq a\,\mathrm{E}\left[X\right]\right] \geq 0$, so subtracting out the last term on the right-hand side can only make it smaller. This gives:

$$\begin{aligned}\mathrm{E}\left[X\right] &\geq \mathrm{E}\left[X \mid X > a\,\mathrm{E}\left[X\right]\right] \Pr\left[X > a\,\mathrm{E}\left[X\right]\right] \\ &> a\,\mathrm{E}\left[X\right] \Pr\left[X > a\,\mathrm{E}\left[X\right]\right],\end{aligned}$$

and dividing both side by $a\,\mathrm{E}\left[X\right]$ gives the desired result.

Another version of Markov's inequality replaces $>$ with $\geq$:

$$\Pr\left[X \geq a\,\mathrm{E}\left[X\right]\right] \leq 1/a.$$

The proof is essentially the same.

12.2.5.1 Example

Suppose that that all you know about the high tide height X is that $\mathrm{E}[X] = 1$ meter and $X \geq 0$. What can we say about the probability that $X > 2$ meters? Using Markov's inequality, we get $\Pr[X > 2 \text{ meters}] = \Pr[X > 2\,\mathrm{E}[X]] < 1/2$.

12.2.5.2 Conditional Markov's inequality

There is, of course, a conditional version of Markov's inequality:

$$\Pr[X > a\,\mathrm{E}[X \mid A] \mid A] < 1/a.$$

This version doesn't get anywhere near as much use as the unconditioned version, but it may be worth remembering that it exists.

12.2.6 The variance of a random variable

Expectation tells you the average value of a random variable, but it doesn't tell you how far from the average the random variable typically gets: the random variables $X = 0$ and $Y = \pm 1,000,000,000,000$ with equal probability both have expectation 0, though their distributions are very different. Though it is impossible to summarize everything about the spread of a distribution in a single number, a useful approximation for many purposes is the **variance** $\mathrm{Var}[X]$ of a random variable X, which is defined as the expected square of the deviation from the expectation, or $\mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$.

Example Let X be 0 or 1 with equal probability. Then $\mathrm{E}[X] = 1/2$, and $(X - \mathrm{E}[X])^2$ is always $1/4$. So $\mathrm{Var}[X] = 1/4$.

Example Let X be the value of a fair six-sided die. Then $\mathrm{E}[X] = 7/2$, and $\mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \frac{1}{6}\left((1-7/2)^2 + (2-7/2)^2 + (3-7/2)^2 + \cdots + (6-7/2)^2\right) = 35/12$.

Computing variance directly from the definition can be tedious. Often it is easier to compute it from $\mathrm{E}[X^2]$ and $\mathrm{E}[X]$:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + (\mathrm{E}[X])^2 \\
&= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.
\end{aligned}$$

The second-to-last step uses linearity of expectation and the fact that $\mathrm{E}[X]$ is a constant.

Example For X being 0 or 1 with equal probability, we have $\mathrm{E}\left[X^2\right] = 1/2$ and $(\mathrm{E}[X])^2 = 1/4$, so $\mathrm{Var}[X] = 1/4$.

Example Let's try the six-sided die again, except this time we'll use an n-sided die. We have

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} i^2 - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\
&= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}.
\end{aligned}$$

When $n = 6$, this gives $\frac{7\cdot 13}{6} - \frac{49}{4} = \frac{35}{12}$. (Ok, maybe it isn't always easier).

12.2.6.1 Multiplication by constants

Suppose we are asked to compute the variance of cX, where c is a constant. We have

$$\begin{aligned}
\mathrm{Var}[cX] &= \mathrm{E}\left[(cX)^2\right] - \mathrm{E}[cX]^2 \\
&= c^2\,\mathrm{E}\left[X^2\right] - (c\,\mathrm{E}[X])^2 \\
&= c^2\,\mathrm{Var}[X].
\end{aligned}$$

So, for example, if X is 0 or 2 with equal probability, $\mathrm{Var}[X] = 4\cdot(1/4) = 1$. This is exactly what we expect given that $X - \mathrm{E}[X]$ is always ±1.

Another consequence is that $\mathrm{Var}[-X] = (-1)^2\,\mathrm{Var}[X] = \mathrm{Var}[X]$. So variance is not affected by negation.

12.2.6.2 The variance of a sum

What is $\operatorname{Var}[X+Y]$? Write

$$\begin{aligned}\operatorname{Var}[X+Y] &= \operatorname{E}\left[(X+Y)^2\right] - (\operatorname{E}[X+Y])^2 \\ &= \operatorname{E}\left[X^2\right] + 2\operatorname{E}[XY] + \operatorname{E}\left[Y^2\right] - (\operatorname{E}[X])^2 - 2\operatorname{E}[X]\cdot\operatorname{E}[Y] - (\operatorname{E}[Y])^2 \\ &= (\operatorname{E}\left[X^2\right] - (\operatorname{E}[X])^2) + (\operatorname{E}\left[Y^2\right] - (\operatorname{E}[Y])^2) + 2(\operatorname{E}[XY] - \operatorname{E}[X]\cdot\operatorname{E}[Y]) \\ &= \operatorname{Var}[X] + \operatorname{Var}[Y] + 2(\operatorname{E}[XY] - \operatorname{E}[X]\cdot\operatorname{E}[Y]).\end{aligned}$$

The quantity $\operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]$ is called the **covariance** of X and Y and is written $\operatorname{Cov}[X,Y]$. So we have just shown that

$$\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}[X,Y].$$

When $\operatorname{Cov}[X,Y] = 0$, or equivalently when $\operatorname{E}[XY] = \operatorname{E}[X]\operatorname{E}[Y]$, X and Y are said to be **uncorrelated** and their variances add. This occurs when X and Y are independent, but may also occur without X and Y being independent.

For larger sums the corresponding formula is

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \operatorname{Var}[X_i] + \sum_{i\neq j} \operatorname{Cov}[X_i, X_j].$$

This simplifies to $\operatorname{Var}[\sum X_i] = \sum \operatorname{Var}[X_i]$ when the X_i are **pairwise independent**, so that each pair of distinct X_i and X_j are independent. Pairwise independence is implied by independence (but is not equivalent to it), so this also works for fully independent random variables.

For example, we can use the simplified formula to compute the variance of the number of heads in n independent fair coin-flips. Let X_i be the indicator variable for the event that the i-th flip comes up heads and let X be the sum of the X_i. We have already seen that $\operatorname{Var}[X_i] = 1/4$, so $\operatorname{Var}[X] = n\operatorname{Var}[X_i] = n/4$.

Similarly, if c is a constant, then we can compute $\operatorname{Var}[X+c] = \operatorname{Var}[X] + \operatorname{Var}[c] = \operatorname{Var}[X]$, since (1) $\operatorname{E}[cX] = c\operatorname{E}[X] = \operatorname{E}[c]\operatorname{E}[X]$ means that c (considered as a random variable) and X are uncorrelated, and (2) $\operatorname{Var}[a] = \operatorname{E}\left[(c - \operatorname{E}[c])^2\right] = \operatorname{E}[0] = 0$. So shifting a random variable up or down doesn't change its variance.

12.2.6.3 Chebyshev's inequality

Variance is an expectation, so we can use Markov's inequality on it. The result is **Chebyshev's inequality**, which like Markov's inequality comes in two versions:

$$\begin{aligned} \Pr\left[\left|X - \mathrm{E}\left[X\right]\right| \geq r\right] &\leq \frac{\mathrm{Var}\left[X\right]}{r^2}, \\ \Pr\left[\left|X - \mathrm{E}\left[X\right]\right| > r\right] &< \frac{\mathrm{Var}\left[X\right]}{r^2}. \end{aligned}$$

Proof. We'll do the first version. The event $|X - \mathrm{E}[X]| \geq r$ is the same as the event $(X - \mathrm{E}[X])^2 \geq r^2$. By Markov's inequality, the probability that this occurs is at most $\frac{\mathrm{E}\left[(X-\mathrm{E}[X])^2\right]}{r^2} = \frac{\mathrm{Var}[X]}{r^2}$. □

Application: showing that a random variable is close to its expectation This is the usual statistical application.

Example Flip a fair coin n times, and let X be the number of heads. What is the probability that $|X - n/2| > r$? Recall that $\mathrm{Var}[X] = n/4$, so $\Pr[|X - n/2| > r] < (n/4)/r^2 = n/(4r^2)$. So, for example, the chances of deviating from the average by more than 1000 after 1000000 coin-flips is less than 1/4.

Example Out of n voters in Saskaloosa County, m plan to vote for Smith for County Dogcatcher. A polling firm samples k voters (with replacement) and asks them who they plan to vote for. Suppose that $m < n/2$; compute a bound on the probability that the polling firm incorrectly polls a majority for Smith.

Solution: Let X_i be the indicator variable for a Smith vote when the i-th voter is polled and let $X = \sum X_i$ be the total number of pollees who say they will vote for Smith. Let $p = \mathrm{E}[X_i] = m/n$. Then $\mathrm{Var}[X_i] = p - p^2$, $\mathrm{E}[X] = kp$, and $\mathrm{Var}[X] = k(p - p^2)$. To get a majority in the poll, we need $X > k/2$ or $X - \mathrm{E}[X] > k/2 - kp$. Using Chebyshev's inequality, this event occurs with probability at most

$$\begin{aligned} \frac{\mathrm{Var}\left[X\right]}{(k/2 - kp)^2} &= \frac{k(p - p^2)}{(k/2 - kp)^2} \\ &= \frac{1}{k} \cdot \frac{p - p^2}{(1/2 - p)^2}. \end{aligned}$$

Note that the bound decreases as k grows and (for fixed p) does not depend on n.

In practice, statisticians will use a stronger result called the **central limit theorem**, which describes the shape of the distribution of the sum of many independent random variables much more accurately than the bound from Chebyshev's inequality. Designers of randomized algorithms are more likely to use **Chernoff bounds**.

Application: lower bounds on random variables Unlike Markov's inequality, which can only show that a random variable can't be too big too often, Chebyshev's inequality can be used to show that a random variable can't be too small, by showing first that its expectation is high and then that its variance is low. For example, suppose that each of the 10^{30} oxygen molecules in the room is close enough to your mouth to inhale with pairwise independent probability 10^{-4} (it's a big room). Then the expected number of oxygen molecules near your mouth is a healthy $10^{30} \cdot 10^{-4} = 10^{26}$. What is the probability that all 10^{26} of them escape your grasp?

Let X_i be the indicator variable for the event that the i-th molecule is close enough to inhale. We've effectively already used the fact that $\operatorname{E}[X_i] = 10^{-4}$. To use Chebyshev's inequality, we also need $\operatorname{Var}[X_i] = \operatorname{E}\left[X_i^2\right] - \operatorname{E}[X_i]^2 = 10^{-4} - 10^{-8} \approx 10^{-4}$. So the total variance is about $10^{30} \cdot 10^{-4} = 10^{26}$ and Chebyshev's inequality says we have $\Pr\left[|X - \operatorname{E}[X]| \geq 10^{26}\right] \leq 10^{26}/(10^{26})^2 = 10^{-26}$. So death by failure of statistical mechanics is unlikely (and the real probability is much much smaller).

But wait! Even a mere 90% drop in O_2 levels is going to be enough to cause problems. What is the probability that this happens? Again we can calculate $\Pr[90\% \text{ drop}] \leq \Pr\left[|X - \operatorname{E}[X]| \geq 0.9 \cdot 10^{26}\right] \leq 10^{26}/(0.9 \cdot 10^{26})^2 \approx 1.23 \cdot 10^{-26}$. So even temporary asphyxiation by statistical mechanics is not something to worry about.

12.2.7 Probability generating functions

For a discrete random variable X taking on only values in $\mathbb{N}$, we can express its distribution using a **probability generating function** or **pgf**:

$$F(z) = \sum_{n=0}^{\infty} \Pr[X = n]\, z^n.$$

These are essentially standard-issue generating functions (see §11.3) with the additional requirement that all coefficients are non-negative and $F(1) = 1$.

A trivial example is the pgf for a Bernoulli random variable (1 with probability p, 0 with probability $q = 1 - p$). Here the pgf is just $q + pz$.

A more complicated example is the pgf for a geometric random variable. Now we have $\sum_{n=0}^{\infty} q^n p z^n = p \sum_{n=0}^{\infty} (qz)^n = \frac{p}{1-qz}$.

12.2.7.1 Sums

A very useful property of pgf's is that the pgf of a sum of independent random variables is just the product of the pgf's of the individual random variables. The reason for this is essentially the same as for ordinary generating functions: when we multiply together two terms $(\Pr[X = n] z^n)(\Pr[Y = m] z^m)$, we get $\Pr[X = n \wedge Y = m] z^{n+m}$, and the sum over all the different ways of decomposing $n + m$ gives all the different ways to get this sum.

So, for example, the pgf of a binomial random variable equal to the sum of n independent Bernoulli random variables is $(q + pz)^n$ (hence the name "binomial").

12.2.7.2 Expectation and variance

One nice thing about pgf's is that the can be used to quickly compute expectation and variance. For expectation, we have

$$F'(z) = \sum_{n=0}^{\infty} n \Pr[X = n] z^{n-1}.$$

So

$$\begin{aligned} F'(1) &= \sum_{n=0}^{\infty} n \Pr[X = n] \\ &= \mathrm{E}[X]. \end{aligned}$$

If we take the second derivative, we get

$$F''(z) = \sum_{n=0}^{\infty} n(n-1) \Pr[X = n] z^{n-1}$$

or

$$\begin{aligned} F''(1) &= \sum_{n=0}^{\infty} n(n-1) \Pr[X = n] \\ &= \mathrm{E}[X(X-1)] \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]. \end{aligned}$$

So we can recover $\mathrm{E}[X^2]$ as $F''(1) + F'(1)$ and get $\mathrm{Var}[X]$ as $F''(1) + F'(1) - (F'(1))^2$.

Example If X is a Bernoulli random variable with pgf $F = (q + pz)$, then $F' = p$ and $F'' = 0$, giving $\mathrm{E}[X] = F'(1) = p$ and $\mathrm{Var}[X] = F''(1) + F'(1) - (F'(1))^2 = 0 + p - p^2 = p(1 - p) = pq$.

Example If X is a binomial random variable with pgf $F = (q + pz)^n$, then $F' = n(q + pz)^{n-1}p$ and $F'' = n(n - 1)(q + pz)^{n-2}p^2$, giving $\mathrm{E}[X] = F'(1) = np$ and $\mathrm{Var}[X] = F''(1) + F'(1) - (F'(1))^2 = n(n - 1)p^2 + np - n^2p^2 = np - np^2 = npq$. These values would, of course, be a lot faster to compute using the formulas for sums of independent random variables, but it's nice to see that they work.

Example If X is a geometric random variable with pgf $p/(1 - qz)$, then $F' = pq/(1 - qz)^2$ and $F'' = 2pq^2/(1 - qz)^3$. So $\mathrm{E}[X] = F'(1) = pq/(1 - q)^2 = pq/p^2 = q/p$, and $\mathrm{Var}[X] = F''(1) + F'(1) - (F'(1))^2 = 2pq^2/(1-q)^3 + q/p - q^2/p^2 = 2q^2/p^2 + q/p - q^2/p^2 = q^2/p^2 + q/p$. The variance would probably be a pain to calculate by hand.

Example Let X be a Poisson random variable with rate λ. We claimed earlier that a Poisson random variable is the limit of a sequence of binomial random variables where $p = \lambda/n$ and n goes to infinity, so (cheating quite a bit) we expect that X's pgf $F = \lim_{n\to\infty}((1 - \lambda/n) + (\lambda/n)z)^n = (1 + (-\lambda + \lambda z)/n)^n = \exp(-\lambda + \lambda z) = \exp(-\lambda)\sum \lambda^n z^n/n!$. We can check that the total probability $F(1) = \exp(-\lambda + \lambda) = e^0 = 1$, that the expectation $F'(1) = \lambda \exp(-\lambda + \lambda) = \lambda$, and that the variance $F''(1) + F'(1) - (F'(1))^2 = \lambda^2 \exp(-\lambda + \lambda) + \lambda - \lambda^2 = \lambda$. These last two quantities are what we'd expect if we calculated the expectation and the variance directly as the limit of taking n Bernoulli random variables with expectation λ/n and variance $(\lambda/n)(1 - \lambda/n)$ each.

12.2.8 Summary: effects of operations on expectation and variance of random variables

$$\begin{aligned}
\mathrm{E}[X+Y] &= \mathrm{E}[X] + \mathrm{E}[Y] & \mathrm{Var}[X+Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X,Y] \\
\mathrm{E}[aX] &= a\,\mathrm{E}[X] & \mathrm{Var}[aX] &= a^2\,\mathrm{Var}[X] \\
\mathrm{E}[XY] &= \mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{Cov}[X,Y]
\end{aligned}$$

For the second line, a is a constant. None of these formulas assume independence, although we can drop $\mathrm{Cov}[X, Y]$ (because it is zero) whenever X and Y are independent. There is no simple formula for $\mathrm{Var}[XY]$.

The expectation and variance of $X - Y$ can be derived from the rules for addition and multiplication by a constant:

$$\begin{aligned}
\mathrm{E}\left[X - Y\right] &= \mathrm{E}\left[X + (-Y)\right] \\
&= \mathrm{E}\left[X\right] + \mathrm{E}\left[-Y\right] \\
&= \mathrm{E}\left[X\right] - \mathrm{E}\left[Y\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}\left[X - Y\right] &= \mathrm{Var}\left[X + (-Y)\right] \\
&= \mathrm{Var}\left[X\right] + \mathrm{Var}\left[-Y\right] + 2\,\mathrm{Cov}\left[X, -Y\right] \\
&= \mathrm{Var}\left[X\right] + \mathrm{Var}\left[Y\right] - 2\,\mathrm{Cov}\left[X, Y\right].
\end{aligned}$$

12.2.9 The general case

So far we have only considered discrete random variables, which avoids a lot of nasty technical issues. In general, a **random variable** on a probability space $(\Omega, \mathcal{F}, P)$ is a function whose domain is Ω that satisfies some extra conditions on its values that make interesting events involving the random variable elements of $\mathcal{F}$. Typically the codomain will be the reals or the integers, although any set is possible. Random variables are generally written as capital letters with their arguments suppressed: rather than writing $X(\omega)$, where $\omega \in \Omega$, we write just X.

A technical condition on random variables is that the inverse image of any measurable subset of the codomain must be in $\mathcal{F}$—in simple terms, if you can't nail down ω exactly, being able to tell which element of $\mathcal{F}$ you land in should be enough to determine the value of $X(\omega)$. For a discrete random variables, this just means that $X^{-1}(x) \in \mathcal{F}$ for each possible value x. For real-valued random variables, the requirement is that the event $[X \leq x]$ is in $\mathcal{F}$ for any fixed x. In each case we say that X is **measurable** with respect to $\mathcal{F}$ (or just "measurable $\mathcal{F}$").[7] Usually we will not worry about this issue too much, but it may come up if we are varying $\mathcal{F}$ to represent different amounts of information available to different observers (e.g., if X and Y are the values of two dice, X is measurable to somebody who can see both dice but not to somebody who can only see the sum of the dice).

The **distribution function** of a real-valued random variable describes the probability that it takes on each of its possible values; it is specified

[7] The detail we are sweeping under the rug here is what makes a subset of the codomain measurable. The essential idea is that we also have a σ-algebra $\mathcal{F}'$ on the codomain, and elements of this codomain σ-algebra are the measurable subsets. The rules for simple random variables and real-valued random variables come from default choices of σ-algebra.

by giving a function $F(x) = \Pr[X \leq x]$. The reason for using $\Pr[X \leq x]$ instead of $\Pr[X = x]$ is that it allows specifying continuous random variables such as a random variable that is uniform in the range $[0, 1]$; this random variable has a distribution function given by $F(x) = x$ when $0 \leq x \leq 1$, $F(x) = 0$ for $x < 0$, and $F(x) = 1$ for $x > 1$.

For discrete random variables the distribution function will have discontinuous jumps at each possible value of the variable. For example, the distribution function of a variable X that is 0 or 1 with equal probability is $F(x) = 0$ for $x < 0$, $1/2$ for $0 \leq x < 1$, and 1 for $x \geq 1$.

Knowing the distribution of a random variable tells you what that variable might do by itself, but doesn't tell you how it interacts with other random variables. For example, if X is 0 or 1 with equal probability then X and $1-X$ both have the same distribution, but they are connected in a way that is not true for X and some independent variable Y with the same distribution. For multiple variables, a **joint distribution** gives the probability that each variable takes on a particular value; for example, if X and Y are two independent uniform samples from the range $[0, 1]$, their distribution function $F(x, y) = \Pr[X \leq x \wedge Y \leq y] = xy$ (when $0 \leq x, y \leq 1$). If instead $Y = 1 - X$, we get the distribution function $F(x, y) = \Pr[X \leq x \wedge Y \leq y]$ equal to x when $y \geq 1 - x$ and 0 when $y < 1 - x$ (assuming $0 \leq x, y \leq 1$).

We've seen that for discrete random variables, it is more useful to look at the **probability mass function** $f(x) = \Pr[X = x]$. We can always recover the probability distribution function from the probability mass function if the latter sums to 1.

12.2.9.1 Densities

If a real-valued random variable is **continuous** in the sense of having a distribution function with no jumps (which means that it has probability 0 of landing on any particular value), we may be able to describe its distribution by giving a **density** instead. The density is the derivative of the distribution function. We can also think of it as a probability at each point defined in the limit, by taking smaller and smaller regions around the point and dividing the probability of landing in the region by the size of the region.

For example, the density of a uniform $[0, 1]$ random variable is $f(x) = 1$ for x in $[0, 1]$, and $f(x) = 0$ otherwise. For a uniform $[0, 2]$ random variable, we get a density of $\frac{1}{2}$ throughout the $[0, 2]$ interval. The density always integrates to 1.

Some distributions are easier to describe using densities than using distribution functions. The **normal distribution**, which is of central importance in statistics, has density

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Its distribution function is the integral of this quantity, which has no closed-form expression.

Joint densities also exist. The **joint density** of a pair of random variables with joint distribution function $F(x, y)$ is given by the partial derivative $f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$. The intuition here again is that we are approximating the (zero) probability at a point by taking the probability of a small region around the point and dividing by the area of the region.

12.2.9.2 Independence

Independence is the same as for discrete random variables: Two random variables X and Y are **independent** if any pair of events of the form $X \in A$, $Y \in B$ are independent. For real-valued random variables it is enough to show that their joint distribution $F(x, y)$ is equal to the product of their individual distributions $F_X(x) F_Y(y)$. For real-valued random variables with densities, showing the densities multiply also works. Both methods generalize in the obvious way to sets of three or more random variables.

12.2.9.3 Expectation

If a continuous random variable has a density $f(x)$, the formula for its expectation is

$$\mathrm{E}[X] = \int x f(x) \, dx.$$

For example, let X be a uniform random variable in the range $[a, b]$. Then $f(x) = \frac{1}{b-a}$ when $a \le x \le b$ and 0 otherwise, giving

$$\begin{aligned}
\mathrm{E}[X] &= \int_a^b x \frac{1}{b-a} \, dx \\
&= \left. \frac{x^2}{2(b-a)} \right|_{x=a}^{b} \\
&= \frac{b^2 - a^2}{2(b-a)} \\
&= \frac{a+b}{2}.
\end{aligned}$$

For continuous random variables without densities, we land in a rather swampy end of integration theory. We will not talk about this case if we can help it. But in each case the expectation depends only on the distribution of X and not on its relationship to other random variables.