Z. Abel, B. Chapman, E. Demaine
revised Tuesday 7th May, 2024

Lecture 21: Random Variables

1 Random variables

Definition 1. *A* random variable *(RV) is a total function whose domain is the sample space.*

For instance, let's suppose that our sample space corresponds to flipping three fair independent coins. What are some possible RVs?

1. The *value of the 1st coin* (H or T, or we could encode it as 0/1, or however we want).
2. The *number of heads* (we'll call this one R).
3. The function that's 1 if all three coin flips match and 0 otherwise (we'll call this M).

A 0/1-valued RV is called an *indicator*.

RVs naturally give rise to events: for an RV f and a value x, we define the event $f = x$ by the set of outcomes ω in the sample space for which $f(\omega) = x$. (Remember: an event is a *set*. An RV is a *function*). Vice versa, every event A corresponds to an indicator RV $\mathbf{1}[A]$ which is equal to 1 for all $\omega \in A$, and 0 for all $\omega \notin A$.

Another type of event we can define for an RV: the event $f \geq x$, defined as the set of outcomes ω such that $f(\omega) \geq x$. We can write the probability of such an event using the sum rule

$$\Pr[f \geq x] = \sum_{y \geq x} \Pr[f = y].$$

Or more generally, if we have some subset T of values in the range of X, we can define the event $f \in T$ by the set of outcomes ω for which $f(\omega) \in T$.

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

2 Conditioning and independence

Of course, these events can be conditioned on just like any other event. E.g.

$$\Pr[R = 2 \mid M = 1] = \frac{\Pr[R = 2 \cap M = 1]}{\Pr[M = 1]} = 0.$$

We can also extend the notion of independence to RVs. Careful: this notion is a bit different than what we did for events!

Definition 2. *Two RVs X, Y are independent if for* all *values x, y in their range,* $\Pr[X = x \cap Y = y] = \Pr[X = x] \cdot \Pr[Y = y]$.

Or equivalently, if for all values x, y, either $\Pr[Y = y] = 0$*, or* $\Pr[X = x \mid Y = y] = \Pr[X = x]$.

So we're saying that two RVs are independent if *all* the pairs of events $X = x$ and $Y = y$ are independent. Intuitively, this says that learning a value for Y, no matter what value you learn, reveals no additional information about X.

Example: R and M are not independent! We just showed that $\Pr[R = 2 \cap M = 1] = 0$, but $\Pr[R = 2] \neq 0$ and $\Pr[M = 1] \neq 0$.

Another example: suppose we roll two fair dice to get values D_1, D_2 (these are our two RVs). Let $S = D_1 + D_2$. This is an RV too. Let's define the r.v. $T = \mathbf{1}[S = 7]$. This is 1 if the two rolls add up to 7, and 0 otherwise. (It's an indicator!)

Are D_1 and S independent? No: intuitively, if you learn that $D_1 = 6$, it's impossible for S to be any smaller than 7. Or by the definition, $0 = \Pr[S = 5 \cap D_1 = 6] \neq \Pr[S = 5] \Pr[D_1 = 6]$.

What about S and T? Obviously not since T is just a function of S.

What about T and D_1? It turns out they are independent!

$$\Pr[T = 1 \mid D_1 = d] = 1/6 = \Pr[T = 1].$$

This is because for each value of D_1, there is always exactly one value for D_2 that causes them to add up to 7. So it's caused by a special symmetry of the problem.

Definition 3. *A collection of RVs $X_1, \dots, X_n$ is* mutually independent *if for all values $x_1, \dots, x_n$, it holds that*

$$\Pr[X_1 = x_1, \dots, X_n = x_n] = \Pr[X_1 = x_1] \dots \Pr[X_n = x_n].$$

Note that we don't need to check all subsets of all size. This is because the condition is implied by the above: we can obtain the analogous equations for subsets of smaller size by *summing* over values. E.g.

$$\sum_{x_3} \Pr[X_1 = x_1, X_2 = x_2, X_3 = x_3] = \sum_{x_3} \Pr[X_1 = x_1] \Pr[X_2 = x_2] \Pr[X_3 = x_3]$$
$$\Pr[X_1 = x_1, X_2 = x_2] = \Pr[X_1 = x_1] \Pr[X_2 = x_2] \cdot 1.$$

3 Distributions, PMFs and CDFs

For any RV X, define the *probability mass function*

$$f(x) = \Pr[X = x],$$

and the *cumulative distribution function*

$$F(x) = \sum_{y \leq x} \Pr[X = y].$$

These are two different (equivalent!) ways to express the *probability distribution* of an RV. It's often to describe random variables by their distributions.

Some common cases:

- Indicator random variables, which are also called *Bernoulli*.

$$f(0) = p, f(1) = 1 - p, F(0) = p, F(1) = 1.$$

- A uniform random variable on $\{1, 2, \ldots, n\}$

$$\forall i \in \{1, 2, \ldots, n\}, f(i) = \frac{1}{n}, F(i) = \frac{i}{n}.$$

4 Two envelope problem

An example of where randomness is useful in solving a task. Suppose I prepare two envelopes each containing an unknown integer between 0 and 100 dollars (and suppose the values are not equal). I hand them to you and ask you to choose an envelope. What's your chance of choosing the envelope with the greater number Clearly, no better than 1/2.

Now, suppose I let you peek inside the envelope you chose, learning the number inside it. I now offer you the opportunity to *switch* envelopes. Should you switch? Sometimes? Always? Never?

Note that I don't promise to pick the numbers in any predetermined random way. I am your *adversary*: you want to figure out a way to play this game that will work *no matter* how I selected the numbers.

If you think about this for a bit, it becomes clear that *always* and *never* switching are both equally good, and neither does any better than 1/2. For if you were always going to do the same thing, you could have done that *before* you got the additional information about the envelope you chose! So you must use the number you observed somehow.

Here's another piece of intuition: suppose (by magic) that you happened to know a number z that lies halfway between the two numbers in the envelope. Then it's clear what to do: if the envelope you saw is above z, you stay; otherwise, you switch! But how can you learn z?

One strategy: if you don't know, guess! That is, let z be a uniformly random number from some set, and see what happens. Let's use our knowledge of random variables to analyze how well this strategy does.

1. Let x_0, x_1 be the two numbers in the envelopes. We have no idea how they're chosen—it's some distribution that the adversary controls.

2. Let z be our randomly chosen threshold. We will choose it uniformly from $\{0.5, 1.5, \ldots, 99.5\}$. Why half-integers? Because we want to avoid z being equal to one of the two amounts. (Just think of the z's as "dividers" on the number line, if you prefer.)

3. Let $r \in \{0, 1\}$ be the index of the randomly chosen envelope 0 or 1. The player is revealed the amount x_r.

4. Our strategy: stick if $x_r > z$, and switch otherwise.

What's the chance that this strategy succeeds? Well observe:

1. If z is *between* x_0, x_1, we succeed with certainty.

2. Otherwise, we succeed with probability 1/2, since in this case we either always switch or never switch regardless of the value of r.

Now what's the chance that the former occurs? For any fixed values of x_0, x_1, there's at least 1 value of z that works. So it's always at least 1/100. So we succeed in this game with probability at least

$$\frac{1}{100} \cdot 1 + \frac{99}{100} \cdot \frac{1}{2} = \frac{1}{2} + \frac{1}{200}.$$

Which is better than random guessing!

Once you understand the calculations above, it's worth dwelling on the *conceptual* novelty of what we just did here. So far in this class, we've used probability purely to *model* uncertainty in the world. Here, we used probability to *design an algorithm* to solve a task! This turns out to be an extremely useful technique in many areas of computer science. We'll see more of it in 6.1210 and 6.1220.

5 The binomial distribution

An extremely common probability distribution that arises often in CS is the *binomial* distribution. This distribution has two *parameters*: n and p. This models a bunch of things

1. Flip n independent coins, each of which gives heads with probability p. The random variable X that counts the number of heads is binomial.

2. Suppose you have n components, each of which fails with probability p. The random variable X that counts the number of failures is also binomial.

What's the PMF for this distribution?

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

Intuitively, each outcome that has exactly k heads has probability $p^k(1-p)^{n-k}$ of occurring, and there are $\frac{n}{k}$ of them.

What about the CDF?

$$F_{n,p}(k) = \sum_{j=0}^{k} \Pr[X = j] = \sum_{j=0}^{k} \binom{n}{j} p^j k (1-p)^{n-j}.$$

The following was not covered in lecture and is optional: These formulas are unwieldy. It is good to use Stirling to approximate them to get a sense of what's going on.

$$\begin{aligned}
f_{n,p}(\alpha n) &= \frac{n!}{(\alpha n)!((1-\alpha)n)!} p^{\alpha n}(1-p)^{(1-\alpha)n} \\
&\sim \frac{\sqrt{2\pi n}}{\sqrt{2\pi\alpha n}\sqrt{2\pi(1-\alpha)n}} \frac{(n/e)^n}{(\alpha n/e)^{\alpha n}((1-\alpha)n/e)^{(1-\alpha)n}} p^{\alpha n}(1-p)^{(1-\alpha)n} \\
&= \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} \frac{1}{(1-\alpha)^{(1-\alpha)n}\alpha^{\alpha n}} p^{\alpha n}(1-p)^{(1-\alpha)n} \\
&= \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} \left(\frac{p}{\alpha}\right)^{\alpha n} \left(\frac{1-p}{1-\alpha}\right)^{(1-\alpha)n} \\
&= \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} 2^{(\alpha \log \frac{p}{\alpha} + (1-\alpha)\log\frac{1-p}{1-\alpha})n}.
\end{aligned}$$

It turns out this formula is also an exact upper bound. The exponent turns out to always be negative unless $p = \alpha$. So the max occurs there. What we get when we plot this is a little bump at $p = \alpha$, which gets exponentially smaller everywhere else.

If we try this for $p = 0.5, n = 100$.

1. For $k = 50$, $f(50) \approx 0.08$—pretty small.
2. For $k = 25$, $f(25) \approx 1.9 \cdot 10^{-7}$—extreeeemely small!

We will discuss in the last lecture of term how to get bounds on the CDF, but it turns out that the "tails" of this distribution ($F(k)$ for small k, or $1 - F(k)$ for large k) are *very* small. So it's more likely, if I flip 100 coins, that I'll get exactly 25 heads, than I'll get < 25 heads! (And both are vanishingly unlikely.)