CS174 Lecture 1 John Canny

Randomized Birthday Search

From the table below, copy the number under the month of your birthday onto a piece of paper.

Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sept	Oct	Nov	Dec
323	106	261	13	75	137	354	292	230	168	44	199

Now if your birthday is in the first half of the month, use this table to lookup a second number:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
104	137	168	200	232	264	296	328	112	144	176	208	240	272	304	336

Or if your birthday is in the second half of the month, use this table:

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
120	152	184	216	248	280	312	344	128	160	192	224	256	288	320

Now add the two numbers. If the total is bigger than 372, subtract 372 from it, to get a number in the range 1-372. This is a very simple hash function of your birthday.

The next step is a survey of how many people have numbers in the following intervals:

1-31	32-62	63-93	94-124	125-155	156-186
187-217	218-248	249-279	280-310	311-341	342-372

In a typical class, its usually possible to find one of these groups with at least 9 people in it. Such a group has a good chance (prob. better than 73%) of having a shared birthday in it. The large group exists because the number of students in each group is a random variable with a mean (which is number of students/12) but a “tail” of larger and smaller values which is quite probable.

Question: The protocol given in class allows the rest of the class to notice which two students had the same birthday. Come up with a variation where the two students can discover this, but the rest of the class cannot. Hint: the process is the same except in the student’s choice of what number they announce. You can suppose this hash function were more complicated, and difficult to invert.

This example illustrates several points:

- Probability implies that its very likely certain things happen (two people in class have the same birthday).
- We can use the “tail” of a random variable (number of birthdays in one bucket) to show that what we’re looking for is likely somewhere.
- We can use random analysis as a fast “filter” to hone in on a likely solution.
- One step of the algorithm is fast and probably correct. We can also modify it to be correct (exhaustive) at the expense of speed.

- We can use encryption to hide information from onlookers.
- We can use randomization to selectively share information.

Definitions

Experiment: perform an action once, e.g. toss a 6-sided die.

Sample Space: The set of possible outcomes, consisting of individual outcomes or sample points with known probability. e.g. for the dice experiment, $S = \{1, 2, 3, 4, 5, 6\}$ and each of $1, 2, 3, \ldots$ is a sample point. For a fair die, the probability of any sample point is $1/6$.

An Event E (subset of S): Is a subset of sample points, e.g. even die tosses $\{2, 4, 6\}$.

Random Variables: A random variable X is a function $X : S \to \mathbb{R}$ from a sample space to $\mathbb{R}$, the real numbers. A random variable assigns a real value to every possible outcome of an experiment. e.g.

$$X_1 = i, \quad \text{where } i \text{ is the number on the die.}$$

X_1 has domain and range $\{1, \ldots, 6\}$.

$$X_2 = \begin{cases} 1 & \text{if the die comes up even} \\ 0 & \text{otherwise} \end{cases}$$

X_2 has domain $\{1, \ldots, 6\}$ and range $\{0, 1\}$.

The second random variable is called an **indicator random variable**, because it is 0-1 valued. A 0-1 valued random variable naturally describes an event, which is the set of sample points where the variable is 1. In this case, X_2 describes the event that the die toss is even.

Random variables inherit a probability distribution from the sample space. The probability $\Pr[X = i]$ is the sum of the probabilities for all sample points where $X = i$. So for the examples above:

$$\Pr[X_1 = i] \text{ is } 1/6 \text{ for } i \text{ in } \{1, \ldots, 6\}$$

$$\Pr[X_2 = 0] = \Pr[X_2 = 1] = 1/2$$

It follows that if we take the sum $\sum \Pr[X = v]$ with v ranging over the range of X, we are summing the probabilities of all the sample points. So

$$\sum_{v \in \text{range}(X)} Pr[X = v] = 1 \text{ always}$$

We can also have a random variable with an infinite domain, e.g. $\Pr[X = i] = 1/2^i$ for $i = 1, 2, 3, \ldots$ and we still have:

$$\sum_{i=1}^{\infty} \Pr[X = i] = 1$$

Joint probability

We will often want to talk about the probability of two events happening at the same time. For example, the probability that a die toss is both even and a multiple of 3. The notation for this is $\Pr[X_1 = u, X_2 = v]$ and it is called a joint probability. It means the probability of the set of outcomes where both X_1 has value u and X_2 has value v. So you can think of the comma as an AND operator.

Arithmetic on Random Variables
Since random variables are real-valued functions, we can do arithmetic on them. The result is another random variable. For example, if we write $Z = X + Y$, then Z is a random variable. Its value at any point in the sample space is the sum of the values of X and Y at that sample point. Similarly, $W = X \times Y$ is a random variable whose value at a sample point is product of X and Y at that point.

A Caution About the Sample Space
Sometimes random variables are defined on different sample spaces. For instance, let X be the value on the top of a fair die toss. Let Y be the value on the top of a *different* toss. There are actually two different sample spaces. But we can think of them being part of a larger sample space that contains both experiments. That is, an experiment is a pair of throws of the die. Then X depends only on the first toss, and Y depends only on the second. If we do this, we are able to define $Z = X + Y$. The table below shows the probability distribution of Z. The first row is the value of Z, the next row is the probability, and the last row is the corresponding pairs of (X, Y) values.

Z	2	3	4	5	6	7	8	9	10	11	12
Pr[Z]	1/12	2/12	3/12	4/12	5/12	6/12	5/12	4/12	3/12	2/12	1/12
(X,Y)	(1,1)	(1,2) (2,1)	(1,3) (2,2) (3,1)	(1,4) (2,3) (3,2) (4,1)	(1,5) (2,4) (3,3) (4,2) (5,1)	(1,6) (2,5) (3,4) (4,3) (5,2) (6,1)	(2,6) (3,5) (4,4) (5,3) (6,2)	(3,6) (4,5) (5,4) (6,3)	(4,6) (5,5) (6,4)	(5,6) (6,5)	(6,6)

Independence
A very important concept for this course is independence of RV's. X_1 and X_2 are independent random variables if $\Pr[X_1 = u, X_2 = v] = \Pr[X_1 = u]\Pr[X_2 = v]$ for all u and v in the ranges of X_1 and X_2.

Example 1
For a single toss of a fair die, let $X_1 = 1$ if the number on the die is even, $X_1 = 0$ otherwise. Let X_2 be 1 if the same die toss gives a 4, and 0 otherwise. Then X_1 and X_2 are **not** independent. We need only disprove the identity in one place, e.g. $\Pr[X_1 = 1, X_2 = 1]$ is the probability that the die toss is even and a four, in other words a four. Thus $\Pr[X_1 = 1, X_2 = 1] = 1/6$. But $\Pr[X_1 = 1] = 1/2$ and $\Pr[X_2 = 1] = 1/6$ and the product of these two does not equal $\Pr[X_1 = 1, X_2 = 1]$.

Example 2
Now suppose we toss a fair die twice, and let $X_1 = 1$ if the number on the *first* die is even,

$X_1 = 0$ otherwise. Let X_2 be 1 if the *second* die toss gives a 4, and 0 otherwise. In this case X_1 and X_2 **are** independent. The outcomes where $X_1 = X_2 = 1$ are the pairs of tosses $(2, 4)$, $(4, 4)$ and $(6, 4)$. The total number of outcomes is 36, so the $\Pr[X_1 = 1, X_2 = 1] = 3/36 = 1/12$. This does match the product of $\Pr[X_1 = 1] = 1/2$ and $\Pr[X_2 = 1] = 1/6$. You can check yourself that probabilities for other values of X_1 and X_2 match also.

Examples 1 and 2 appear to have the same definitions for their R.V.'s. But the sample spaces are different. Be careful when using random variables. Make sure you understand both the definition of the variable *and* the sample space on which it is defined.

Conditional Probability and Independence
The conditional probability that $X_1 = u$ given $X_2 = v$ is written $\Pr[X_1 = u|X_2 = v]$ and is defined as:

$$\Pr[X_1 = u|X_2 = v] = \frac{\Pr[X_1 = u, X_2 = v]}{\Pr[X_2 = v]}$$

it means the probability that $X_1 = u$ within the smaller sample space where $X_2 = v$. Because we are in the smaller sample space, we divide by the probability of that space $\Pr[X_2 = v]$.

Conditional probability gives an alternative definition of independence: Random variables X_1 and X_2 are independent if and only if:

$$\Pr[X_1 = u|X_2 = v] = \Pr[X_1 = u] \quad \text{for all } u, v$$

In other words, X_1 and X_2 are independent if conditioning by X_2 has no effect on the probability of $X_1 = u$.

Expected Value
Associated with a random variable is its expected value $\mathrm{E}[X]$, defined by

$$\mathrm{E}[X] = \sum_{v \in \mathrm{Range}(X)} v\Pr[X = v]$$

Linearity
Another important idea for this course: Expected value is linear, i.e. it satisfies $\mathrm{E}[X_1 + X_2] = \mathrm{E}[X_1] + E[X_2]$ or more generally:

$$\mathrm{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathrm{E}[X_i]$$

Note: Linearity of expectation doesnt require independence. It is always true.

Proof:
We do the proof only for $n = 2$. The general case follows easily by induction on n. To compute the value for the random variable $Y = X_1 + X_2$, we would ordinarily compute its range. But in fact its equivalent to work separately over the ranges of X_1 and X_2. That is, what we actually want is

$$\mathrm{E}[Y] = \sum_{w \in \mathrm{Range}(Y)} w\Pr[Y = w]$$

But since the $\Pr[Y = w]$ is the sum of $\Pr[X_1 = u, X_2 = v]$ for all pairs u, v such that $u + v = w$, the above sum is equivalent to:

$$\begin{aligned}
\mathrm{E}[X_1 + X_2] &= \textstyle\sum_{u \in \mathrm{Range}(X_1)} \sum_{v \in \mathrm{Range}(X_2)} (u+v) \Pr[X_1 = u, X_2 = v] \\
&= \textstyle\sum_u \sum_v u \Pr[X_1 = u, X_2 = v] + \sum_u \sum_v v \Pr[X_1 = u, X_2 = v] \\
&= \textstyle\sum_u u \sum_v \Pr[X_1 = u, X_2 = v] + \sum_v v \sum_u \Pr[X_1 = u, X_2 = v] \\
&= \textstyle\sum_u u \Pr[X_1 = u] + \sum_v v \Pr[X_2 = v] \\
&= \mathrm{E}[X_1] + \mathrm{E}[X_2]
\end{aligned}$$

QED, and nowhere did we use the independence property

Products
The rule for products of RV's is what you might expect. However, it requires independence of the RV's.

Theorem:
If X_1 and X_2 are independent, then $\mathrm{E}[X_1 X_2] = \mathrm{E}[X_1]\mathrm{E}[X_2]$

Proof: We can start like we did for sums:

$$\mathrm{E}[X_1 X_2] = \sum_{u \in \mathrm{Range}(X_1)} \sum_{v \in \mathrm{Range}(X_2)} uv \Pr[X_1 = u, X_2 = v]$$

We can move u, but we get stuck here unless we use independence:

$$\mathrm{E}[X_1 X_2] = \sum_u u \sum_v v \Pr[X_1 = u, X_2 = v]$$

applying the independence rule will allow us to go further:

$$\mathrm{E}[X_1 X_2] = \sum_u u \sum_v v \Pr[X_1 = u] \Pr[X_2 = v]$$

Now we have a "constant" ($\Pr[X_1 = u]$) that can be moved outside the sum over v:

$$\mathrm{E}[X_1 X_2] = \sum_u u \Pr[X_1 = u] \sum_v v \Pr[X_2 = v]$$

which we recognize as:

$$\mathrm{E}[X_1 X_2] = \sum_u u \Pr[X_1 = u] \mathrm{E}[X_2] = \mathrm{E}[X_1]\mathrm{E}[X_2]$$

QED