Introduction

In probability and statistics we develop the tools for studying data that is uncertain or noisy and processes that have a random element. Probability and statistics is important to computer scientists of all types; some computer scientists use computers to manipulate data, or to make statistical inferences about data in, for example, machine learning, other computer scientists make products that need to be tested and statistics is needed to interpret the results of user testing.

Imaging a gambler wants to decide if a coin is fair; imagine they toss it five times and get a head each time. Can they decide that the coin is fair. It is clear that they can't say that it is impossible that a fair coin would turn up heads five times; it is just pretty unlikely. They could say that there is only a one in

$$2^5 = 32 \tag{1}$$

chance that a fair coin would produce that result. Since $(1 - 1/32) * 100 \approx 97$ it would seem that gambler could be 97% certain the coin was unfair. This isn't the end of the story though: five harps would have been equally surprising; so perhaps the thing to say is that there is only a one in

$$2^4 = 16 \tag{2}$$

the coin would produce such an unlikely result. Does this means the gambler can only be 94% certain the coin is unfair? The idea of this unit it to learn how to calculate probabilities, the probability of five heads is an easy example of this, and how to make inferences, such as deciding the chance the coin is unfair.

It is very easy to confuse yourself with probability; here is a confusing problem. An evil genius has decided to tattoo their names on some peoples foreheads and in the usual over-elaborate evil genius manner they do it according to a diabolic game. A person is selected and forced into the tattoo parlour, the evil genius rolls two dice[1] and if they show two sixes, they tattoo the person and the game ends, otherwise the lucky person is released and the evil genius selects nine new people. Again, the evil genius rolls the dice, again, if it gives two sixes then they tattoo everyone and the game ends, if not, they let them leave, select 90 people and roll the dice. The process repeats, multiplying the number each time by ten, until the dice comes up with two sixes, the people in the parlour are tattooed and the game ends. If you are forced into the perilous parlour, what is the chance you'll end up with a tattoo on your forehead. Is it the relatively modest one in 36, or, given that 90% of people who enter the room end up tattooed, is it nine in ten?

This is a problem for you to think about yourselves, the point is that in statistics and probability it isn't enough to be good at calculating probabilities, though this is useful, it is important to know what you are trying to calculate. To avoid getting confused we need some notation and some mathematical machinery and that's what we are going to do now.

Probability

To start with we need to terminology and notation and this begins with a **sample space** X, this is the set of points; the idea is that these points are the possible outcomes of an experiment,

[1] Traditionally die is the singular of dice: "she rolled two dice, both die came up one, the roll known as 'snake eyes"', but these days often people use dice in the singular and dices in the plural; either is, of course, fine.

or, it is said, the outcome of a **trial**. Initially it is useful to think of discrete sample spaces with individual points, for example, for the coin where the experiment is tossing it once, the sample space has two points:

$$S = \{H, T\} \tag{3}$$

where the two points H and T correspond to heads and harps, the two possible outcomes of flipping the coin. If the experiment is tossing the coin twice then the sample space has four points:

$$S = \{HH, HT, TH, TT\} \tag{4}$$

unless the outcome of a trial only records the number of heads and harps, not the order, in which case the sample space has three points:

$$S = \{HH, HT, TT\} \tag{5}$$

which HT stand for heads then harps as well as harps then heads. Later we will look at continuous sample spaces where the experiment takes a value in a continuum, so we could have, for example

$$\mathcal{L} = [0, \infty) \tag{6}$$

as the sample space for the experiment of measuring the length of a song on the radio.

An **event** in a discrete sample space X is a subset $E \subset X$. If the sample space was $\{1, 2, 3, 4, 5, 6\}$ corresponding to the face values of a dice, then the set of even sides $\{2, 4, 6\}$ is an example event. This is a useful notation because we might want to calculate the probability for an event as well as for the individual points; we might wonder, what is the chance the dice will show an even value, that is, what is the chance the outcome is an element of the event.

Now, we want to define a **probability** on a sample space X. Obviously a probability is the chance of something happening, but before doing that part, we need to look at some of the properties that a probability needs to have. Formally, a probability is a map P from events to real numbers such that

1. $P(A) \geq 0$ for all events.
2. $P(X) = 1$
3. If $A \cap B = \emptyset$ for two events A and B then

$$P(A \cup B) = P(A) + P(B) \tag{7}$$

where $A \cap B$ and $A \cup B$ are the intersection and union of A and B and $\emptyset$ is the empty set. This lists the properties we expect a probability to have; the key thing being if two events are disjoint then the probability of the joint event is the sum of the probabilities of the individual events. For example the probability of getting one or two on a dice is a third, the probability of getting a three is a sixth; putting these together, the probablity of getting a one, two or three is a half.

On a discrete sample space a probability and a **probability mass function** are similar but subtely different objects. A probability, as defined above, is a map from events to real numbers, a probability mass function is a map from points in the sample space to real numbers: the probability mass function $p(x)$ for x in X has the properties:

1. $p(x) \geq 0$ for all $x \in X$

2. $\sum_{x\in X} p(x) = 1$

and the two can be related:

$$P(A) = \sum_{x\in A} p(x) \tag{8}$$

This might seem like a silly distinction and, in fact, people are often very casual in distinguishing the two. The distinction is, however, more important when dealing with more challenging sample spaces.

Counting things

The is a class of problems where calculating probabilities can be reduced to counting things; this basically works using the rule above that says $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$. If, for example, all the points in a sample space have the same probability, say q, then this rule means that the probability of an event is is

$$P(A) = (\text{number of points in } A) \times q \tag{9}$$

To see this notice that if $A = \{a_1, a_2, \ldots, a_k\}$ are the points in A we can write A as the union of little sets each containing just one point:

$$A = \{a_1\} \cup \{a_2\} \cup \ldots \cup \{a_k\} \tag{10}$$

and so

$$P(A) = P(\{a_1\}) + P(\{a_2\}) + \ldots + P(\{a_k\}) = q + q + \ldots + q \tag{11}$$

From the rule $P(X) = 1$ it also follows that

$$q = 1/(\text{total number of points in the sample space}) \tag{12}$$

Incidentally, a few different notations are used for the number of points in a set, I will use

$$\#(A) := \text{number of points in } A \tag{13}$$

where ':=' means 'is defined as' or 'is used to mean'. Another common notation is

$$|A| := \text{number of points in } A \tag{14}$$

One standard example for calculating probabilities is the problem of working out the probability of getting different card hands in poker; for definiteness lets consider five card stud poker where the player is dealt five cards. A pair is the hand where two cards are the same and all the rest of different. What is the probability of getting a pair. The total number of different hands in poker is 52 choose five:

$$\left(\begin{array}{c} 52 \\ 5 \end{array}\right) = \frac{52 \times 51 \times 50 \times 49 \times 48}{1 \times 2 \times 3 \times 4 \times 5} = 2598960 \tag{15}$$

where the binomial coefficient

$$\left(\begin{array}{c} n \\ r \end{array}\right) = \frac{n!}{r!(n-r)!} \tag{16}$$

counts the number of subsets of size r in a set of n objects and

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \tag{17}$$

How many pairs are there; well there are 13 possible values for the pair in two suits out of four, giving six possible pairs of suits, there are 12 choose three ways of picking out the other three values, there is no constraint on the suit of the non-pair cards, giving

$$\#(\text{set of pairs}) = 13 \times 6 \times \frac{12 \times 11 \times 10}{6} \times 4^3 = 1098240 \tag{18}$$

Thus

$$P(\text{pair}) = \frac{1098240}{2598960} \approx 0.42 \tag{19}$$

A classic problem is the birthday party problem; a common intuition is that it is very rare to find people in a group with the same birthday; it isn't the case: in a group of 23 people there is a one in two chance that two with share a birthday. Roughly speaking the point is that there are lots of pairs of people. Now, lets ignore leap years and the impact the long long nights of January have on the birth rate in September. So, say there are 20 people in a group, a point in the sample space is a list of 20 birthdays. The number of possible lists, and hence the size of the sample space is 365^{20}. However, the number of lists where all the birthdays are different is

$$\#(\text{lists with different birthdays}) = 365 \times 364 \times \ldots \times 347 \times 346 \tag{20}$$

This means the probability no pair has a shared birthday is

$$P(\text{no shared birthday}) = \frac{365}{365}\frac{364}{365}\cdots\frac{347}{365}\frac{346}{365} \approx 0.5886 \tag{21}$$

and hence

$$P(\text{at least one shared birthday}) = 1 - P(\text{no shared birthday}) \approx 0.414 \tag{22}$$

Put another way, the probability of a shared birthday in a group of n people is

$$P(\text{at least one shared birthday}, n) = 1 - \prod_{i=0}^{n-1} \frac{365-i}{365} \tag{23}$$

where the big pi symbol is the produce symbol, like the big sigma used for summing but for products. The probability graph can be seen in Fig. 1.

When we were defining the binomial coefficient, we used the factorial $n!$. The factorial also has a counting function, it counts the number of ways of ordering a set of size n. A similar function is the permutation

$$P_r^n = \frac{n!}{(n-r)!} \tag{24}$$

This counts the number of ways of selecting r objects from n where ordering is taken in to account. Say you had 26 lettered tiled, all different and you made a random word by choosing five lettered tiles one by one at random and lining them up as you picked them out. The number of possible words would be

$$P_5^{26} = 26 \times 25 \times 25 \times 24 \times 23 = 7893600 \tag{25}$$

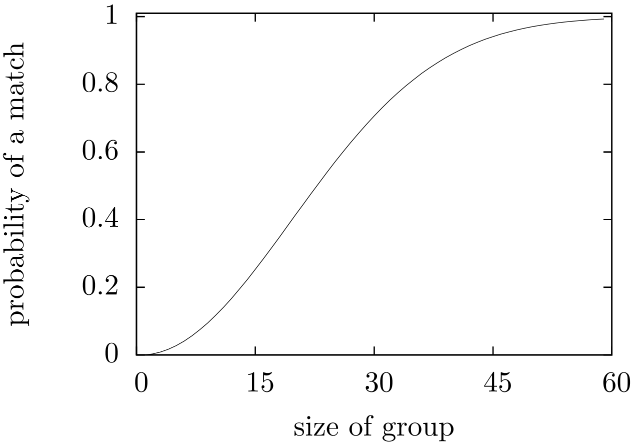

Figure 1: The probability of at least two people having the same birthday in a group of people, plotted against the size of the group. The code for making this graph is `birthday.jl`

Another useful combinatorial function is the partition function

$$\left(\begin{array}{c} n \\ n_1, n_2, \ldots, n_r \end{array} \right) = \frac{n!}{n_1! n_2! \ldots n_r!} \tag{26}$$

where $n_1 + n_2 + \ldots + n_r = n$. The partition function counts the number of ways a set of n objects can be split up into r subgroups of sizes n_1, n_2 and so on to n_r.

Imagine there are 20 people who are given four different jobs; the best job, easy work, lots of pay, requires six people, the second, inferior job requires four and the other two, even worse jobs, require five each. The people are supposed to be distributed among the jobs randomly, but the supervisor's four siblings are all given the best job. How likely is that to be just good fortune for the sibs? Well the total number of ways of distributing the jobs is

$$\left(\begin{array}{c} 20 \\ 6, 4, 5, 5 \end{array} \right) = \frac{20!}{6!4!5!5!} \tag{27}$$

Now, if the supervisor's four siblings are in the first group, there are 16 labourers left to distribute, with two places left for the first job. This means the number of ways of assigning the jobs so that the supervisor's siblings all get the first job is

$$\left(\begin{array}{c} 16 \\ 2, 4, 5, 5 \end{array} \right) = \frac{16!}{2!4!5!5!} \tag{28}$$

Hence, if giving out the jobs was truly done randomly, the chance the supervisor's siblings all got the good job is

$$p = \frac{16!}{2!4!5!5!} \bigg/ \frac{20!}{6!4!5!5!} = \frac{16!}{20!} \frac{6!}{2!} = \frac{6 \times 5 \times 4 \times 3}{20 \times 19 \times 18 \times 17} \approx 0.0031, \tag{29}$$

so not very likely.

Summary

- A **sample space** is a set of point, they are the possible **outcomes** of a **trial**.
- An **event** is a subset of a sample space.
- A **probability** is a map from events to real numbers such that
  1. $P(A) \geq 0$ for all events.
  2. $P(X) = 1$
  3. If $A \cap B = \emptyset$ for two events A and B then
  $$P(A \cup B) = P(A) + P(B) \tag{30}$$
- A **probability mass function** is a map from points in the sample space to real numbers such that
  1. $p(x) \geq 0$ for all $x \in X$
  2. $\sum_{x \in X} p(x) = 1$
- $P(A) = \sum_{x \in A} p(x)$
- If all the points in a sample space have the same probability then
  $$P(A) = \frac{\text{number of points in } A}{\text{number of points in } X} = \frac{\#A}{\#X} \tag{31}$$
  where $\#(A)$ means the number of points in A.
- The **binomial coefficient**
  $$\left(\begin{array}{c} n \\ r \end{array}\right) = \frac{n!}{r!(n-r)!} \tag{32}$$
  counts the number of subsets of size r in a set of n objects and
  $$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \tag{33}$$
- The **partition function**
  $$\left(\begin{array}{c} n \\ n_1, n_2, \ldots, n_r \end{array}\right) = \frac{n!}{n_1! n_2! \ldots n_r!} \tag{34}$$
  where $n_1 + n_2 + \ldots + n_r = n$ counts the number of ways a set of n objects can be split up into r subgroups of sizes n_1, n_2 and so on to n_r.

Example question

A rich man has six gold coins he splits into three piles randomly, so that each pile has at least one coin. The first pile he gives to his daughter, the second to his son and the third he throws in the lake for good luck. What chance does the daughter have of getting three or more coins.
solution The sample space is made up of all possible divisions in piles

$$\begin{aligned} X &= \{(4,1,1),(3,2,1),(3,1,2),(2,3,1),(2,1,3), \\ &\quad (2,2,2),(1,4,1),(1,1,4),(1,3,2),(1,2,3)\} \end{aligned} \tag{35}$$

The event that the daughter gets at least three coints is

$$E = \{(4,1,1),(3,2,1),(3,1,2)\} \tag{36}$$

If all divisions are equally likely

$$P(E) = \frac{\#E}{\#X} = \frac{3}{10} \tag{37}$$