# Probability & Statistics

## Probability

Probability deals with events which are not certain. The probability of an event represents a measure of the strength of our belief that an event will occur. It can intuitively be thought of by imagining the experiment being repeated infinitely many times in such a way that we have the same belief of the event happening each time. For example, if a coin is fair, then when we toss it a large number of times, approximately half of the times will be heads, and half will be tails. We therefore say that the probability of heads is $\frac{1}{2}$. This situation where we have equally likely outcomes is common. We then use methods from enumerative combinatorics to determine how many outcomes are included in the event we are interested in. For example, if we roll two fair dice, there are 36 equally likely outcomes; if we want to determine the probability that the sum of the two dice is 6, we need to count the number of outcomes where the sum is 6. In ths case, there are 5 such outcomes, so the probability is $\frac{5}{36}$.

Sometimes we can't divide into equally likely outcomes. In this case, we can assign a probability to each outcome, then calculate the probability of an event as the sum of the probabilities of the outcomes covered by that event. For example, if a die is loaded so that '6' has a probability $\frac{2}{7}$ and all other numbers have a probability $\frac{1}{7}$, then the probability of rolling an even number is $\frac{2}{7} + \frac{1}{7} + \frac{1}{7} = \frac{4}{7}$.

When we have two or more different experiments, the probabilities of combined results depend on how one experiment influences another. However, in the case where one experiment does not influence the other (in this case, we say they are *independent*), we obtain the probability of the combined outcome as the product of the probabilities of the individual outcomes. For example, if we roll a fair die and toss a fair coin, then the probability of rolling a '5' and getting tails is $\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$.

One important rule of probability which is often misunderstood is Bayes' theorem. This considers the problem of how to revise probabilities of events based on new evidence. For example, suppose a person receives a test for a disease which affects one person in 1000, and that the test falsely indicates that a person has the disease when they do not 1% of the time (and always gives a positive result if they do have the disease). If the test result is positive, the chance that the person actually has the disease is still only about 9% (more accurately, it is $\frac{100}{1099}$). The reason is that the initial probability of having the disease is very low, so even after the positive test result, it is more likely that the person does not have the disease. Another way to look at this is that for every person with the disease who is tested, there are 999 people without the disease tested, and on average, about 10 of them will give false positive tests. This kind of situation, where we need to revise probabilities based on evidence is very common — it also occurs in legal trials, for example forensic evidence such as DNA testing can give a false positive result, so if it is the only evidence against a suspect, then it should not usually be enough to convict them. Failure

to understand this has led to many wrong convictions.

Related to Bayes' rule is the Monty Hall problem. On a certain game-show, the contestant was given a choice of three doors, one of which concealed a prize, and the other two concealed worthless prizes. After the contestant chose a door, before openning it, the host would open one of the other doors, which concealed a worthless prize, and give the contestant a chance to switch to the other unopenned door. [Actually the host only did this some of the time, often when the contestant had chosen the correct door, but for the mathematical problem, we assume that the host would make this offer every time.] Intuitively, many people think that the prize is equally likely to be behind either of the two unopenned doors, but this is not true, because the probability that the original choice was correct is $\frac{1}{3}$, and regardless of the original choice, the host can find another door to open that does not reveal a prize, so the host's doing this does not change the original probability. Therefore, the probability that the prize is behind the original door is $\frac{1}{3}$, and the probability that the prize is behind the other door is $\frac{2}{3}$. It is worth noting that this depends upon the host's intentions. If the host chose to open one of the other two doors at random, regardless of what was behind, then it would not be advantageous to change, because the fact that the host did not accidentally reveal the prize makes it more likely that the prize is behind the door originally chosen.

## Statistics

Statistics is the study of how to interpret data which is influenced by some random effect. For example, if a random sample of people are surveyed, how much can we infer from their answers? Our interpretation of the data is usually based upon a belief that the underlying randomness is controlled by a certain probability distribution. For example, in an opinion poll, we might hope that every person should have the same chance of being surveyed, and that whether one individual is surveyed or not should have no influence on another individual's probability of being included in the survey (the events are called *independent*).

Given the wide variety of situations where we need to make informed decisions based on data, statistics is an area with many applications across all areas of life.

There are several problems we might consider in statistics. Firstly, we might want a method for estimating unknown parameters from our data. For example, in an opinion poll, the unknown parameter we would be interested in is the proportion of the total population that supports a particular political party. Other examples might include the total population of a particular species of animal or plant. A popular method for doing this is to estimate the parameter which would lead to the largest probability of observing the data that we actually observed.

An estimate of the unknown parameters is of limited use on its own, because we do not know how reliable it is. An alternative is to give not just a single estimate, but a range of plausible values, such that we believe that the true value should be within this range a large proportion of the time (95% is a commonly

used value). This is called a confidence interval. For example, the opinion poll might report that there is a 95% chance that the support for a given party is between 45% and 49%. This can give us an idea whether the experiment gave sufficient data to reach reasonable conclusions, or whether further experiments will be necessary.

One important use of data is to determine whether a particular conclusion is valid. For example, to answer questions such as "Is this new drug more effective than existing drugs?" Answering such questions is called *hypothesis testing*. To answer such questions, we need to establish how plausible it is that the observed data was just a coincidence. To achieve this, we form two hypotheses: a *null hypothesis* which asserts that there is nothing interesting happenning, for example "the new drug has the same effectiveness as the old one."; and an *alternative hypothesis* which we are testing, for example, "The new drug is more effective than the old one." We then use the data to obtain a *test statistic*, and choose a cutoff value of the test statistic so that if the null hypothesis is true, the test statistic will be less than that cutoff value (or more than that cutoff value if the statistic is likely to be smaller under the alternative hypothesis) a fixed percentage of the time (95% is common, so is 99%). The proportion of the time that the test statistic would be more than the cutoff value under the null hypothesis is called the *significance level* of the test. The idea is that the significance level gives the proportion of times that the hypothesis test will wrongly support the alternative hypothesis. The choice of significance level will depend on how expensive such a mistake could be.

An important point to remember in hypothesis testing is that often any individual result is very unlikely, so we need to group results together to reach a good conclusion. For example, the probability of any one individual winning the lottery is very small, but the probability that someone will win the lottery is fairly large, so if we just looked at the probability of the result that we observed (a particular person winning the lottery), we would appear to have witnessed an unlikely event, but in fact the event is not so surprising, when we take it together with all similar results. This of course depends upon our initial intentions. Another example: suppose we roll a die three times and get three '2's; the probability of this is $\frac{1}{216}$, but there is the same probability for any three outcomes, so to say that the observed outcome was significant, we have to decide what other outcomes we would view as similar. If our initial belief had been that '2' was more likely than any other number, then we would not consider any other result as significant as this one; but if our belief was just that the die wasn't fair, then we would consider any result with all three numbers the same as equally significant, so the relevant probability would be $\frac{6}{216} = \frac{1}{36}$; and if our belief was that '6' was more likely, then this result would not be at all significant, since it does not indicate that '6' is more likely.