Die Welt ist alles, was der Fall ist.
The world is all that is the case.

1 Some basics

These are the opening words of Tractatus Logico-Philosophicus, the monumental work by Ludwig Wittgenstein whom many regard as the greatest philosopher of the twentieth century (and perhaps one of the greatest enigmas). Our universes, however, will be smaller, and ones in which there will be outcomes of experiments like the draw of a card from a deck or toss of a die. Sometimes nature performs the experiment and the outcome is whether there will be rain on July 14 or life on Mars. One way or another we assign likelihoods or probabilities to all of these, and sometimes the probabilities depend on who we are, but these probabilities all obey certain rules.

The probability of an event is a number between 0 and 1, written \( p(\text{event}) \). Roughly speaking, \( p(\text{event}) = 0 \) means it won’t happen (but not that it can’t happen) and \( p(\text{event}) = 1 \) means that it will happen (but not that it is certain to happen). That may seem contradictory, but we’ll make sense of it later. The illustrations have been of several different types of probability. As to rain on July 14 at City Hall, we can consult weather bureau records as far back as they go, and if they show that on 20% of all July 14s on record there was measurable rain then we may say that \( p(\text{rain on July 14}) = .20 \); this is empirical probability, probability derived from observation. If we were asked, say, on April 2, What are the chances of rain on next July 14? our answer might be “one in five” (but if we were asked on July 13 and had read the forecast it might be quite different). If, by extraordinary chance, it had never in the past rained on July 14 then our empirical \( p \) would be zero, but that wouldn’t imply that it could never rain on July 14. As for the toss of the die, if we assume that it was perfectly formed (not “loaded” or biased in any way) and therefore that all faces have an equal chance of showing, then the probability that any one of ‘1’ to ‘6’ will appear will be 1/6, the probability that the face showing will be even will be 1/2 (since 1/2 of the possibilities are even and each has equal likelihood). This is theoretical probability, a probability derived from certain assumptions and the nature of the experiment. We can say the same thing for the draw of a
card from a theoretically well-shuffled deck: the probability of drawing the ace of spades is 1/52 (although I wouldn’t bet on it with some of my friends), and of drawing a red card is 1/2. Finally, as to life on Mars, we each have our own subjective probability; some might say, one chance in a million (p = .000001), others emphatically, No way! (p = 0). The value depends on who you are (and can be drawn out of you by asking certain questions).

Whatever foundations you put under your concept of probability, once we decide what the universe is that we are exploring and the kind of experiment that we will perform, there are certain basic rules. If A is a possible event (of the kind with which we are dealing, like rain or toss of a die) and ∼ A (read “not A”) the set of all possibilities in our universe other than A, then

\[ 0 \leq p(A) \leq 1 \quad \text{and} \quad p(\sim A) = 1 - p(A). \]

Frequently, when the value of p is understood, one writes q for 1 − p, so one could write \( p(\sim A) = q \). Suppose that A and B are possible events. Then we write \( A \cup B \) (read “A or B”) for the “union” of the two events; this is the non-exclusive “or” – it includes the possibility that both may happen. For example, if A is “the card drawn is the ace of spades” and B is “rain on July 14”, then
\[ A \cup B \] in “the card drawn is the ace of spades or rain on July 14 or both”.
Similarly, \( A \cap B \) (“A and B”) denotes the “intersection” of the two events, “the card drawn is the ace of spades and rain on July 14, both”. The next rule is that If A and B are mutually exclusive events then \( p(A \cup B) = p(A) + p(B) \). To write this in short mathematical notation it is common to introduce the symbol “∅” for the “empty set”, the event with no possibilities, for which we must have, of course \( p(\emptyset) = 0 \) since the empty set includes none of the possibilities which may happen. We may also use \( \mathcal{U} \) for the “universe”, the set of all possibilities, for which we have \( p(\mathcal{U}) = 1 \); something must happen. For the throw of a die we may think of \( \mathcal{U} \) as consisting of the numbers \{1, 2, 3, 4, 5, 6\} and for the weather on July 14 as consisting of \{rain, dry\} or even \{0, 1\} where 0 codes for rain and 1 for dry (or the other way around, if you prefer). The notation \( A \setminus B \) denotes all things in A which are not in B. (This is the same as \( A \cap \sim B \) since we only take away from A things which are simultaneously in A and in B) Then \( \mathcal{U} \setminus A = \sim A \), the complement of A. The second rule can then be written

\[ A \cap B = \emptyset \quad \text{implies} \quad p(A \cup B) = p(A) + p(B) \quad (1) \]

If we have three mutually exclusive events, \( A_1, A_2, A_3 \) and in the preceding equation put \( A = A_1, B = A_2 \cup A_3 \) and then use the equation a second time to compute \( p(A_1 \cup A_2) \) we will get

\[ A_1 \cap A_2 = A_1 \cap A_3 = A_2 \cap A_3 = \emptyset \quad \text{implies} \quad p(A_1 \cup A_2 \cup A_3) = p(A_1) + p(A_2) + p(A_3) \]

We can go on this way for any finite number of mutually exclusive events but we will assume (since we can’t prove) the following statement for an infinite sequence of events:

if \( A_1, A_2, A_3, \ldots \) are mutually exclusive, then
\[ p(A_1 \cup A_2 \cup A_3 \cup \ldots) = p(A_1) + p(A_2) + p(A_3) + \cdots \quad (2) \]
When events $A$ and $B$ are mutually exclusive they are often called disjoint and then $A \cup B$ is called a disjoint union. From (1) it is a simple exercise to show that
\[
p(A \cup B) = p(A) + p(B) - p(A \cap B) \quad (3)
\]
and that if $A_1, A_2, A_3$ are three events then
\[
p(A_1 \cup A_2 \cup A_3) = p(A_1) + p(A_2) + p(A_3) - p(A_1 \cap A_2) - p(A_1 \cap A_3) - p(A_2 \cap A_3) + p(A_1 \cap A_2 \cap A_3).
\]

You should be able to guess the general formula.

## 2 Conditional probability

The knowledge that an event $B$ has occurred may influence our computation of the probability of an event $A$. For example, rain frequently persists for more than one day, so if we know that it rained on July 13 then our estimate of the probability of rain on July 14 should increase. The notation $p(A|B)$ denotes “the probability of $A$ given $B$”, i.e., the probability of $A$ knowing that $B$ has occurred. Since
\[
p(A \cap B) = p(B)p(A|B)
\]
(“to get into the intersection one must first get into $B$, and given that one is in $B$ then get into $A$”) so we have
\[
p(A|B) = \frac{p(A \cap B)}{p(B)} \quad \text{if} \quad p(B) \neq 0.
\]

Notice that $A \cap B$ is the same thing as $B \cap A$, so we can interchange the roles of $B$ and $A$ to get the useful formula
\[
p(A|B)p(B) = p(A \cap B) = p(B|A)p(A) \quad (5)
\]

There are important situations where $p(B|A)$ is known and we want to compute $p(A|B)$. Here is an abstract example. Suppose that we have two urns, labeled ‘Urn1’ and ‘Urn2’, each containing white and green balls, well stirred, but in Urn1 $1/2$ of the balls are white and the rest green, while in Urn2 only $1/4$ of the balls are white and the rest green. If someone picked a ball at random from Urn1, the probability that it would be white would be $1/2$, but if they picked from Urn2 the probability of getting a white ball would be only $1/4$. People are blindfolded, walked to the table holding the urns and told to pick a ball from one of them. Suppose that because of the way the urns are placed people normally take a ball from Urn1 $4/5$ of the time and from Urn2 only $1/5$ of the time. Now one of the blindfolded subjects in this experiment picks a ball from one of the urns, the blindfold is removed, and it turns out that the ball is green. What is the probability that the ball came from Urn1? Does the fact that it was green and that Urn2 has a greater proportion of green balls change the initial probability?
Here is one way to answer this question. Suppose that Urn1 contains 400 balls, half white and half green, and that Urn2 contains 100 balls, 25 white and 75 green, reflecting the given probabilities. Mark the balls so that when we toss them all into one large urn we can identify their source. This urn now contains 500 balls, \( \frac{4}{5} \) from Urn1 and \( \frac{1}{5} \) from Urn2, so if we stir them well and pick one at random, then just as before, the probability that it came from Urn1 is \( \frac{4}{5} \) and from Urn2 is \( \frac{1}{5} \). (The issue of stirring well is quite serious; find and read the discussion of the 1970 draft lottery, which for many young men was a matter of life and death!) So now we no longer need two urns; all the balls are in one, one ball has been drawn, and it is green. What is the probability that it came from Urn1? In the one big urn there are now 275 green balls, 200 from Urn1, 75 from Urn2, and we know that a green ball has been chosen. The white balls now are of no consequence. The probability of drawing one of Urn1’s green balls is now \( \frac{200}{275} = \frac{8}{11} \) and of Urn2’s is \( \frac{3}{11} \). So knowing that the ball drawn is green has indeed changed the probability that it came from Urn1; it was .8 originally and now it is only \( \frac{8}{11} \), which is a little less than .73. We can summarize this experiment as follows. Let \( p(A) \) denote the original or a priori or simply prior probability of drawing a ball from Urn1, and \( p(A|\text{green}) \) denote the probability of having drawn from Urn1, knowing that the ball drawn was green. We call this the a posteriori or posterior probability. Then the prior probability was \( p(A) = .8 \); the posterior probability was \( p(A|\text{green}) = .73 \).

While our trick of putting 400 balls in one urn and 100 in the other solved the problem, and in principle every such problem can be solved in a similar way, there is a general formula called Bayes’ Theorem, (which the Reverend Thomas Bayes never actually published in his lifetime). It is one of the oldest theorems in probability theory and one which has led to much controversy in its use (and misuse).

Let \( A \) denote the event of picking a ball from urn Urn1. Since there are only two urns, we can denote the event of picking from Urn2 by \( \sim A \). Then we have the prior probabilities \( p(A) = .8, p(\sim A) = .2 \). Let \( B \) denote the event of picking a green ball. We are told in advance that \( p(B|A) = 1/2 \) and that \( p(B|\sim A) = 1/4 \), and we want to determine \( p(A|B) \), the probability that the ball came from Urn1, given that it was green. From (4) one has

\[
p(A|B) = \frac{p(B|A)p(A)}{p(B)}
\]

The numerator is known. To compute the denominator, observe that there are two mutually exclusive possibilities for drawing a green ball, either it came from Urn1, \( A \), or from Urn2, \( \sim A \), so \( p(B) \) must be the sum of the probabilities that it came from each, namely \( p(B) = p(B \cap A) + p(B \cap \sim A) \). But \( p(B \cap A) = p(B|A)p(A) \) from (4) and similarly \( p(B \cap \sim A) = p(B|\sim A)p(\sim A) \). Therefore we can write \( p(B) = p(B|A)p(A) + p(B|\sim A)p(\sim A) \), and on the right side all the numbers are known. We therefore finally have the simple form of Bayes’
theorem:
\[ p(A|B) = \frac{p(B|A)p(A)}{(p(B|A)p(A) + p(B|\sim A)p(\sim A))} \] (6)

If there were more than two urns, say a total of \( n \), and the event of picking from urn \( i \) is denoted \( A_i \) then the complete form of Bayes’ theorem is

**Theorem 1 (Bayes)**
\[ p(A_i|B) = \frac{p(B|A_i)p(A)}{(p(B|A_1)p(A_1) + p(B|A_2)p(A_2) + \cdots + p(B|A_n)p(A_n))} \] (7)

If we insert the probabilities we started with into the simple form of Bayes’ theorem we get
\[
p(\text{Urn1|green}) = p(A|B) = \frac{\frac{1}{2} \times \frac{4}{5}}{\frac{1}{2} \times \frac{4}{5} + \frac{3}{4} \times \frac{1}{5}} = \frac{\frac{4}{A + 0.15}}{\frac{4}{A + 0.15}} = 0.727,\]

exactly what we got before.

The problem with Bayes’ theorem is that we can rely on it only when the prior probabilities are known but they are often uncertain. Nevertheless, it has been argued that a prosecutor should be allowed to present it to a jury; see the debate between Finkelstein and Fairley arguing for and Tribe against that raged through several issues of the Harvard Law Review. Human intuitive assessment of probabilities is notoriously bad and very carefully handled Bayes’ theorem might make a significant contribution, but should a prosecutor be allowed to say something like the following in his summation?

Ladies and Gentlemen of the Jury:
You have heard evidence placing the defendant at the scene of the burglary. There is no DNA and there are no fingerprints. Evidently the thief wore gloves. But you have seen that there were some strands of hair recovered from the frame of the window against which he must have brushed when he entered the house. These were rather unusual – kinky and orange – only one person in one hundred has natural hair like that. And these strands match the defendant’s hair. Now other evidence you have heard certainly implicates the defendant but perhaps it doesn’t yet convince you beyond reasonable doubt. Maybe you think that there is, perhaps, still one chance in ten that this defendant isn’t the person who broke into the house, sent the elderly couple living there to the hospital, and robbed them not only of their few possessions but also of their peace of mind. But let me show you what scientists can conclude from these strands of hair.

[Prentes Bayes’ theorem]
So you see, there is barely one chance in a thousand that with all the other evidence and that with these tell-tale strands of hair someone other than the defendant sitting there is guilty. Ladies and gentlemen, you must convict.
3 Independence

When the probability of an event $A$ does not depend on whether or not an event $B$ has occurred, that is, if $p(A|B) = p(A)$, then we say that $A$ and $B$ are independent. Since $p(A|B) = p(A \cap B)/p(B)$ this is the same as saying that

$$A \quad \text{and} \quad B \quad \text{are independent if} \quad p(A \cap B) = p(A)p(B).$$

That is, events are independent if the probability of their joint occurrence is the product of their individual probabilities. Warning: For three events $A$, $B$ and $C$ to be independent it is not sufficient that all pairs be independent; we need in addition that $p(A \cap B \cap C) = p(A)p(B)p(C)$. And for four events to be independent, we need, in addition to having all subsets of three be independent, that the probability of the joint occurrence of all four is the product of their individual probabilities. It may therefore become very difficult, empirically, to determine that the events in a large set are independent because the product of their probabilities may be small and you would need an overwhelming number of observations to verify that it is the same as the probability of the joint occurrence of the events. Here is a famous case in which this appeared: People v. Collins, 438 P. 2d 33 (68 Cal. 2d 319 1968). (You can find it through Wikipedia, where the entry will also give you the references to the first two papers in the debate between Flinkelstein plus Fairley and Tribe mentioned above. Find and read the decision and read these papers.) At the time of the Collins case Lawrence Tribe, later a law Professor at Harvard who had studied math as an undergraduate, was clerking for the justice of the California Supreme Court who wrote the opinion. The statistical appendix to the opinion (unique, as far as I know) was written (but of course not) signed by Tribe. The opinion contains three reasons for overturning Malcolm Collins’ original conviction of robbery. Two are valid and compelling, but there is a logical error in the third and most sophisticated, which involves the so-called Poisson distribution. I’ll discuss it when we get to that part of the course. The history of this distribution, now absolutely essential in statistics, is itself fascinating. Poisson, a brilliant French mathematician of the 19th century who also produced an equation fundamental in physics, at one point became interested in what would happen to the rate of reversals when the rules of the highest court in France changed. He developed the distribution named after him and predicted the result.

4 Odds; Bayes’ theorem restated

Frequently instead of the probability of an event $A$ being stated directly one is given the ‘odds’. If someone says, I’ll give you odds of 3 to 2 that Slowpoke will win the race, he is saying, I’ll put 3 on the table, you put down 2, and if slowpoke wins then I collect all 5, if not you get the 5. If honest (a major assumption in the circumstances) he is saying that his (subjective) probability that Slowpoke will win is 3/5 and he is willing to bet on it. (More likely, he thinks that the
probability is higher; he will ‘lay off’ the bet on someone willing to bet the other way and as the ‘house’, in this way hopes to make a steady profit from the individual betters. ) We can recover the probability from the odds: If we express the odds on an event as a ratio, \( r \), then its probability is \( p = \frac{1}{1 + r} \).

In the Slowpoke example \( r = \frac{3}{2}, 1 + r = \frac{5}{2}, \) and \( p = \frac{(3/2)/(5/2)}{3/5} = \frac{3}{5} \). To compute odds from probability, set \( r = \frac{p}{1 - p} \), as long as \( p \neq 1 \) (and if \( p = 1 \) you don’t want to be betting).

Important medical statistics are often presented in terms of an ‘odds ratio’. Suppose we are comparing the survival rates of long term heavy smokers (2 packs per day or more for 20 years or more) with those who never smoked and find that 70% of the non-smoker survived to age 80 but only 20% of the smokers. (These numbers ae fictitious. Go online to see if you can find the true statistics.) The odds of a non-smoker surviving to age 80 are then \( 7:3 \) or \( 7/3 \), while those for a non smoker are only \( 2:8 \) or \( 1/4 \). The odds ratio is \( (7/3)/(1/4) = 28/3 \); with these statistics the odds of surviving to age 80 are better than nine times as good for non-smokers than they are for smokers. If we put it the other way, the odds on dying before 80 are 3:7 for non-smokers and 4:1 for smokers, giving an odds ratio of \( 3/28 \), the reciprocal of what we got before.

The use of odds give an elegant way to restate Bayes’ theorem. Suppose again that we have an events \( A \) and \( \sim A \) (choosing from Urn1 or Urn2, which contain mixtures of white and green balls), and an event \( B \) (the ball chosen was green), that we know the prior probabilities, \( p(A), p(\sim A) \) (of choosing from each urn), and that we know \( p(B|A), pB(\sim A) \) (the proportion of green to white balls in each of the tow urns). From Bayes’ theorem (applied to both \( A \) and \( \sim A \)) we have

\[
p(A|B) = \frac{p(B|A)p(A)}{(p(B|A)p(A) + p(B|\sim A)p(\sim A))}, \tag{8}
\]

\[
p(\sim A|B) = \frac{p(B|\sim A)p(\sim A)}{(p(B|A)p(A) + p(B|\sim A)p(\sim A))}, \tag{9}
\]

(Notice that these probabilities sum to 1.) The right sides of the two equations have the same denominators so if we take the quotient of the left sides, which must equal the quotient of the right sides, on the right the denominators will cancel. The result is

**Theorem 2** *Theorem/Restatement of Bayes*

\[
\frac{p(A|B)}{p(\sim A|B)} = \frac{p(A|B)}{p(B|\sim A)} \cdot \frac{p(A)}{p(\sim A)}
\]

The second factor on the right, \( p(A)/p(\sim A) \) is just the prior odds on \( A \). The left side is the posterior odds (the odds given that \( B \) has occurred). The theorem says that we get the posterior odds by multiplying the prior odds by \( p(B|A)/p(B|\sim A) \). Applying this to the example in the previous section where we had \( p(A) = .8, p(\sim A) = .2, p(B|A) = .5, p(B|\sim A) = .75 \), we got
\[ p(A|B) = \frac{.4}{.55} = .72\ldots, \text{ so } p(\sim A|B) = \frac{.15}{.55}, \text{ giving a posterior odds ratio of } \frac{.4}{.15} = \frac{8}{3}. \] We could have computed this as follows, using the Restatement. The prior odds were \( \frac{.8}{.2} = 4 \) and \( \frac{p(B|A)}{p(B|\sim A)} = \frac{.5}{.75} = \frac{2}{3} \), so the posterior odds must be \( 4 \times \frac{2}{3} = \frac{8}{3} \). To get the actual posterior probabilities we must then take

\[ p = \frac{\frac{8}{3}}{1 + \left(\frac{8}{3}\right)} = \frac{\frac{8}{3}}{\frac{11}{3}} = \frac{8}{11} = .72\ldots \]

just as before.