Probability Theory
JOHN WORRALL

“Probability is the very guide of life.” Cicero *De Natura,* 5,12

1. Motivating the axioms – an example

Historically speaking, the mathematical theory of probability was developed by analysing games of chance. So let’s motivate the basic principles by taking a very simple game: drawing a single card ‘at random’ out of a well-shuffled pack of playing cards.

(a) The Outcome Space and Basic Events

Intuitively so long as the cards are well-shuffled and the draw is made non-systematically, each card has the same chance of being drawn. So there is an *outcome space*: the set of all possible outcomes right the way through from 2 of Clubs to the Ace of Spades. Each of the 52 members of the outcome space plausibly has the same probability: 1 in 52. (Probability theory is *not* constricted to the assumption of equal probabilities for all members of the outcome space, but it is easier to see what is going on in cases where that assumption is true.)

Call any of the outcomes in the outcome space, a ‘basic event’.

(b) Disjunctions : the Addition Law

Other *events* can then be characterised on the basis of those basic events. So the event of drawing a club is the event of drawing one the 13 individual cards that are clubs; the event of drawing a black card is the event of drawing one – any - of the 26 cards that are either clubs or spades. Set theoretically, the event of drawing a club is the union of the set of points in the sample space corresponding to the individual club cards. Intuitively the probability of drawing a club, any club, is 13 in 52,i.e. ¼. In other words we are adding here to get the probability of the complex event from the probabilities of the basic ones: P(either 2 of Clubs or 3 of Clubs or ..or Ace of Clubs) = P(2 of clubs) + P(3 of clubs) + ….+P(Ace of Clubs) = 1/52 + 1/52 …. + 1/52 (13 times in all) = 1/4 . Similarly P(Black card) = 1/52 + …+ 1/52 (26 times) = 26/52= ½.

But think of the event of drawing a card that is *either* a club *or* a face card. We just saw that P(Club) = ¼. As for P(face card), there are 12 face cards in all, 3 in each suit, so P(face card) = 12/52 = 3/13. So applying “the addition law” (naively, as it will turn out) we get P(either club or face card) = ¼ + 3/13 = 25/52.

But how many basic events satisfy this description?: 13 (clubs *including the club face cards)* + 3 (the 3 diamond face cards) + 3(the 3 heart face cards) + 3 (spades) = 22. So the right probability here should be 22/52 and 22/52 ≠ 25/52.

So the addition law does not always apply straightforwardly. Thinking about what went wrong here it is easy to see that in arriving at the incorrect answer of 25/52 we mistakenly *counted the club face cards twice*! (Once just as clubs and then again as (club) face cards.)

So the ***correct law for probabilities of disjunctions (or unions***) is

P(A or B) = P(A) + P(B) – P(A&B) (or set-theoretically P (A U B) = P(A) + P(B) – P(A ∩ B)).

(E.g. P (either club or face card) = P(club) + P(face card) – P(club & face card) = 13/52 + 12/52 – 3/52 = 22/52 which is as we saw the right answer.)

Where the two events in the disjunction are *mutually exclusive* – as they are for example for any of the basic events or for example the event of getting a club or a heart (you can't draw a card that is both), then the event of being both never occurs, its probability is therefore 0 and we get as a ***special case***:

When the events A and B are mutually exclusive, then P(A or B) = P(A) + P(B).

So for example P(either a club or a heart) = 13/52 + 13/52 (since P(club & heart) = 0) which gives the clearly correct answer: ½.

(c) Conjunctions: the Multiplication law?

What is the probability of drawing a card that is *both* a club *and* at the same time at least as high in value as a Jack? It is easy to work out what this probability must be in terms of basic events: only 4 cards – Jack, Queen, King and Ace of Clubs –satisfy both conditions (I am assuming throughout the standard Bridge ordering of values where 'aces are high'). So the correct probability is 4/52= 1/13.

But how does this probability, P(Club & ≥Jack), relate to the separate probabilities P(Club) and P(≥Jack)?

Let's take a simpler example. Suppose the game is instead that of tossing a fair coin twice. The 4 outcomes in this derivative outcome space are HH, HT, TH, TT (with of course H for heads and T for tails). The probabilities for each individual toss (given that the coin is fair) are P(H) = P(T) = ½. The probability of the *joint event* P(H(1), H(2)) = ¼ (HH is one of 4 equally probable outcomes in the joint sample space). So it looks like we get the joint probability (both of two events occur) by multiplying the individual probabilities:

(Conjectured Law for Joint Probabilities): for any two events A and B, P(A&B) = P(A).P(B)

This gives the correct result sometimes – it does so for the coin toss case as we saw and also for our most recent card case:
P(Club & ≥Jack) = 4/52;
P(Club) = 13/52 = ¼ and P(≥Jack) = 16/52 (there are a total of 16 cards in the pack that are ≥Jack – 4 from each of the 4 suits) = 4/13
So , applying this 'rule' gives P(Club & ≥Jack) = ¼ . 4/13 – yielding the correct result of 4/52.

But what is the probability of drawing, say, a card that is both a heart and a red card? Well there are 13 cards that are hearts, all of which are red, and none of the other red cards (i.e. the diamonds) are hearts (!) so there are 13 chances in 52 that the card will be both a heart and red and that's ¼.

But applying the conjectured law we get P(Heart & Red) = P(Heart). P(Red) = ¼ . ½= 1/8 ≠ ¼

Intuitively the problem is that there is a 'connection' between being a heart and being red (being a heart actually entails being red), whereas in the cases where the 'rule' gave the correct result the events concerned – turning up heads in the first toss and turning up heads in the second or being a club and being ≥Jack - are unconnected. If you know that the first toss produced a head, then (assuming that the coin is indeed fair) this gives you no information about whether the second toss will give you a head or a tail and so P(H(2)) remains ½; similarly knowing that a card is a club gives you no information about whether or not it is ≥Jack – *think this through*; on the other hand, knowing that the card that was drawn was red (you may just have glimpsed it quickly) changes the probability that it was a heart: this is now ½ (there are 26 red cards, 13 of which are hearts) rather than the original ¼; and conversely if you know that the card you drew was a heart, the probability that it was a red card, which was initially ½, is obviously changed to certainty, P = 1.

(d) Conditional probability and the correct multiplication law

These intuitive considerations lead us to an important further concept in probability theory: that of ***conditional probability***. We can talk about the probability of an event A occurring *given* that some other event B has occurred (for sure). What is the probability that the second toss of this fair coin will be H, given that the first toss was H? What is the probability that the card drawn from the pack was a heart, given that it was red? What is the probability that it will rain in London tomorrow, given that it rained in London today?

These are conditional probabilities, written P(A/B) where A is the event whose probability we are still seeking, but now given, or on the assumption, that the conditioning event B has occurred.

There are two importantly different cases, one where conditioning 'doesn't change the probability' and one where it does.

So P(H(2)/H(1)) – the probability of the second toss of the fair coin giving a head, given that the first did - = P(H(2)) =1/2. In such a case knowing the second event does not change the probability of the first and the two events are said to be *probabilistically independent*. That is, events A and B are probabilistically independent just in case P(A/B) = P(A).

In the card case, P(Card is a heart/Card is red) = ½ ≠ P(card is a heart) = ¼. (Similarly P(Card is red/Card is a heart) = 1 ≠ P(Card is red) = ½.) Where P(A/B) ≠P(A) the events A and B are said to be *probabilistically* ***de****pendent*.

So the
Correct Law for Joint Probabilities: for any two events A and B, P(A&B) = P(A).P(B/A)

Read intuitively: the probability that both A and B will occur is the probability that A will occur multiplied by the probability that B will occur, given that A has.

Notice
1. this again yields a **special case**: where the two events A and B *are* independent then, since that means that P(B/A) = P(B), our Law reduces to P(A&B) = P(A).P(B). [P(H(1) & H(2)) = ½.½= ¼.]
2. But of course the correct law also gives the right answer in cases where the 'Conjectured' law went wrong: P(Card is a heart & Card is red) should be ¼; and that is the answer we get from P(Card is a heart & Card is red)= P(heart) (=1/4) . P(red/heart) (=1).

3. Since the event A&B is the same event as the event B&A we would expect that P(A&B) should also be equal to P(B).P(B /A) and that in turn must imply the result that P(A/B)= P(A) if and only if P(B/A) = P(B). We shall in fact prove that this is true shortly.
4. The Correct Law for Joint Probabilities clearly yields an expression for P(B/A) as P(A&B)/P(B) [notice that this could also be written P(B&A)/P(B)] = but ***only if*** P(B) ≠ 0 [since the result of dividing any quantity by zero is undefined.]

[*Exercises*:
1. What is the probability that a card drawn at random is either a club or a black card? Show that the correct addition law yields the correct relationship between that probability and those of being a club and of being a black card.
2. What is the probability that a card drawn at random is either a club or a black card *and* a club? Apply both the addition law and law for joint probabilities to show that they yield the correct relationship between that probability and those of being a club and of being a black card.

2. The Probability Axioms and some important theorems

Formally a probability P is a non-negative, real-valued function having values on all the possible subsets of some basic outcome space S. In order to be a probability function P must satisfy:
1. P(A) ≥ 0 for all A in the domain of P
2. If S is the whole space, then P(S) = 1
3. P(A U B) = P(A) + P(B) – P(A ∩ B) (hence when A and B are mutually exclusive: P(A U B) = P(A) + P(B).)

3. generalises to a principle called 'countable additivity' and in higher powered applications in physics even to 'non-countable additivity' but these need not concern us.

Axioms 1.-3. generate the theory of absolute (or one place) probabilities, but as we saw there is also an important category of two-place, conditional probabilities. The two are connected via a 4^{th} axiom (sometimes also regarded as a 'definition'):

4. P(A/B) = P(A ∩ B)/P(B), IF P(B) ≠ 0
[I just switched round A and B compared to the result toward the end of section **1**.]

Since the algebra of sets under complementation, union and intersection is isomorphic to (i.e. structurally identical to) the algebra of propositions under negation, disjunction and conjunction, we could also regard these axioms as applying to propositions in which case 1 still applies, while 2, 3 and 4 become

2' If T is a tautology then P(T) = 1
3' P (A v B) = P(A) + P(B) – P(A&B) (hence when A and B are mutually exclusive: P(A v B) = P(A) + P(B))

4' P(A/B) = P(A&B)/P(B) where P(B) ≠0.

Those of you more familiar with basic logic than set theory will find it easier to follow the proofs of theorems and work the exercises by thinking of the axioms in this way. And so this is the way I will follow below.

Some easy theorems:

1. P(¬A) = 1 – P(A)
Proof: A and ¬A are mutually exclusive, hence P(Av¬A) = P(A) + P(¬A). But by axiom 2' since Av¬A is a tautology P(Av¬A) =1. Hence P(A) + P(¬A) = 1 and so P(¬A)= 1 – P(A)

2. If X is a contradiction P(X) = 0
Proof: ¬X is a tautology and hence by axiom 2' P(¬X)=1. But then by theorem 1, P(X)=0

3. If A and B are logically equivalent (A≡B) then P(A) = P(B)
Proof: If A≡B then A v ¬B is a tautology (check by truth tables if you are unclear!). Hence P(Av¬B) =1. But if A≡B then A and ¬B are mutually exclusive (contradictory), and so P(Av¬B) = P(A) + P(¬B) =1.Hence by theorem 1, P(A) + (1- P(B)) =1 from which it follows that P(A) =P(B).

4. If A logically entails B then P(A) ≤P(B) (No statement has higher probability than any of its logical consequences.)
Proof: If A logically entails B then B≡ (A v (B & ¬A)). [Exercise: show that this is true.] So by theorem 3, P(B) = P(A v (B & ¬A). But A and (B& ¬A) are mutually exclusive (contradictory) and so by axiom 3,
(*) P(B) = P(A) + P(B &¬A).
And finally, since that second term on the RHS of (*) must be ≥ 0 (by axiom 1), (*) implies that P(B)≥P(A), i.e. that P(A) ≤P(B).
Theorem 4 directly entails that for any A, P(A) ≤1 , so that , in conjunction with Axiom 1, we have the result that all probabilities lie between 0 and 1 inclusive.

5. For any A, $P(A) \leq 1$.
Proof: Any A logically entails any tautology T. [*Exercise*: explain why] Hence by theorem 4, $P(A) \leq P(T)$ and by axiom 2', $P(T) = 1$.

Although countable additivity would be an extra and somewhat controversial extra axiom, finite additivity is readily proved to hold from the basic axioms
6. Suppose the propositions A_i (i ε (1, …n) for some (any) finite n), are mutually exclusive, that is for any Ai, Aj, $i \neq j$, A_i entails $\neg A_j$, then
$P(A_1 v A_2 \ldots v A_n) = P(A_1) + P(A_2) \ldots + P(A_n)$.
Proof: Assume n = 1, in that case the result reduces to P(A1) = P(A1) and so it holds trivially.
Assume it holds for all k up to and including n-1, that is that
(*) $P((A_1 v A_2 \ldots v A_{n-1}) = P(A_1) + \ldots + P(A_n)$,
then since $P(A_1 v A_2 \ldots v A_n) = P((A_1 v A_2 \ldots v A_{n-1}) v A_n)$ and since $((A_1 v A_2 \ldots v A_{n-1})$ and A_n are mutually exclusive (each of A_1 through A_{n-1} entails ¬An and so, therefore does the disjunction – *Exercise*: check - $P((A_1 v A_2 \ldots v A_{n-1}) v A_n)) = P((A_1 v A_2 \ldots v A_{n-1}) + P(A_n)$
So by (*) $P((A_1 v A_2 \ldots v A_{n-1} v A_n) = P(A_1) + \ldots + P(A_{n-1}) + P(A_n)$

7. ('Theorem of Total Probability'): Let Ai be as in Theorem 6 ('mutually exclusive') and let $P(A_1 v A_2 \ldots v A_n) = 1$ (so that the Ai's 'cover the whole space'): then
for any B, $P(B) = P(B \& A_1) + \ldots P(B \& A_n)$
Proof: Each $(B \& A_j)$ is mutually exclusive of all the others, since each of the A_j's are (usual *exercise*). Moreover
$B \equiv (B \& A_1) v \ldots v (B \& A_n) v (B \& \neg(A_1 v A_2 \ldots v A_n))$ (again usual *exercise*).
Letting that last disjunction be A, we have from theorem 3:
$(P(B) = P(B \& A_1) + \ldots + P(B \& A_n) + P(B \& \neg(A))$.
But, by assumption, $P(A_1 v A_2 \ldots v A_n) = 1$ ie $P(A)=1$ so $P(\neg A)=0$ and so, since $B \& \neg A$ logically entails $\neg A$, $P(B \& \neg A)$ is also =0 and hence drops out of (*) leaving the result.
Notice that as a ***special case*** of theorem 7, when we just break down the possibilities into A and ¬A, we get that for any B, $P(B) = P(B \& A) + P(B \& \neg A)$.

Important Exercise:
Apply axiom 4' to theorem 7 to prove that
If the A_i are as in theorem 7 ($P(A_1 v A_2 \ldots v A_n) = 1$ and the A_j's mutually exclusive) and moreover $P(A_j) > 0$ for any j, then
for any B, $P(B) = P(A_1) P(B/A_1) + P(A_2)P(B/A_2) \ldots + P(A_n)P(B/A_n)$

Finally, one result about conditional probabilities mentioned earlier.
8. $P(A/B) \neq P(A)$ if and only if $P(B/A) \neq P(B)$

Proof: Suppose P(A/B)≠P(A) – say P(A/B)>P(A) (an identically structured proof goes through if P(A/B)<P(A)). Then P(A&B)/P(B) > P(A). So P(A&B)>P(A).P(B). So, dividing both sides now by P(A) , P(A&B)/P(A) >P(B), but since A&B ≡B&A, this is the same as P(B&A)/P(A)>P(B) – i.e.P(B/A)> P(B).

So this yields the result that if A is probabilistically dependent on B , then B is automatically probabilistically dependent on A. (Knowing that a card was red affects the probability that it is a heart, but equally knowing that it is heart affects the probability that it is red.)

3. "Inverse Probability": Bayes' Theorem

Probably the most important theorem in Probability – certainly the most discussed – is Bayes' Theorem.
In its simplest form it states that for any two events A and B,
P(A/B) = P(B/A).P(A)/ P(B).

Its proof, as we shall see in a moment, is straightforward and on the face of it, the theorem does not look particularly exciting. However it is the basis from which so called 'inverse' or sometimes 'indirect' probabilities can be calculated.

Let's go back to coin-tossing. If we assume that a coin is fair, then we can readily work out the probabilities (in this context often referred to as the direct probabilities) of various outcomes – suppose the trial is to be that of tossing the coin 4 times and the statistic we are interested in the number of tosses out of the 4 that are heads. Then those probabilities are :
P(4 heads) = P(0 heads) = $(½)^4$ = 1/16
P(3 heads) = P(1 head) = 4/16
P(2 heads) = 6/16
(Each specific outcome: H first, followed by a T, followed by etc has the same probability viz 1/16, and then you just use the formula P(r heads out of 4) = ${}^4C_r$/16, where is ${}^4C_r$ the number of ways of selecting r things from a total of 4 things – this of course generalises to any n. And it is easy to prove that ${}^nC_r$ = n!/(r!(n-r)!))

Now consider another coin which we assume or know to be biased. Let's suppose that the probability of getting a head on any single toss of coin 2 is ¾. Then of course we get different probabilities ('direct probabilities') for the various possible outcomes. Now:
P(4 heads) = $(¾)^4$ = 81/256

$P(3heads) = 3.(1/4).(3/4)^3$
$P(2heads) = 6.(1/4)^2.(3/4)^2$
$P(1\ head) = 3.\ (1/4)^3.(3/4)$
$P(0\ heads) = (1/4)^4$

But next suppose that we have the two coins in a box, that we cannot distinguish them physically, and that we select one at random and observe the outcome of tossing it 4 times. Can we work out, on the basis of that observation, what the probability is that that outcome was produced by the fair coin (or by the biased coin)? Bayes' theorem allows us to do so.

Let A be the coin is fair (so ¬A is the coin has p=3/4 of producing heads) and B be, say, the event of producing 1 head out of 4. Assume moreover that the fact that we simply chose one coin out of the box at random means that the initial ('prior') probability of A , P(A) (=P(¬A)) =1/2.

We already calculated the value of P(B/A) – the probability of getting 1 head out of 4 with the fair coin – it is ¼. What we would like to know is P(A/B) – that is, the probability that it was the fair coin that produced this result. (That is we want to 'invert' the probability (P(B/A) that we know.) Inspection of Bayes' theorem shows that since we know P(B/A) (=1/4) and P(A) (=1/2 - we drew a coin at random), we need one further probability namely P(B) - the probability of getting 1 head out of 4 in this whole experiment. Well there are two ways that we could have got 1 head out of 4 – with the fair coin and with the biased coin, and there was the same chance, ½, of drawing either coin.
In general, P(B) = P(A).P(B/A) + P(¬A). P(B/¬A)
and so here P(B) = ½ P(B/A)+ ½ P(B/¬A). If you think about this a little, it is obviously the right result – there was a half chance of getting the fair coin, and in that case the probability of the observed result was ¼, and there was a half chance of getting the biased coin and in that case the probability of the observed result was, see earlier, $3.\ (1/4)^3.(3/4)(= 9/4^4)$; so the overall probability of 1 head is $½\ .\ ¼ + ½\ .\ 3.\ 9/4^4= 91/512$. (This is in fact an application of theorem of total probability, as you will see if you apply axiom 4', the 'definition' of conditional probability to the special case of theorem 7.)

Finally then P(A/B)= ¼. ½/ 91/512 = 64/91. So, despite the fact that something has happened which is relatively unlikely if the hypothesis A is correct, its probability has gone up from a 'prior' of ½ to a 'posterior' close to 2/3. If this seems surprising reflect on the fact that the observed outcome (1 Head in 4, remember) would be even less likely were the alternative (P(H) = ¾) true.

Exercise Apply Bayes' theorem to this same setup to calculate the 'posterior' probability of A given the outcome B of 0 heads. Again you will find that this greater than A's prior, despite the fact that A attributes minimal probability to that outcome .

Proof of Bayes' Theorem

Theorem: P(A/B) = P(B/A). P(A)/P(B)

Proof:

By 4', P(B&A) = P(B). P(A/B).

So (*), P(A/B) = P(B&A)/P(B)

But, equally, P(B/A) = P(B&A)/P(A) and so ($) P(B&A) = P(B/A).P(A)

Substituting the expression for P(B&A) in ($) into (*) gives

P(A/B) = P(B/A). P(A)/P(B) as required.

As this shows, Bayes' theorem is a straightforward theorem of probability theory and hence uncontroversial. A whole general account of scientific reasoning has, however, been based on the theorem – called Bayesianism. And this although powerful and powerfully supported *has* proved controversial.

The basic approach to evidence underwritten by Bayesian is the intuitively very appealing principle that evidence e supplies evidence for a hypothesis h if, and to the extent that, it 'raises h's probability' of being true. That is e supports h if P(h/e)>P(h) and it yields the more support, the greater the difference - that is the higher the value of P(h/e) –P(e). Written in terms of h and e, Bayes's theorem becomes:

P(h/e) = P(e/h).P(h)/P(e)

P(e/h) is usually called the likelihood term, measuring how likely some piece of evidence is in the light of h, while P(h) is the prior probability of h. P(e) is the probability of the evidence, as it were independently of h. Because P(h/e) is inversely proportional to P(e), Bayes' theorem tells us that P(h/e) is higher, other things being equal, the lower the probability of e. This underwrites an intuitive principle about evidence that we will come across a number of times.

How should we think about P(e)? The following version of the theorem yields a good insight (it really just elaborates on the reasoning involved in our coin 'inverse probability' example).

Bayes' Theorem (2nd version)

Suppose that the only possible hypotheses at issue are h_1, …, h_n so that P(h_1v …v h_n) = 1, suppose also that each of the h_i entails the negation of any of the others (they are 'mutually exclusive' as well as exhaustive) and finally that both P(h_i) for any i and P(e) are >0, then

$P(h_k/e) = (P(e/h_k)P(h_k))/\Sigma_i P(e/h_i)P(h_i)$

This version reveals P(e) as the weighted average of its likelihood on all the different rival hypotheses – weighted by the prior probability of those hypotheses. So P(e) is the higher (and therefore has less effect on the posterior probability of any particular h) the more likely e is in the light of plausible rivals to h; and conversely P(e) is lower (even if e is very probable given h) and hence gives a bigger 'boost' to the credibility of h, the smaller its probability in the light of any rivals that have any real plausibility (reasonable priors).

Reading

The above treatment is based to a good extent on that in Howson and Urbach, 3rd edition, chapter 2.

Further Reading

Other parts of that same book.