18

The A_p distribution and rule of succession

Inside every Non-Bayesian, there is a Bayesian struggling to get out.
Dennis V. Lindley

Up to this point, we have given our robot fairly general principles by which it can convert information into numerical values of prior probabilities, and convert posterior probabilities into definite final decisions; so it is now able to solve lots of problems. But it still operates in a rather inefficient way in one respect. When we give it a new problem, it has to go back into its memory (this proposition that we have denoted by X or I, which represents everything it has ever learned). It must scan its entire memory archives for anything relevant to the problem before it can start working on it. As the robot grows older this gets to be a more and more time-consuming process.

Now, human brains don't do this. We have some machinery built into us which summarizes our past conclusions, and allows us to forget the details which led us to those conclusions. We want to see whether it is possible to give the robot a definite mechanism by which it can store general conclusions rather than isolated facts.

18.1 Memory storage for old robots

Note another thing, which we will see is closely related to this problem. Suppose you have a penny and you are allowed to examine it carefully, and convince yourself that it is an honest coin; i.e. accurately round, with head and tail, and a center of gravity where it ought to be. Then you're asked to assign a probability that this coin will come up heads on the first toss. I'm sure you'll say $1/2$. Now, suppose you are asked to assign a probability to the proposition that there was once life on Mars. Well, I don't know what your opinion is there, but on the basis of all the things that I have read on the subject, I would again say about $1/2$ for the probability. But, even though I have assigned the same 'external' probabilities to them, I have a very different 'internal' state of knowledge about those propositions.

To see this, imagine the effect of getting new information. Suppose we tossed the coin five times and it comes up tails every time. You ask me what's my probability for heads on the next throw; I'll still say $1/2$. But if you tell me one more fact about Mars, I'm ready to change my probability assignment completely. There is something which makes

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

my state of belief very stable in the case of the penny, but very unstable in the case of Mars.[1]

This might seem to be a fatal objection to probability theory as logic. Perhaps we need to associate with a proposition not just a single number representing plausibility, but two numbers: one representing the plausibility, and the other how stable it is in the face of new evidence. And so, a kind of two-valued theory would be needed. In the early 1950s, the writer gave a talk at one of the Berkeley statistical symposiums, expounding this viewpoint.

But now, with more mature reflection we think that there is a mechanism by which our present theory automatically contains all these things. So far, all the propositions we have asked the robot to think about are 'Aristotelian' ones of two-valued logic: they had to be either true or false. Suppose we bring in new propositions of a different type. It doesn't make sense to say the proposition is either true or false, but still we are going to say that the robot associates a real number with it, which obeys the rules of probability theory. Now, these propositions are sometimes hard to state verbally; but we noticed before that if we give the probabilities conditional on X for all propositions that we are going to use in a given problem, we have told you everything about X which is relevant to that mathematical problem (although of course, not everything about its meaning and significance to us, that may make us interested in the problem). So, we introduce a new proposition A_p, defined by

$$P(A|A_pE) \equiv p, \tag{18.1}$$

where E is any additional evidence. If we had to render A_p as a verbal statement, it would come out something like this:

$$\begin{aligned} A_p \equiv\ & \text{regardless of anything else you may have been told,} \\ & \text{the probability of } A \text{ is } p. \end{aligned} \tag{18.2}$$

Now, A_p is a strange proposition, but if we allow the robot to reason with propositions of this sort, Bayes' theorem guarantees that there's nothing to prevent it from getting an A_p worked over onto the left side in its probabilities: $P(A_p|E)$. What are we doing here? It seems almost as if we are talking about the 'probability of a probability'.

Pending a better understanding of what that means, let us adopt a cautious notation that will avoid giving possibly wrong impressions. We are not claiming that $P(A_p|E)$ is a 'real probability' in the sense that we have been using that term; it is only a number which is to obey the mathematical rules of probability theory. Perhaps its proper conceptual meaning will be clearer after getting a little experience using it. So let us refrain from using the prefix symbol p; to emphasize its more abstract nature, let us use the bare bracket symbol notation $(A_p|E)$ to denote such quantities, and call it simply 'the density for A_p, given E'.

We defined A_p by writing an equation. You ask what it means, and we reply by writing more equations. So let's write the equations: if X says nothing about A except that it is

[1] Note in passing a simple counter-example to a principle sometimes stated by philosophers, that theories cannot be proved true, only false. We seem to have just the opposite situation for the theory that there was once life on Mars. To prove it false, it would not suffice to dig up every square foot of the surface of Mars; to prove it true one needs only to find a single fossil.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

possible for A to be true, and also possible for it to be false, then, as we saw in case of the 'completely ignorant population' in Chapter 12,

$$(A_p|X) = 1, \qquad 0 \le p \le 1. \tag{18.3}$$

The transformation group arguments of Chapter 12 apply to this problem. As soon as we have this, we can use Bayes' theorem to compute the density for A_p, conditional on the other things. In particular,

$$(A_p|EX) = (A_p|X)\frac{P(E|A_pX)}{P(E|X)} = \frac{P(E|A_p)}{P(E|X)}. \tag{18.4}$$

Now,

$$P(A|E) = \int_0^1 \mathrm{d}p\,(AA_p|E). \tag{18.5}$$

The propositions A_p are mutually exclusive and exhaustive (in fact, every A_p flatly and dogmatically contradicts every other A_q), so we can do this. We're just going to apply all of our mathematical rules with total disregard of the fact that A_p is a funny kind of proposition. We believe that these rules form a consistent way of manipulating propositions. But now we recognize that consistency is a purely *structural* property of the rules, which could not depend on the particular semantic meaning you and I might attach to a proposition. So now we can blow up the integrand of (18.5) by the product rule:

$$P(A|E) = \int_0^1 \mathrm{d}p\; P(A|A_pE)(A_p|E). \tag{18.6}$$

But from the definition (18.1) of A_p, the first factor is just p, and so

$$P(A|E) = \int_0^1 \mathrm{d}p\; p\,(A_p|E). \tag{18.7}$$

The probability which our robot assigns to proposition A is just the *first moment* of the density for A_p. Therefore, the density for A_p should contain more information about the robot's state of mind concerning A, than just the probability for A. Our conjecture is that the introduction of propositions of this sort solves both of the problems mentioned, and also gives us a powerful analytical tool for calculating probabilities.

18.2 Relevance

To see why we propose our conjecture, let's note some lemmas about relevance. Suppose this evidence E consists of two parts, $E = E_aE_b$, where E_a is relevant to A and, given E_a, E_b is not relevant:

$$P(A|E) = P(A|E_aE_b) = P(A|E_a). \tag{18.8}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

By Bayes' theorem, it follows that, given E_a, A must also be irrelevant to E_b, for

$$P(E_b|AE_a) = P(E_b|E_a)\frac{P(A|E_bE_a)}{P(A|E_a)} = P(E_b|E_a). \tag{18.9}$$

Let's call this property 'weak irrelevance'. Now, does this imply that E_b is irrelevant to A_p? Evidently not, for (18.8) says only that the first moments of $(A_p|E_a)$ and $(A_p|E_aE_b)$ are the same. But suppose that, for a given E_b, (18.8) holds independently of what E_a might be; call this 'strong irrelevance'. Then we have

$$P(A|E) = \int_0^1 \mathrm{d}p\, p\,(A_p|E_aE_b) = \int_0^1 \mathrm{d}p\, p\,(A_p|E_a). \tag{18.10}$$

But if this is to hold for all $(A_p|E_a)$, the integrands must be the same:

$$(A_p|E_aE_b) = (A_p|E_a), \tag{18.11}$$

and from Bayes' theorem it follows as in (18.9) that A_p is irrelevant to E_b:

$$P(E_b|A_pE_a) = P(E_b|E_a) \tag{18.12}$$

for all E_a.

Now, suppose our robot gets a new piece of evidence, F. How does this change its state of knowledge about A? We could expand directly by Bayes' theorem, which we have done before, but let's use our A_p this time:

$$P(A|EF) = \int_0^1 \mathrm{d}p\, p\,(A_p|EF) = \int_0^1 \mathrm{d}p\, p\,(A_p|E)\frac{P(F|A_pE)}{P(F|E)}. \tag{18.13}$$

In this likelihood ratio, any part of E that is irrelevant to A_p can be struck out; because, by Bayes' theorem, it is equal to

$$\frac{P(F|A_pE_aE_b)}{P(F|E_aE_b)} = \frac{P(F|A_pE_a)\left[\frac{P(E_b|FA_pE_a)}{P(E_b|A_pE_a)}\right]}{P(F|E_a)\left[\frac{P(E_b|FE_a)}{P(E_b|E_a)}\right]} = \frac{P(F|A_pE_a)}{P(F|E_a)}, \tag{18.14}$$

where we have used (18.12).

Now if E_a still contains a part irrelevant to A_p, we can repeat this process. Imagine this carried out as many times as possible; the part E_{aa} of E that is left contains nothing at all that is irrelevant to A_p. E_{aa} must then be some statement only about A. But then, by definition (18.1) of A_p, we see that A_p automatically cancels out E_{aa} in the numerator: $(F|A_pE_{aa}) = (F|A_p)$. And so we have (18.13) reduced to

$$P(A|EF) = \frac{1}{P(F|E_{aa})}\int_0^1 \mathrm{d}p\, p\,(A_p|E)P(F|A_p). \tag{18.15}$$

The weak point in this argument is that we have not proved that it is always possible to resolve E into a completely relevant part and completely irrelevant part. However, it is easy to show that in many applications it *is* possible. So, let's just say that the following results

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

apply to the case where the prior information is ‘completely resolvable’. We have not shown that it is the most general case; but we do know that it is not an empty one.

18.3 A surprising consequence

Now, $(F|E_{aa})$ is a troublesome thing which we would like to eliminate. It’s really just a normalizing factor, and we can eliminate it the way we did in Chapter 4: by calculating the odds on A instead of the probability. This is just

$$O(A|EF) = \frac{P(A|EF)}{P(\overline{A}|EF)} = \frac{\int_0^1 \mathrm{d}p\, p\, (A_p|E)P(F|A_p)}{\int_0^1 \mathrm{d}p (A_p|E)P(F|A_p)(1-p)}. \tag{18.16}$$

The significant thing here is that the proposition E, which for this problem represents our prior information, now appears only in the density $(A_p|E)$. This means that *the only property of E which the robot needs in order to reason out the effect of new information is this density* $(A_p|E)$. Everything that the robot has ever learned which is relevant to proposition A may consist of millions of isolated separate facts. But when it receives new information, it does not have to go back and search its entire memory for every little detail of its information relevant to A. Everything it needs in order to reason about A from that past experience is contained summarized in this one function, $(A_p|E)$.

So, for each proposition A about which it is to reason, the robot can store a density function $(A_p|E)$ like that in Figure 18.1. Whenever it receives new information F, it will be well advised to calculate $(A_p|EF)$, and then it can erase the previous $(A_p|E)$ and for the future store only $(A_p|EF)$. By this procedure, every detail of its previous experience is taken into account in future reasoning about A.

This suggests that in a machine which does inductive reasoning, the memory storage problem may be simpler than it is in a machine which does only deductive reasoning. This does not mean that the robot is able to throw away all of its past experience, because there is

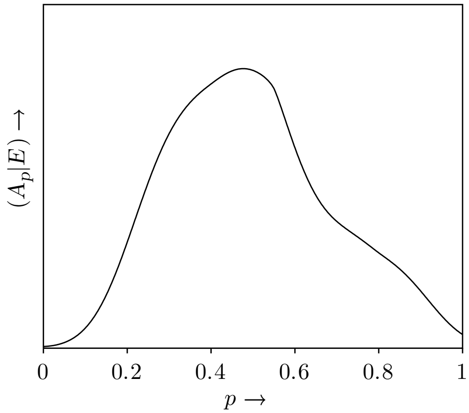

Fig. 18.1. An example A_p distribution.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

always a possibility that some new proposition will come up which it has not had to reason about before. And, whenever this happens, then of course it *will* have to go back into its original archives and search for every scrap of information it has relevant to this proposition.

With a little introspection, we would all agree that this is just what goes on in our minds. If you are asked how plausible you regard some proposition, you don't go back and recall all the details of everything that you ever learned about this proposition. You recall your previous state of mind about it. How many of us can still remember the argument which first convinced us that $\mathrm{d}\sin(x)/\mathrm{d}x = \cos(x)$? But, unlike the robot, when you or I are confronted with some entirely new proposition Z, we do not have the ability to carry out a full archival search.

Let's look once more at (18.15). If the new information F is to make any appreciable change in the probability of A, we can see from this integral what has to happen. If the density $(A_p|E)$ was already very sharply peaked at one particular value of p, then $P(F|A_p)$ will have to be even more sharply peaked at some other value of p, if we are going to get any appreciable change in the probability. On the other hand, if the density $(A_p|E)$ is very broad, any small slope in $P(F|A_p)$ can make a big change in the probability which the robot assigns to A.

So, the stability of the robot's state of mind when it has evidence E is determined, essentially, by the *width* of the density $(A_p|E)$. There does not appear to be any single number which fully describes this stability. On the other hand, whenever it has accumulated enough evidence so that $(A_p|E)$ is fairly well peaked at some value of p, then the variance of that distribution becomes a pretty good measure of how stable the robot's state of mind is. The greater amount of previous information it has collected, the narrower its A_p-distribution will be, and therefore the harder it will be for any new evidence to change that state of mind.

Now we can see the difference between the penny and Mars. In the case of the penny, my $(A_p|E)$ density, based on my prior knowledge, is represented by a curve something like that shown in Figure 18.2(a). In the case of previous life on Mars, my state of knowledge is

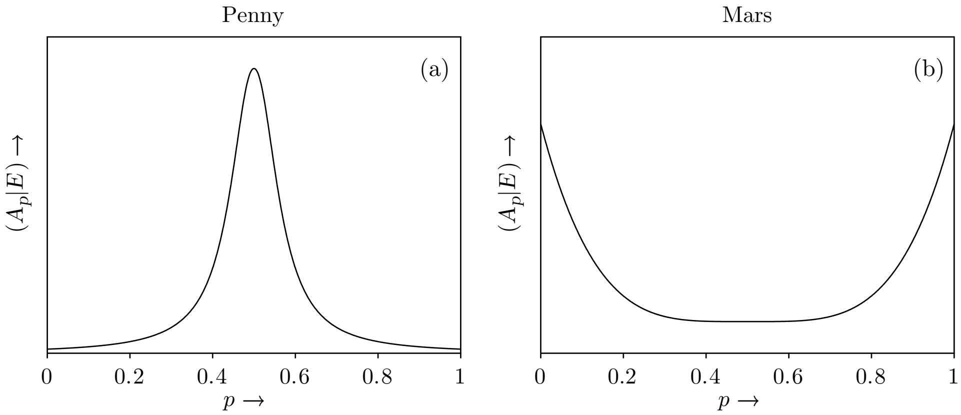

Fig. 18.2. Two A_p distributions having the same first moments, but representing very different states of knowledge.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

described by an $(A_p|E)$ density something like that shown in Figure 18.2(b), qualitatively. The first moment is the same in the two cases, so I assign probability $1/2$ to either one; nevertheless, there's all the difference in the world between my state of knowledge about those two propositions, and this difference is represented in the $(A_p|E)$ densities.

Ideas very much like this have arisen in other contexts. While the writer was first speculating on these ideas, a newspaper story appeared entitled: 'Brain Stockpiles Man's Most Inner Thoughts'. It starts out:

> Everything you have ever thought, done, or said – a complete record of every conscious moment – is logged in the comprehensive computer of your brain. You will never be able to recall more than the tiniest fraction of it to memory, but you'll never lose it either. These are the findings of Dr Wilder Penfield, Director of the Montreal Neurological Institute, and a leading Neurosurgeon. The brain's ability to store experiences, many lying below consciousness, has been recognized for some time, but the extent of this function is recorded by Dr Penfield.

Now, there are several examples given, of experiments on patients suffering from epilepsy. Stimulation of a definite location in the brain recalled a definite experience from the past, which the patients had not been able to recall to memory previously. Here are the concluding sentences of the article. Dr Penfield now says:

> This is not memory as we usually use the word, although it may have a relation to it. No man can recall by voluntary effort such a wealth of detail. A man may learn a song so he can sing it perfectly, but he cannot recall in detail any one of the many times he heard it. Most things that a man is able to recall to memory are generalizations and summaries. If it were not so, we might find ourselves confused by too great a richness of detail.

This is exactly the hint we needed to form a clearer idea of what the A_p density means conceptually.

18.4 Outer and inner robots

We know from overwhelming evidence, of which the above is only a small part, that human brains have two different functions: a conscious mind and a subconscious one. They work together in some kind of cooperation. The subconscious mind is probably at work continually throughout life. It solves problems and communicates information to the conscious mind under circumstances not under our conscious control; everyone who has done original thinking about difficult problems has experienced this, and many (Henri Poincaré, Jacques Hadamard, Wm. Rowan Hamilton, Freeman Dyson) have recorded the experience for others to read. A communication from the subconscious mind appears to us as a sudden inspiration that seems to come out of nowhere when we are relaxed and not thinking consciously about the problem at all; instantly, we feel that we understand the problem that has perplexed us for weeks.[2]

[2] The writer has experienced this several times when, in unlikely situations like riding a tractor on his farm, he suddenly saw how to prove something long conjectured. But the inspiration does not come unless the conscious mind has prepared the way for it by intense concentration on the problem.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

Now, if the human brain can operate on two different levels, so can our robot. Rather than trying to think of a 'probability of a probability', we may think of two different levels of reasoning: an 'outer robot' in contact with the external world and reasoning about it; and an 'inner robot' who observes the activity of the outer robot and thinks about it. The conventional probability formulas that we used before this chapter represent the reasoning of the outer robot; the A_p density represents the inner robot at work. But we would like our robot to have one advantage over the human brain. The outer robot should not be obliged as we are to wait for the inspiration from within; it should have the power to call at will upon the services of the inner robot.

Looking at the A_p distribution this way makes it much less puzzling conceptually. The outer robot, thinking about the real world, uses Aristotelian propositions referring to that world. The inner robot, thinking about the activities of the outer robot, uses propositions that are not Aristotelian in reference to the outer world, but they are still Aristotelian in its context, in reference to the thinking of the outer robot; so, of course, the same rules of probability theory will apply to them. The term 'probability of a probability' misses the point, since the two probabilities are at different levels.

Having had this much of a glimpse of things, our imagination races on far beyond it. The inner robot may prove to be more versatile than merely calculating and storing A_p densities; it may have functions that we have not yet imagined. Furthermore, could there be an 'inner inner' robot, twice removed from the real world, which thinks about the activity of the inner one? What prevents us from having a nested hierarchy of such robots, each inner to the next? Why not several parallel hierarchies, concerned with different contexts?

Questions like this may seem weird, until we note that just this same hierarchy has evolved already in the development of computers and computer programming methods. Our present microcomputers operate on three discernible hierarchical levels of activity, the inner 'BIOS' code which contacts the machine hardware directly, the 'COMMAND SHELL' which guards it from the outer world while sending information and instructions back and forth between them, and the outer level of human programmers who provide the 'high level' instructions representing the conscious ultimate purpose of the machine level activity. Furthermore, the development of 'massively parallel' computer architecture has been underway for several years.

In the evolution of computers this represented such a natural and inevitable division of labor that we should not be surprised to realize that a similar division of labor occurred in the evolution of the human brain. It has an inner 'BIOS' level which in some way exerts direct control over the body's biological hardware (such as rate of heartbeat and levels of hormone secretion), a 'COMMAND SHELL' which receives 'high level' instructions from the conscious mind and converts them into the finely detailed instructions needed to execute such complex activities as walking or playing a violin, without any need for the conscious mind to be aware of all those details. Then in some aspects of the present organization of the brain, not yet fully understood, we may be seeing some aspects of the future evolution of computers; in particular of our robot.

The idea of a nested hierarchy of robots, each thinking about propositions on a different level, is in some ways similar to Bertrand Russell's 'theory of types', which he

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

introduced as a means of avoiding some paradoxes that arose in the first formulation of his *Principia Mathematica*. There may be a relationship between them; but these efforts at what Peano and Poincaré called 'logistic' made in the early 20th century, are now seen as so flawed and confused – with an unlimited proliferation of weird and self-contradictory definitions, yet with no recognition of the concept of information – that it seems safest to scrap this old work entirely and rebuild from the start using our present understanding of the role of information and our new respect for Kronecker's warnings, so appropriate in an age of computers, that *constructibility* is the first criterion for judging whether a newly defined set or other mathematical object makes any sense or can serve any useful purpose.

Our opening quotation from Dennis Lindley (made during a talk at a Bayesian seminar in the early 1980s) fits in nicely with these considerations and with our remarks in Chapter 5 about visual perception. There we noted that any reasoning format whose results conflict with Bayesian principles would place a creature at a decided survival disadvantage, so evolution by Darwinian natural selection would automatically produce brains which reason in the Bayesian format. But the outer brain can become corrupted by false indoctrination from contact with the outer world – even to the point of becoming anti-Bayesian – while the inner brain, protected from this, retains its pristine Bayesian purity. Thus, Lindley's remark, made as a kind of joke, may be quite literally true.

Here, however, we are treading on the boundaries of present knowledge, so the above material is necessarily a tentative, preliminary exploration of a possibly large new territory (call it wild speculation if you prefer), rather than expounding a well-established theory. With these cautions in mind, let us examine some concrete examples which follow from the above line of thought, but can also be justified independently.

18.5 An application

Now let's imagine that a 'random' experiment is being performed. From the results of the experiment in the past, we want to do the best job we can of predicting results in the future. To make the problem a definite one, introduce the propositions:

$$X \equiv \text{For each trial we admit two prior hypotheses: } A \text{ true, and } A \text{ false.}$$

The underlying 'causal mechanism' is assumed the same at every trial. This means, for example, that (1) the probability assigned to A at the nth trial does not depend on n, and (2) evidence concerning the results of past trials retains its relevance for all time; thus for predicting the outcome of trial 100, knowledge of the result of trial 1 is just as relevant as is knowledge of the result of trial 99. There is no other prior evidence.

$$N_n \equiv A \text{ true } n \text{ times in } N \text{ trials in the past.}$$
$$M_m \equiv A \text{ true } m \text{ times in } M \text{ trials in the future.}$$

The verbal statement of X suffers from just the same ambiguities that we have found before, and which have caused so much trouble and controversy in the past. One of the important points we want to put across here is that we have not defined the prior information precisely until we have given, not just verbal statements, but equations, which show how we have

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

translated them into mathematics by specifying the prior probabilities to be used. In the present problem, this more precise statement of X is, as before,

$$(A_p|X) = 1, \quad 0 \le p \le 1, \tag{18.17}$$

with the additional understanding (part of the prior information for this particular problem) that the *same* A_p distribution is to be used for calculations pertaining to all trials. What we are after is $P(M_m|N_n)$. Firstly, note that by many repetitions of our product and sum rules in the same way that we found Eq. (9.34), we have the binomial distributions

$$\begin{aligned} P(N_n|A_p) &= \binom{N}{n} p^n (1-p)^{N-n}, \\ P(M_m|A_p) &= \binom{M}{m} p^m (1-p)^{M-m}, \end{aligned} \tag{18.18}$$

and at this point we see that, although A_p sounds like an awfully dogmatic and indefensible statement to us the way we introduced it, this is actually the way in which probability *is* introduced in almost all present textbooks. One postulates that an event possesses some intrinsic, 'absolute' or 'physical' probability, whose numerical value we can never determine exactly. Nevertheless, no one questions that such an 'absolute' probability exists. Cramér (1946, p. 154), for example, takes it as his fundamental axiom. That is just as dogmatic a statement as our A_p; and we think it is, in fact, just our A_p. The equations we see in current textbooks are all like the two above; whenever p appears as a *given* number, an adequate notation would show that there is an A_p hiding invisibly in the right-hand side of the probability symbols.

Mathematically, the main functional differences between what we are doing here and what is done in current textbooks are: (1) we recognize the existence of that right-hand side of *all* probabilities, whether or not an A_p is hiding in them; and (2) thanks to Cox's theorems, we are not afraid to use Bayes' theorem to work any proposition – including A_p – back and forth from one side of our symbols to the other. In refusing to make free use of Bayes' theorem, orthodox writers are depriving themselves of the most powerful single principle in probability theory. When a problem of inference is studied long enough, sometimes through a string of *ad hockeries* for decades, one is always forced eventually to a conclusion that could have been derived in three lines from Bayes' theorem. But those cases refer to 'external' probabilities at the interface between the robot and the outside world; now we shall see that Bayes' theorem is equally powerful and indispensable for manipulating 'inner' probabilities.

Now we need to find the prior probability $P(N_n|X)$. This is determined already from $(A_p|X)$, for our trick of resolving a proposition into mutually exclusive alternatives gives us

$$P(N_n|X) = \int_0^1 \mathrm{d}p\,(N_n A_p|X) = \int_0^1 \mathrm{d}p\, P(N_n|A_p)(A_p|X) = \binom{N}{n} \int_0^1 \mathrm{d}p\, p^n (1-p)^{N-n}. \tag{18.19}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

The integral we have to evaluate is the complete Beta-function:

$$\int_0^1 \mathrm{d}x\, x^r (1-x)^s = \frac{r!s!}{(r+s+1)!}. \tag{18.20}$$

Thus, we have

$$P(N_n|X) = \begin{cases} \dfrac{1}{N+1} & 0 \le n \le N \\ 0 & N < n, \end{cases} \tag{18.21}$$

i.e. just the uniform distribution of maximum entropy; $P(M_m|X)$ is found similarly. Now we can turn (18.18) around by Bayes' theorem:

$$(A_p|N_n) = (A_p|X)\frac{P(N_n|A_p)}{P(N_p|X)} = (N+1)P(N_n|A_p), \tag{18.22}$$

and so finally the desired probability is

$$P(M_m|N_n) = \int_0^1 \mathrm{d}p\,(M_m A_p|N_n) = \int_0^1 \mathrm{d}p\, P(M_m|A_p N_n)(A_p|N_n). \tag{18.23}$$

Since $P(M_m|A_p N_n) = P(M_m|A_p)$ by the definition of A_p, we have worked out everything in the integrand. Substituting into (18.23), we have again an Eulerian integral, and our result is

$$P(M_m|N_n) = \frac{\dbinom{n+m}{n}\dbinom{N+M-n-m}{N-n}}{\dbinom{N+M+1}{M}}. \tag{18.24}$$

Note that this is not the same as the hypergeometric distribution (3.22) of sampling theory. Let's look at this result first in the special case $M = m = 1$; it then reduces to the probability of A being true in the next trial, given that it has been true n time in the previous N trials. The result is

$$P(A|N_n) = \frac{n+1}{N+2}. \tag{18.25}$$

We recognize Laplace's rule of succession, which we found before and discussed briefly in terms of urn sampling in (6.29)–(6.46). Now we need to discuss it more carefully, in a wider context.

18.6 Laplace's rule of succession

This rule occupies a supreme position in probability theory; it has been easily the most misunderstood and misapplied rule in the theory, from the time Laplace first gave it in 1774. In almost any book on probability, this rule is mentioned very briefly, mainly in order to warn the reader not to use it. But we must take the trouble to understand it, because in our design of this robot Laplace's rule is, like Bayes' theorem, one of the most important

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

constructive rules we have. It is a 'new' rule (i.e. a rule in addition to the principle of indifference and its generalization, maximum entropy) for converting raw information into numerical values of probabilities, and it gives us one of the most important connections between probability and frequency.

Poor old Laplace has been ridiculed for over a century because he illustrated use of this rule by calculating the probability that the sun will rise tomorrow, given that it has risen every day for the past 5000 years.[3] One obtains a rather large factor (odds of $5000 \times 365.2426 + 1 = 1\,826\,214 : 1$) in favor of the sun rising again tomorrow. With no exceptions at all as far as we are aware, modern writers on probability have considered this a pure absurdity. Even Keynes (1921) and Jeffreys (1939) find fault with the rule of succession.

We have to confess our inability to see anything at all absurd about the rule of succession. We recommend very strongly that you do a little independent literature searching, and read some of the objections various writers have to it. You will see that in every case the same thing has happened. Firstly, Laplace was quoted out of context, and secondly, in order to demonstrate the absurdity of the rule of succession, the author applies it to a case where it does not apply, because there is additional prior information which the rule of succession does not take into account.

But if you go back and read Laplace (1812) himself, you will see that in the very next sentence after this sunrise episode, he warns the reader against just this misunderstanding:

> But this number is far greater for him who, seeing in the totality of phenomena the principle regulating the days and seasons, realizes that nothing at the present moment can arrest the course of it.

In this somewhat awkward phraseology he is pointing out to the reader that the rule of succession gives the probability based *only* on the information that the event occurred n times in N trials, and that our knowledge of celestial mechanics represents a great deal of additional information. Of course, if you have additional information beyond the numbers n and N, then you ought to take it into account. You are then considering a different problem, the rule of succession no longer applies, and you can reach an entirely different answer. Probability theory gives the results of consistent plausible reasoning on the basis of the information *which was put into it*.

It has to be admitted that, in mentioning the sunrise at all, Laplace made a very unfortunate choice of an example – because the rule of succession does not really apply to the sunrise, for just the reason that he points out. This choice has had a catastrophic effect on Laplace's reputation ever since. His statements make sense when the reader interprets 'probability', as Laplace did, as a means of representing a state of partial knowledge. But to those who thought of probability as a real physical phenomenon, existing independently of human knowledge, Laplace's position was quite incomprehensible; and so they jumped to the

[3] Some passages in the Bible led early theologians to conclude that the age of the world is about 5000 years. It seems that Laplace at first accepted this figure, as did everyone else. But it was during Laplace's lifetime that dinosaur remains were found almost under his feet (under the streets of Montmartre in Paris), and interpreted correctly by the anatomist Cuvier. Had he written this near the end of his life, we think that Laplace would have used a figure vastly greater than 5000 years.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

conclusion that Laplace had committed a ludicrous error, without even bothering to read his full statement.

Here are some famous examples of the kind of objections to the rule of succession which may be found in the literature.

(1) Suppose the solidification of hydrogen to have been once accomplished. According to the rule of succession, the probability that it will solidify again if the experiment is repeated is $2/3$. This does not in the least represent the state of belief of any scientist.

(2) A boy is 10 years old today. According to the rule of succession, he has the probability $11/12$ of living one more year. The boy's grandfather is 70; according to this rule he has the probability $71/72$ of living one more year. The rule violates qualitative common sense!

(3) Consider the case $N = n = 0$. It then says that any conjecture without verification has the probability $1/2$. Thus there is probability $1/2$ that there are exactly 137 elephants on Mars. Also there is probability $1/2$ that there are 138 elephants on Mars. Therefore, it is certain that there are at least 137 elephants on Mars. But the rule says also that there is probability $1/2$ that there are *no* elephants on Mars. The rule is logically self-contradictory!

The trouble with examples (1) and (2) is obvious in view of our earlier remarks; in each case, highly relevant prior information, known to all of us, was simply ignored, producing a flagrant misuse of the rule of succession. But let's look a little more closely at example (3). Wasn't the rule applied correctly here? We certainly can't claim that we had prior information about elephants on Mars which was ignored. Evidently, if the rule of succession is to survive example (3), there must be some very basic points about the use of probability theory which we need to emphasize.

Now, what do we mean when we say that there is 'no evidence' for a proposition? The question is not what you or I might mean colloquially by such a statement. The question is: *What does it mean to the robot*? What does it mean in terms of probability theory?

The prior information we used in derivation of the rule of succession was that the robot is told that there are only two possibilities: A is true, or A is false. Its entire 'universe of discourse' consists of only two propositions. In the case $N = 0$, we could solve the problem also by direct application of the principle of indifference, and this will of course give the same answer $P(A|X) = 1/2$, that we obtained from the rule of succession. But, just by noting this, we see what is wrong. Merely by admitting the possibility of one of three different propositions being true, instead of only one of two, we have already specified prior information different from that used in deriving the rule of succession.[4]

If the robot is told to consider 137 different ways in which A could be false, and only one way in which it could be true, and is given no other information, then its prior probability for A is $1/138$, not $1/2$. So, we see that the example of elephants on Mars was, again, a gross misapplication of the rule of succession.

[4] We see here only what should have been obvious: that our conclusions from some data can depend on the size of our hypothesis space. We saw a very similar thing in our study of the marginalization paradox in Chapter 15, in the discussion following Eq. (15.92), where we found that the size of a parameter space can affect our inferences. That is, introducing a new parameter can make a difference in our conclusions, even when we have no knowledge of its numerical value.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

Moral

Probability theory, like any other mathematical theory, cannot give a definite answer unless we ask it a definite question. We should always start a problem with an explicit enumeration of the 'hypothesis space' consisting of the different propositions that we are going to consider in that problem. That is part of the 'boundary conditions' which must be specified before we have a well-posed mathematical problem. If we say, 'I don't know what the possible propositions are', that is mathematically equivalent to saying, 'I don't know what problem I want to solve'. The only answer the robot can give is: 'Come back and ask me again when you do know'.

18.7 Jeffreys' objection

As one would expect, the example used by Jeffreys (1939, p. 107) is more subtle. He writes:

> I may have seen one in 1000 of the 'animals in feathers' in England; on Laplace's theory the probability of the propositions 'all animals with feathers have beaks' would be about $1/1000$. This does not correspond to my state of belief, or anybody else's.

Now, while we agree with everything Jeffreys said, we must point out that he failed to add two important facts. Firstly, it is true that, on this evidence, $P(\text{all have beaks}) \approx 1/1000$ according to Laplace's rule. But also $P(\text{all but one have beaks}) \approx 1/1000$, $P(\text{all but two have beaks}) \approx 1/1000, \ldots,$ etc. More specifically, if there are N feathered animals of which we have seen r (all with beaks), then rewriting (18.24) in this notation we see that $P(\text{all have beaks}) = P_0 = (r+1)/(N+1) \approx 1/1000$, while P (all but n have beaks) is

$$P_n = P_0 \frac{(N-r)!\,(N-n)!}{N!\,(N-n-r)!}, \tag{18.26}$$

and the probability that there are n_0 or more without beaks is

$$\sum_{n=n_0}^{N} P_n = \frac{(N-r)!\,(N-n_0+1)!}{(N+1)!\,(N-n_0-r)!} \approx \exp\{-rn_0/N\}. \tag{18.27}$$

Thus if there are one million animals with feathers, of which we have seen 1000 (all with beaks), this leaves it an even bet that there are at least $1000 \ln(2) = 693$ without beaks; and, of course, an even bet that the number is less than that. If the only relevant information one had was the aforementioned observation, we think that this *would* be just the proper and reasonable inference.

Secondly, Laplace's rule is not appropriate for this problem because we all have additional prior information that it does not take into account: hereditary stability of form, the fact that a beakless feathered animal would, if it existed, be such an interesting curiosity that we all should have heard of it even if we had not seen it (as has happened in the converse case of the duck-billed platypus), etc. To see fairly and in detail what Laplace's rule (18.24) says, we need to consider a problem where our prior information corresponds better to that supposed in its derivation.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

18.8 Bass or carp?

A guide of unquestioned knowledge and veracity assures us that a certain lake contains only two species of fish: bass and carp. We catch ten and find them all to be carp – what is then our state of belief about the percentage of bass? Common sense tells us that, if the fish population were more than about 10% bass, then in ten catches we had a reasonably good chance of finding one; so our state of belief drops off rapidly above 10%. On the other hand, these data D provide no evidence against the hypothesis that the bass population is zero. So common sense without any calculation would lead us to conclude that the bass population is quite likely to be in the range, say, (0%, 15%), but intuition does not tell us quantitatively how likely this is.

What, then, does Laplace's rule say? Denoting the bass fraction by f, its posterior cumulative pdf is $P(f < f_0|DX) = 1 - (1 - f_0)^{11}$. Thus we have a probability of $1 - (1 - 0.15)^{11} = 0.833$, or odds of 5:1, that the bass population is indeed below 15%. Likewise, the data yield a probability of $2/3$, or odds of 2:1, that the lake contains less than 9.5% bass, and odds of 10:1 that it is less than 19.6%, while the posterior median value is

$$f_{1/2} = 1 - \left(\frac{1}{2}\right)^{1/11} = 0.061, \tag{18.28}$$

or 6.1%; it is an even bet that the bass population is less than this. The interquartile range is $(f_{1/4},\ f_{3/4}) = (2.6\%,\ 11.8\%)$; it is as likely to be within as outside that interval. The 'best' estimate of f by the criterion of minimum mean-square error is Laplace's posterior mean value (18.25): $\langle f \rangle = 1/12$, or 8.3%.

Suppose now that our 11th catch is a bass; how does this change our state of belief? Evidently, we shall revise our estimate of f upward, because the data now *do* provide evidence against the hypothesis that f is very small. Indeed, if the bass population were less than 5%, then we would be unlikely to find one in only 11 catches, so our state of belief drops off rapidly below 5%, but less rapidly than before above 10%.

Laplace's rule agrees, now saying that the best mean-square estimate is $\langle f \rangle = 2/13$, or 15.4%, and the posterior density is $P(\mathrm{d}f|DX) = 132 f(1 - f)^{10}\mathrm{d}f$. This yields a median value of 13.6%, raised very considerably because the new datum has effectively eliminated the possibility that the bass population might be below about 3%, which was just the most likely region before. The interquartile range is now (8.3%, 20.9%).

It appears to us that all these numbers correspond excellently to our common sense judgments. This, then, is the kind of problem to which Laplace's rule applies very realistically; i.e. there were known to be only two possibilities at each trial, and our prior knowledge gave no other information beyond assuring us that both were possible. Whenever the result of Laplace's rule of succession conflicts with our intuitive state of belief, we suggest that the reason is that our common sense is making use of additional prior information about the real world situation that is not used in the derivation of the rule of succession.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

18.9 So where does this leave the rule?

Mathematically, the rule of succession is the solution to a certain problem of inference, defined by the prior probability and the data. The 200 year old hangup has been over the question: *what* prior information is being described by the uniform prior probability (18.3)? Laplace was not too clear about this – his discussion of it seemed to invoke the idea of a 'probability of a probability' which may appear to be metaphysical nonsense until one has the notion of an inner and outer robot – but his critics, instead of being constructive and trying to define the conceptual problem more clearly, seized upon this to denounce Laplace's whole approach to probability theory.

Of Laplace's critics, only Jeffreys (1939) and Fisher (1956) seem to have thought it through deeply enough to realize that the unclear definition of the prior information was the source of the difficulty; the others, following the example of Venn (1866), merely produce examples where common sense and Laplace's rule are in conflict, and, without making any attempt to understand the reason for it, reject the rule in any and all circumstances. As we noted in Chapter 16, Venn's criticisms were so unjust that even Fisher (1956) was impelled to come to Laplace's defense on this issue.

In this connection we have to remember that probability theory never solves problems of actual practice, because all such problems are infinitely complicated. We solve only idealizations of the real problem, and the solution is useful to the extent that the idealization is a good one. In the example of the solidification of hydrogen, the prior information, which our common sense uses so easily, is actually so complicated that nobody knows how to convert it into a prior probability assignment. There is no reason to doubt that probability theory is, in principle, competent to deal with such problems; but we have not yet learned how to translate them into mathematical language without oversimplifying rather drastically.

In summary, Laplace's rule of succession provides a definite, useful solution to a definite, real problem. Everybody denounces it as nonsense because it is not also the solution to some different problem. The case where the problem can be reasonably idealized to one with only two hypotheses to be considered, a belief in a constant 'causal mechanism', *and no other prior information*, is the only case where it applies. But we can, of course, generalize it to any number of hypotheses, as follows.

18.10 Generalization

We give the derivation in full detail, to present a mathematical technique of Laplace that is useful in many other problems. There are K different hypotheses, $\{A_1, A_2, \ldots, A_K\}$, a belief that the 'causal mechanism' is constant, and no other prior information. We perform a random experiment N times, and observe A_1 true n_1 times, A_2 true n_2 times, etc. Of course, $\sum_i n_i = N$. On the basis of this evidence, what is the probability that in the next $M = \sum_i m_i$ repetitions of the experiment, A_i will be true exactly m_i times? To find the probability $P(m_1 \cdots m_K|n_1, \ldots, n_K)$ that answers this, define the prior knowledge by a K-dimensional uniform prior A_p density:

$$(A_{p_1} \cdots A_{p_K}|X) = C\delta(p_1 + \cdots + p_K - 1), \qquad p_i \geq 0. \tag{18.29}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

To find the normalization constant C, we set

$$\int_0^\infty \mathrm{d}p_1 \cdots \mathrm{d}p_K \,(A_{p_1} \cdots A_{p_k}|X) = 1 = CI(1), \tag{18.30}$$

where

$$I(r) \equiv \int_0^\infty \mathrm{d}p_1 \cdots \mathrm{d}p_k \,\delta(p_1 + \cdots + p_K - r). \tag{18.31}$$

Direct evaluation of this would be rather messy, because all integrations after the first would be between limits that need to be worked out; so let's use the following trick. Firstly, take the Laplace transform of (18.31):

$$\int_0^\infty \mathrm{d}r \exp\{-\alpha r\}\, I(r) = \int_0^\infty \mathrm{d}p_1 \cdots \mathrm{d}p_K \;\exp\{-\alpha(p_1 + \cdots + p_K)\} = \frac{1}{\alpha^K}. \tag{18.32}$$

Then, inverting the Laplace transform by Cauchy's theorem,

$$\begin{aligned} I(r) &= \frac{1}{2\pi\mathrm{i}} \int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \mathrm{d}\alpha \,\frac{\exp\{\alpha r\}}{\alpha^K} \\ &= \frac{1}{(K-1)!} \frac{\mathrm{d}^{K-1}}{\mathrm{d}\alpha^{K-1}} \exp\{\alpha r\}\Big|_{\alpha=0} \\ &= \frac{r^{K-1}}{(K-1)!}, \end{aligned} \tag{18.33}$$

where, according to the standard theory of Laplace transforms, the path of integration passes to the right of the origin, and is closed by an infinite semicircle over the left half-plane, the integral over which is zero. Thus,

$$C = \frac{1}{I(1)} = (K-1)!\,. \tag{18.34}$$

By this device, we avoided having to consider complicated details about different ranges of integration over the different p_i, that would come up if we tried to evaluate (18.31) directly. The prior $P(n_1 \cdots n_K|X)$ is then, using the same trick,

$$\begin{aligned} P(n_1 \cdots n_K|X) &= \frac{N!}{n_1! \ldots n_K!} \int_0^\infty \mathrm{d}p_1 \cdots \int_0^\infty \mathrm{d}p_K \; p_1^{n_1} \cdots p_K^{n_K} (A_{p_1} \cdots A_{p_K}|X) \\ &= \frac{N!\,(K-1)!}{n_1! \cdots n_K!} \,J(1), \end{aligned} \tag{18.35}$$

where

$$J(r) \equiv \int_0^\infty \mathrm{d}p_1 \cdots \mathrm{d}p_K \; p_1^{n_1} \cdots p_K^{n_K} \,\delta(p_1 + \cdots + p_k - r), \tag{18.36}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

which we evaluate as before by taking the Laplace transform:

$$\begin{aligned}\int_0^\infty \mathrm{d}r\, \mathrm{e}^{-\alpha r} J(r) &= \int_0^\infty \mathrm{d}p_1 \cdots \mathrm{d}p_K\, p_1^{n_1} \cdots p_K^{n_K}\, \exp\{-\alpha(p_1 + \cdots + p_K)\} \\ &= \prod_{i=1}^{K} \frac{n_i\,!}{\alpha^{n_i+1}}.\end{aligned} \tag{18.37}$$

So, as in (18.33), we have

$$J(r) = \frac{n_1! \cdots n_K!}{2\pi \mathrm{i}} \int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \mathrm{d}\alpha\, \frac{\exp\{\alpha r\}}{\alpha^{N+K}} = \frac{n_1! \cdots n_K!}{(N+K-1)!}\, r^{N+K-1} \tag{18.38}$$

and

$$P(n_1 \cdots n_k|X) = \frac{N!\,(K-1)!}{(N+K-1)!}, \quad n_i \geq 0, \quad n_1 + \cdots + n_K = N. \tag{18.39}$$

Therefore, by Bayes' theorem

$$\begin{aligned}(A_{p_1} \cdots A_{p_K}|n_1 \cdots n_K) &= (A_{p_1} \cdots A_{p_K}|X) \frac{P(n_1 \cdots n_K|A_{p_1} \cdots A_{p_K})}{P(n_1 \cdots n_K|X)} \\ &= \frac{(N+K-1)!}{n_1! \cdots n_K!}\, p_1^{n_1} \cdots p_K^{n_K}\, \delta(p_1 + \cdots + p_K - 1),\end{aligned} \tag{18.40}$$

and finally

$$\begin{aligned}&P(m_1 \cdots m_K|n_1 \cdots n_K) \\ &\quad = \int_0^\infty \mathrm{d}p_1 \cdots \mathrm{d}p_K\, P(m_1 \cdots m_K|A_{p_1} \cdots A_{p_K})(A_{p_1} \cdots A_{p_K}|n_1 \cdots n_K) \\ &\quad = \frac{M!}{m_1! \cdots m_K!}\, \frac{(N+K-1)!}{n_1! \cdots n_K!} \int_0^\infty \mathrm{d}p_1 \cdots \mathrm{d}p_K\, p_1^{n_1+m_1} \cdots p_K^{n_K+m_K} \\ &\qquad \times \delta(p_1 + \cdots + p_K - 1).\end{aligned} \tag{18.41}$$

The integral is the same as $J(1)$ except for the replacement $n_i \to n_i + m_i$. So, from (18.38),

$$P(m_1 \cdots m_K|n_1 \cdots n_K) = \frac{M!}{m_1! \cdots m_K!}\, \frac{(N+K-1)!}{n_1! \cdots n_K!}\, \frac{(n_1+m_1)! \cdots (n_K+m_K)!}{(N+M+K-1)!} \tag{18.42}$$

or, reorganizing into binomial coefficients, the generalization of (18.24) is

$$P(m_1 \cdots m_K|n_1 \cdots n_K) = \frac{\binom{n_1+m_1}{n_1} \cdots \binom{n_K+m_K}{n_K}}{\binom{N+M+K-1}{M}}. \tag{18.43}$$

In the case where we want just the probability that A_1 will be true on the next trial, we need this formula with $M = m_1 = 1$, all other $m_i = 0$. The result is the generalized rule

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

of succession:

$$P(A_1|n_1NK) = \frac{n_1 + 1}{N + K}. \tag{18.44}$$

We see that, in the case $N = n_1 = 0$, this reduces to the answer provided by the principle of indifference, which it therefore contains as a special case. If K is a power of 2, this is the same as a method of inductive reasoning proposed by R. Carnap (1942), which he denotes $c^*(h, e)$ in his *Continuum of Inductive Methods*.

Use of the rule of succession in cases where N is very small is rather foolish, of course. Not really wrong; just foolish. Because, if we have no prior evidence about A, and we make such a small number of observations that we have practically no evidence, well, that's just not a very promising basis on which to do plausible reasoning. We can't expect to get anything useful out of it. We do, of course, obtain definite numerical values for the probabilities, but these values are very 'soft,' i.e. very unstable, because the A_p distribution is still very broad for small N. Our common sense tells us that the evidence N_n for small N provides no reliable basis for further predictions, and we'll see that this conclusion also follows as a consequence of the theory we are developing here.

The real reason for introducing the rule of succession lies in the cases where we *do* obtain a significant amount of information from the experiment; i.e. when N is a large number. In this case, fortunately, we can pretty much forget about these fine points concerning prior evidence. The particular initial assignment $(A_p|X)$ will no longer have much influence on the results, for the same reason as in the particle-counter problem of Chapter 6. This remains true for the generalized case leading to (18.43). You see from (18.44) that, as soon as the number of observations N is large compared with the number of hypotheses K, then the probability assigned to any particular hypothesis depends, for all practical purposes, only on what we have observed, and not on how many prior hypotheses there are. If you contemplate this for ten seconds, your common sense will tell you that the criterion $N \gg K$ is exactly the right one for this to be so.

In the literature starting with Venn (1866), those who issued polemical denunciations of Laplace's rule of succession have put themselves in an incredible situation. How is it possible for one human mind to reject Laplace's rule – and then advocate a frequency definition of probability? Anyone who assigns a probability to an event equal to its observed frequency in many trials is doing just what Laplace's rule tells him to do! The generalized rule (18.44) supplies an obviously needed refinement of this, small correction terms when the number of observations is not large compared with the number of propositions.

18.11 Confirmation and weight of evidence

A few new ideas – or rather, connections with familiar old ideas – are suggested by our calculations involving A_p. Although we shall not make any particular use of them, it seems worthwhile to point them out. We saw that the stability of a probability assignment in the face of new evidence is essentially determined by the width of the A_p distribution. If E is

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

prior evidence and F is new evidence, then

$$P(A|EF) = \int_0^1 \mathrm{d}p\, p\, (A_p|EF) = \frac{\int_0^1 \mathrm{d}p\, p\, (A_p|F)(A_p|E)}{\int_0^1 \mathrm{d}p\, (A_p|F)(A_p|E)}. \tag{18.45}$$

We might say that F is *compatible* with E, as far as A is concerned, if having the new evidence, F, doesn't make any appreciable change in the probability of A;

$$P(A|EF) = P(A|E). \tag{18.46}$$

The new evidence can make an enormous change in the distribution of A_p without changing the first moment. It might sharpen it up very much, or broaden it. We could become either more certain or more uncertain about A, but if F doesn't change the center of gravity of the A_p distribution, we still end up assigning the same probability to A.

Now, the stronger property: the new evidence F *confirms* the previous probability assignment, if F is compatible with it, and at the same time, gives us more confidence in it. In other words, we exclude one of these possibilities, and with new evidence F the A_p distribution narrows. Suppose F consists of performing some random experiment and observing the frequency with which A is true. In this case $F = N_n$, and our previous result, Eq. (18.22), gives

$$(A_p|N_n) = \frac{(N+1)!}{n!(N-n)!} p^n (1-p)^{N-n} \approx (\text{constant}) \cdot \exp\left\{-\frac{(p-f)^2}{2\sigma^2}\right\}, \tag{18.47}$$

where

$$\sigma^2 = \frac{f(1-f)}{n}, \tag{18.48}$$

and $f = (n/N)$ is the observed frequency of A. The approximation is found by expanding $\log(A_p|N_p)$ in a Taylor series about its peak value, and is valid when $n \gg 1$ and $(N-n) \gg 1$. If these conditions are satisfied, then $(A_p|N_n)$ is very nearly symmetric about its peak value. Then, if the observed frequency f is close to the prior probability $P(A|E)$, the new evidence N_n will not affect the first moment of the A_p distribution, but will sharpen it up, and that will constitute a confirmation as we have defined it.

This shows one more connection between probability and frequency. We defined the 'confirmation' of a probability assignment according to entirely different ideas than are usually used to define it. We define it in a way that agrees with our intuitive notation of confirmation of a previous state of mind. But it turned out that the *same* experimental evidence would constitute confirmation on either the frequency theory or our theory.

Now, from this we can see another useful notion, which we will call weight of evidence. Consider A_p, given two different pieces of evidence, E and F,

$$(A_p|EF) = (\text{constant}) \times (A_p|E)\,(A_p|F). \tag{18.49}$$

If the distribution $(A_p|F)$ was very much sharper than the distribution $(A_p|E)$, then the product of the two would still have a peak at practically the value determined by F. In this case, we would say intuitively that the evidence F carries much greater 'weight' than the evidence E. If we have F, it doesn't really matter much whether we take E into account or

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

not. On the other hand, if we don't have F, then whatever evidence E may represent will be extremely significant, because it will represent the best we are able to do. So, acquiring one piece of evidence which carries a great amount of weight can make it, for all practical purposes, unnecessary to continue keeping track of other pieces of evidence which carry only a small weight.

Of course, this is the way our minds operate. When we receive one very significant piece of evidence, we no longer pay so much attention to vague evidence. In so doing, we are not being very inconsistent, because it wouldn't make much difference anyway. So, our intuitive notion of weight of evidence is bound up with the sharpness of the A_p distribution. Evidence concerning A that we consider very significant is not necessarily evidence that makes a big change in the probability of A. It is evidence that makes a big change in our density for A_p. Seeing this, we can gain a little more insight into the principle of indifference and also make contact between this theory and Carnap's methods of inductive reasoning.

18.11.1 Is indifference based on knowledge or ignorance?

Before we can use the principle of indifference to assign numerical values of probabilities, there are two different conditions that must be satisfied: (1) we must be able to analyze the situation into mutually exclusive, exhaustive possibilities; (2) having done this, we must then find the available information gives us no reason to prefer any of the possibilities to any other. In practice, these conditions are hardly ever met unless there's some evident element of symmetry in the problem. But there are two entirely different ways in which condition (2) might be satisfied. It might be satisfied as a result of ignorance, or it might be satisfied as a result of positive knowledge about the situation. To illustrate this, let's suppose that a person who is known to be very dishonest is going to toss a coin, and that there are two people watching him. Mr A is allowed to examine the coin. He has all the facilities of the National Bureau of Standards at his disposal. He performs hundreds of experiments with scales and calipers, magnetometers and microscopes, X-rays and neutron beams, and so on. Finally, he is convinced that the coin *is* perfectly honest. Mr B is not allowed to do this. All he knows is that a coin is being tossed by a shady character. He suspects the coin is biased, but he has no idea in which direction. Condition (2) is satisfied equally well for both of them. Each would start out by assigning probability one-half to each face. The same probability assignment can describe a condition of complete ignorance or a condition of very great knowledge. This has seemed paradoxical for a long time. Why doesn't Mr A's extra knowledge make any difference? Well, of course, it *does* make a difference. It makes a very important difference, but one that doesn't show up until we start performing this experiment. The difference is not in the probability for A, but in the density for A_p.

Suppose the first toss is heads. To Mr B, that constitutes evidence that the coin is biased to favor heads. And so, on the next toss, he would assign new probabilities to take that into account. But to Mr A, the evidence that the coin is honest carries overwhelmingly greater weight than the evidence of one throw, and he'll continue to assign a probability of $1/2$.

You see what's going to happen. To Mr B, every toss of the coin represents new evidence about its bias. Every time it's tossed, he will revise his assignment for the next toss; but,

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

after several tosses, his assignment will get more and more stable, and in the limit $n \to \infty$ they will tend to the observed frequency of heads. To observer A, the prior evidence of symmetry continues to carry greater weight than the evidence of almost any number of throws, and he persists in assigning the probability $1/2$. Each has done consistent plausible reasoning on the basis of the information available to him, and our theory accounts for the behavior of each.

If you assumed that Mr A had perfect knowledge of symmetry, you might conclude that his A_p distribution is a δ-function. In that case, his mind could never be changed by any amount of new data. Of course, that's a limiting case that's never reached in practice. Not even the Bureau of Standards can give us evidence that good.

18.12 Carnap's inductive methods

The philosopher Rudolph Carnap (1952) gives an infinite family of possible 'inductive methods' by which one can convert prior information and frequency data into a probability assignment and an estimate of frequencies for this future. His *ad hoc* principle (that is, a principle that is found from intuition rather than from the rules of probability theory) is that the final probability assignment $P(A|N_n X)$ should be a weighted average of the prior probability $P(A|X)$ and the observed frequency, $f = n/N$. Assigning a weight N to the 'empirical factor' f, and an arbitrary weight λ to the 'logical factor' $P(A|X)$ leads to the method which Carnap denotes by $c_\lambda(h, e)$. Introduction of the A_p distribution accounts for this in more detail; the theory developed here includes all of Carnap's methods as special cases corresponding to different prior densities $(A_p|X)$, and leads us to reinterpret λ as the weight of prior evidence. Thus, in the case of two hypotheses, the Carnap λ method is the one you can calculate from the prior density $(A_p|X) = (\text{constant}) \cdot [p(1-p)]^r$, with $2r = \lambda - 2$. The result is

$$P(A|N_n X) = \frac{2n + \lambda}{2N + 2\lambda} = \frac{(n + r) + 1}{(N + 2r) + 2}. \tag{18.50}$$

Greater λ thus corresponds to a more sharply peaked $(A_p|X)$ density.

In our coin-tossing example, Mr A from the Bureau of Standards reasons according to a Carnap method with λ of the order of, perhaps, thousands; while Mr B, with much less prior knowledge about the coin, would use a λ of perhaps 5 or 6. (The case $\lambda = 2$, which gives Laplace's rule of succession, is much too broad to be realistic for coin tossing; for Mr B surely knows that the center of gravity of a coin can't be moved by more than half its thickness from the geometrical center. Actually, as we saw in Chapter 10, this analysis isn't always applicable to tossing of real coins, for reasons having to do with the laws of physics.)

From the second way we wrote Eq. (18.50), we see that the Carnap λ method corresponds to a weight of prior evidence which would be given by $(\lambda - 2)$ trials, in exactly half of which A was observed to be true. Can we understand why the weighting of prior evidence is $\lambda =$ (number of prior trials $+2$), while that of the new evidence N_p is only (number of new trials) $= N$? Well, look at it this way. The appearance of the $(+2)$ is the robot's way of

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

telling us this: prior knowledge that is *possible* for A to be either true or false is equivalent to knowledge that A has been true at least once, and false at least once. This is hardly a derivation, but it makes reasonably good sense.

Let's pursue this line of reasoning a step further. We started with the statement X: it is *possible* for A to be either true or false at any trial. But that is still a somewhat vague statement. Suppose we interpret it as meaning that A has been observed true exactly once, and false exactly once. If we grant that this state of knowledge is correctly described by Laplace's assignment $(A_p|X) = 1$, then *what was the 'pre-prior' state of knowledge X_0 before we had the data X?* To answer this, we need only to apply Bayes' theorem backwards, as we did in the method of imaginary results in Chapter 5 and in urn sampling in Chapter 6. The result is: our 'pre-prior' A_p distribution must have been

$$(A_p|X_0)\,\mathrm{d}p = (\text{constant})\frac{\mathrm{d}p}{p(1-p)}. \tag{18.51}$$

This is just the quasi-distribution representing 'complete ignorance', or the 'basic measure' of our parameter space, that we found by transformation groups in Chapter 12 and which Haldane (1932) had suggested long ago. So, here is another line of thought that could have led us to this measure. By the same line of thought we found the discrete version of (18.51) already in Chapter 6, Eq. (6.49).

It appears, then, that if we have definite prior evidence that it *is* possible for A to be either true or false on any one trial, then Laplace's rule $(A_p|X) = 1$ is the appropriate one to use. But if initially we are so completely uncertain that we're not even sure whether it is *possible* for A to be true on some trials and false on others, then we should use the prior (18.51).

How different are the numerical results which the pre-prior assignment (18.51) gives us? Repeating the derivation of (18.22) with this pre-prior assignment, we find that, provided n is not zero or N,

$$(A_p|N_nX_0) = \frac{(N-1)!}{(n-1)!\,(N-n-1)!}p^{n-1}(1-p)^{N-n-1}, \tag{18.52}$$

which leads, instead of to Laplace's rule of succession, to the mean-value estimate of p:

$$P(A|N_nX_0) = \int_0^1 \mathrm{d}p\, p\,(A_p|N_n) = \frac{n}{N}, \tag{18.53}$$

equal to the observed frequency, and identical with the maximum likelihood estimate of p. Likewise, provided $0 < n < N$, we find instead of (18.24) the formula

$$P(M_m|N_nX_0) = \frac{\binom{m+n-1}{m}\binom{M-m+N-n-1}{M-m}}{\binom{N+M-1}{M}}. \tag{18.54}$$

All of these results correspond to having observed one less success and one less failure.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

18.13 Probability and frequency in exchangeable sequences

We are now in a position to say quite a bit more about connections between probability and frequency. There are two main types of connections: (a) given an observed frequency in a random experiment, convert this information into a probability assignment; and (b) given a probability assignment, predict the frequency with which some condition will be realized. We have seen, in Chapters 11 and 12, how the principles of maximum entropy and transformation groups lead to probability assignments which, if the quantity of interest happens to be the result of some 'random experiment', correspond automatically to predicted frequencies, and thus solve problem (b) in some situations.

The rule of succession gives us the solution to problem (a) in a wide class of problems. If we have observed whether A was true in a very large number of trials, *and the only knowledge we have about A is the result of this random experiment, and the consistency of the 'causal mechanism',* then it says that the probability we should assign to A at the next trial becomes practically equal to the observed frequency. Now, in fact, this is exactly what people who define probability in terms of frequency do: one postulates the existance of an unknown 'absolute' probability, whose numerical value is to be found by performing random experiments. Of course, you must perform a very large number of experiments. Then the observed frequency of A is taken as the estimate of the probability. As we saw earlier in this chapter, even the $+1$ and $+2$ in Laplace's formula turn up when the 'frequentist' refines his methods by taking the center of a confidence interval. So, I don't see how even the most ardent advocates of the frequency theory of probability can damn the rule of succession without thereby damning his own procedures; after all polemics, there remains the simple fact that in his own procedures, he is doing exactly what Laplace's rule of succession tells him to do. Indeed, to define probability in terms of frequency is equivalent to saying that the rule of succession is the *only* rule which can be used for converting observational data into probability assignments.

18.14 Prediction of frequencies

Now let's consider problem (b) in this situation: to reason from a probability to a frequency. This is simply a problem of parameter estimation, not different in principle from any other. Suppose that instead of asking for the probability that A will be true in the next trial, we wish to infer something about the relative frequency of A in an indefinitely large number of trials, on the basis of the evidence N_n. We must take the limit of (18.24) as $M \to \infty$, $m \to \infty$, in such a way that $(m/M) \to f$. Introducing the proposition

$$A_f = \text{the frequency of } A \text{ true in the indefinitely large number of trials is } f, \tag{18.55}$$

we find in the limit that the probability density of A_f, given N_n, is

$$P(A_f|N_n) = \frac{(N+1)!}{N!(N-n)!} f^n (1-f)^{N-n}, \tag{18.56}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

which is the same as our $(A_p|N_n)$ in (18.22), with f numerically equal to p. According to (18.55), the most probable frequency is equal to (n/N), the observed frequency in the past. But we have noted before that in parameter estimation (if you object to my calling f a 'parameter', then let's just call it 'prediction'), the most probable value is usually a poorer estimate than the mean value in the small sample case, where they can be appreciably different. The mean-value estimate of the frequency is

$$\overline{f} = \int_0^1 \mathrm{d}f f P(A_f|N_n) = \frac{n+1}{N+2}, \tag{18.57}$$

i.e. just the same as the value of $P(A|N_n)$, (18.25), given by Laplace's rule of succession. Thus, we can interpret the rule in either way; *the probability which Laplace's theory assigns to A at a single trial is numerically equal to the estimate of frequency which minimizes the expected square of the error.* You see how nicely this corresponds with the relationship between probability and frequency which we found in the maximum entropy and transformation group arguments.

Note also that the distribution $P(A_f|N_n)$ is quite broad for small N, confirming our expectation that no reliable predictions should be possible in this case. As a numerical example, if A has been observed true once in two trials, then $\overline{f} = P(A|N_n) = 1/2$, but according to (18.55) it is still an even bet that the true frequency f lies outside the interval $0.326 < f < 0.674$. With no evidence at all ($N = n = 0$), it would be an even bet that f lies outside the interval $0.25 < f < 0.75$. More generally, the variance of (18.55) is

$$\mathrm{var} P(A_f|N_n) = \overline{f^2} - \overline{f}^2 = \frac{\overline{f}(1-\overline{f})}{N+3}, \tag{18.58}$$

so that the expected error in the estimate (18.56) decreases like $1/\sqrt{N}$. More detailed conclusions about the reliability of predictions, which we could make from (18.56), are, for all practical purposes, identical with those the statistician would make by the method of confidence intervals.

All these results hold also for the generalized rule of succession. Taking the limit of (18.43) as $M \to \infty$, $m_i/M \to f_i$, we find the joint probability density function for A_i to occur with frequency f_i to be

$$P(f_1 \cdots f_k|n_1 \cdots n_k) = \frac{(N+K)!}{n_1! \cdots n_k!}(f_1^{n_1} \cdots f_k^{n_k})\delta(f_1 + \cdots + f_k - 1). \tag{18.59}$$

The probability that the frequency f_1 will be in the range $\mathrm{d}f_1$ is found by integrating (18.59) over all values of $f_2, \ldots, f_k$ compatible with $f_1 \geq 0$, $(f_2 + \cdots + f_k) = 1 - f_1$. This can be carried out by application of Laplace transforms in a well-known way, and the result is

$$P(f_1|n_1 \cdots n_k) = \frac{(N+K-1)!}{n_1!(N-n_1+K-2)!} f_1^{n_1}(1-f_1)^{N-n_1+K-2}, \tag{18.60}$$

from which we find the most probable value to be

$$(\hat{f}_1) = \frac{n_1}{N+K-2} \tag{18.61}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

and the mean value to be

$$\overline{f_1} = \frac{n_1 + 1}{N + K}, \tag{18.62}$$

which is Laplace's rule of succession (18.44).

Another interesting result is found by taking the limit of $P(M_m|A_p)$ in (18.18) as $M \to \infty$, $(m/M) \to f$; we find

$$P(M_m|A_p) = \delta(f - p). \tag{18.63}$$

Likewise, taking the limit of $(A_p|N_n)$ in (18.22) as $N \to \infty$, we find

$$(A_p|A_f) = \delta(f - p), \tag{18.64}$$

which follows from (18.63) by application of Bayes' theorem. Therefore, if B is any proposition, we have, from our standard argument,

$$\begin{aligned} P(B|A_f) &= \int_0^1 \mathrm{d}p\,(BA_p|A_f) = \int_0^1 \mathrm{d}p\; P(B|A_pA_f)(A_p|A_f) \\ &= \int_0^1 \mathrm{d}p\; P(B|A_p)\delta(p - f). \end{aligned} \tag{18.65}$$

In the last step we used the property (18.1) that A_p automatically neutralizes any other statement about A. Thus, if f and p are numerically equal, we have $P(B|A_p) = P(B|A_p)$; A_p and A_f are *equivalent* statements in their implication for plausible reasoning.

To verify this equivalence in one case, note that in the limit $N \to \infty$, $(n/N) \to f$, $P(M_m|N_n)$ in Eq. (18.24) reduces to the binomial distribution $P(M_m|A_p)$ as given by (18.18). The generalized formula (18.43), in the corresponding limit, goes into the multinomial distribution,

$$P(m_1 \cdots m_k|f_1 \cdots f_k) = \frac{m!}{m_1! \cdots m_k!} f_1^{m_1} \cdots f_k^{m_k}. \tag{18.66}$$

This equivalence shows why it is so easy to confuse the notion of probability and frequency, and why in many problems this confusion does no harm. Whenever the available information consists of observed frequencies in a large sample, and constancy of the 'causal mechanism', Laplace's theory becomes mathematically equivalent to the frequency theory. Most of the 'classical' problems of statistics (life insurance, etc.) are of just this type; and as long as one works only on such problems, all is well. The harm arises when we consider more general problems.

Today, physics and engineering offer many important applications for probability theory in which there is an absolutely essential part of the evidence which cannot be stated in terms of frequencies, and/or the quantities about which we need plausible inference have nothing to do with frequencies. The axiom (probability) ≡ (frequency), if applied consistently, would prevent us from using probability theory in these problems.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

18.15 One-dimensional neutron multiplication

Our discussion so far has been rather abstract; perhaps too much so. In order to make amends for this, I would like to show you a specific physical problem where these equations apply. This was first described in a short note by Bellman, Kalaba and Wing (1957) and further developed in a more recent book by Wing (1962). Neutrons are traveling in fissionable material, and we want to estimate how many new neutrons will be produced in the long run as a consequence of one incident trigger neutron. In order to have a tractable mathematical problem, we make some drastic simplifying assumptions as follows.

(a) The neutrons travel only in the $\pm x$ direction at a constant velocity.
(b) Each time a neutron, traveling either to the right or the left, initiates a fission reaction, the result is exactly two neutrons, one traveling to the right, one to the left. The net result is therefore that any neutron will from time to time emit a progeny neutron traveling in the opposite direction.
(c) The progeny neutrons are immediately able to produce still more progeny in the same manner.

We fire a single trigger neutron into a thickness x of fissionable material from the left, and the problem is to predict the number of neutrons that will emerge from the left and from the right, over all time, as a consequence. At least, that is what we would *like* to calculate. But, of course, the number of emerging neutrons is not *determined* by any of the given data, and so the best we can do is to calculate the *probability* that exactly n neutrons will be transmitted or reflected. We want to make a detailed comparison of the Laplace theory and the frequency theory of probability, as applied to the initial formulation of this problem. We are concerned mainly with the underlying rationale by which we relate probability theory to the physical model.

Many proponents of the frequency theory berate the Laplace theory on purely philosophical grounds that have nothing to do with its success or failure in applications. There is a more defensible position, held by some, who recognize that the present state of affairs gives them no reason for smugness, and a good reason for caution. While they believe that at present the frequency theory is superior, they also say, as one of my correspondents did to me, ‘I will most cheerfully renounce the frequency thoery for any theory that yields me a better understanding and a more efficient formalism.’ The trouble is that the current statistical literature gives us no opportunity to see the Laplace theory in actual use so that valid comparisions could be made; and that is the situation we are trying to correct here.

18.15.1 The frequentist solution

Firstly, let us formulate the problem as it would be done using the frequency theory. Here is the way in which the ‘frequentist’ would reason:

The experimentalists have measured for us the relative frequency $p = a\Delta$ of fission in a very small thickness Δ of this material. This means that they have fired N trigger neutrons at a thin film of thickness Δ, and observed fission in n cases. Since N is finite, we cannot find the exact value of p from this, but it is approximately equal to the observed

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

frequency (n/M). More precisely, we can find confidence limits for p. In similar situations, we can expect about $k\%$ of the time, the limits (Cramér, 1946, p. 515)

$$\frac{N}{N+\lambda^2}\left[\frac{2n+\lambda^2}{2N}\pm\lambda\sqrt{\frac{n(N-n)}{N^3}+\frac{\lambda^2}{4N^2}}\right] \tag{18.67}$$

will include the true value of p, where λ is the $(100-k)\%$ value of a normal deviate. For example, with $\lambda=\sqrt{2}$, the range

$$\frac{n+1}{N+2}\pm\frac{N}{N+2}\sqrt{\frac{2n(N-n)}{N^3}+\frac{1}{N^2}}=\frac{n+1}{N+2}\pm\sqrt{\frac{2n(N-n)}{N^3}} \tag{18.68}$$

will cover the correct p in about 84% of similar cases. [Again, there's that +1 and +2 of Laplace's rule of succession!] In general, the connection between λ and k is given by

$$\frac{1}{\sqrt{2\pi}}\int_{-\lambda}^{\lambda}\mathrm{d}x\exp\left\{-\frac{x^2}{2}\right\}=\frac{k}{100}. \tag{18.69}$$

Equation (18.67) is an approximation valid when the numbers n and $(N-n)$ are sufficiently large; the exact confidence limits are difficult to express analytically, and for small N one should consult the graphs of E.S. Pearson and Clopper (1934). The number p is, of course, a definite, but imperfectly known, *physical constant* characteristic of the fissionable material.

Now, in order to calculate the relative frequency with which n neutrons will be reflected from a thickness x of this material, we have to make some additional assumptions. We assume that the probability of fission per unit length is always the same for each neutron independent of its history. Due to the complexity of the causes operating, it seems reasonable to assume this; but the real test of whether it is a valid assumption can come only from comparison of the final results of our calculation with experiment. This assumption means that the probabilities of fission in successive slabs of thickness Δ are independent, so that, for example, the probability that an incident neutron will undergo fission in the second slab of thickness Δ, but not in the first, is the product $p(1-p)$.

At this point, we turn to the mathematics and solve the problem by any one of several possible techniques, emerging with the relative frequencies $p_n(x)$, $q_n(x)$ for reflection or transmission of n neutrons, respectively. [Actually, the analytical solution has not yet been found, but Wing (1962) gives the results of numerical integration, which is equally good for our purposes.]

We now compare these predictions with experiment. When the first trigger neutron is fired into the thickness x, we observe r_1 neutrons reflected and t_1 neutrons transmitted. These data do not in any way affect the assignments $p_n(x)$ and $q_n(x)$, since the latter have no meaning in terms of a single experiment, but are predictions only of limiting frequencies for an indefinitely large number of experiments. We therefore must repeat the

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

experiment many times, and record the numbers r_i, t_i for each experiment. If we find that the frequency of cases for which $r_i = n$ tends sufficiently close to $p_n(x)$ ('sufficiently close' being determined by certain significance tests such as chi-squared), then we conclude that the theory is satisfactory; or at least that it is not rejected by the data. If, however, the observed frequencies show a wide departure from $p_n(x)$, then we know that there is something wrong with our initial set of assumptions.

Now, of course, the theory is either right or wrong. If it is wrong, then in principle the entire theory is demolished, and we have to start all over again, trying to find the right theory. In practice, it may happen that only one minor feature of the theory has to be changed, so that most of the old calculations will be useful in the new theory.

18.15.2 The Laplace solution

Now let's state this same problem in terms of Laplace's theory. We regard it simply as an exercise in plausible reasoning, in which we make the best possible guesses as to the outcome of a *single* experiment, or of a finite number of them. We are not concerned with the prediction, or even the existence, of limiting frequencies; because any assertion about the outcome of an impossible experiment is obviously an empty statement, and cannot be relevant to any applications. We reason as follows.

The experimentalists have provided us with the evidence N_n, by firing N neutrons at a thin film of thickness Δ, and observing fission in n cases. Since by hypothesis the only prior knowledge was that a neutron either will or will not undergo fission, we have just the situation where Laplace's rule of succession applies and the probability, on this evidence, of fission for the $(N+1)$th neutron in thickness Δ, is

$$p \equiv P(F_{N+1}|N_n) = \frac{n+1}{N+2}, \tag{18.70}$$

where

$$F_m \equiv \text{the } n\text{th neutron will undergo fission.} \tag{18.71}$$

Whether N is large or small, the question of the 'accuracy' of this probability does not arise – it is exact by definition. Of course, we will prefer to have as large a value of N as possible, since this increases the weight of the evidence N_n and makes the probability p not more *accurate*, but more *stable*. The probability p is manifestly *not* a physical property of the fissionable material, but is only a means of describing our state of knowledge about it, on the basis of the evidence N_n. If the preliminary experiment had yielded a different result N'_n, then we would of course assign a different probability p'; but the properties of the fissionable material would remain the same.

We now fire a neutron at a thickness $x = M_\delta$, and define the propositions

$$F^n \equiv \text{the neutron will cause fission in the } n\text{th slab of thickness } \Delta;$$
$$f^n \equiv \text{the neutron will not cause fission in the } n\text{th slab.}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

The probability of fission in slab 1 is then

$$p \equiv P(F_1|N_n) = \frac{n+1}{N+2}. \tag{18.72}$$

But now the probability that fission will occur in the second but not the first slab is *not* $p(1-p)$ as in the first treatment. At this point we see one of the fundamental differences between the theories. From the product rule, we have

$$\begin{aligned} P(F^2 F^1|N_n) &= P(F^2|F^1 N_n)P(F^1|N_n) \\ &= \frac{n+1}{N+2}\left[1 - \frac{n+1}{N+2}\right] \\ &= \frac{(n+1)(N-n+1)}{(N+2)(N+3)}. \end{aligned} \tag{18.73}$$

The difference is that, in calculating the probability $P(F^2|F^1 N_n)$, we must take into account the evidence, F^1, that a neutron has passed through one more thickness Δ without fission. This amounts to one more experiment in addition to that leading to N_n. The evidence F^1 is fully as cogent as N_n, and it would be clearly inconsistent to take one into account and ignore the other. Continuing in this way, we find that the probability that the incident neutron will emit exactly m first-generation progeny in passing through thickness $M\Delta$ is just the expression

$$P(M_m|N_n) = \binom{M}{m} \frac{(n+m)!(N+1)!(N+M-n-m)!}{n!(N-n)!(N+M+1)!}, \tag{18.74}$$

which we have derived before, Eq. (18.24). Now, if N is not a very large number, this may differ appreciably from the value

$$P(M_m|A_p) = \binom{M}{m} p^m (1-p)^{M-m}, \tag{18.75}$$

which one obtains in the frequency approach. However, note again that, as the weight of evidence N_n increases, we find $(A_{p'}|N_n) \to \delta(p' - n/N)$, and

$$P(M_n|N_n) \to P(M_m|A_p) \tag{18.76}$$

in the limit $N \to \infty$, $(n/N) \to p$. The difference in the two results is negligible whenever $N \gg M$; i.e. when the weight of evidence N_n greatly exceeds M_m. Now let's study the difference between (18.74) and (18.75) more closely. From (18.74) we have, for the mean-value estimate of m, on the Laplace theory,

$$\overline{m} = M\frac{n+1}{N+2}. \tag{18.77}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

To state the accuracy of this estimate, we can calculate the variance of the distribution (18.74). This is most easily done by using the representation (18.23):

$$\begin{aligned}\overline{m^2} &= \sum_{m=0}^{M} m^2 \int_0^1 \mathrm{d}p P(M_m|A_p)(A_p|N_n) \\ &= \frac{(N+1)!}{n!(N-n)!}\int_0^1 \mathrm{d}p \left[M_p + M(M-1)p^2\right] p^n(1-p)^{N-n} \\ &= M\frac{n+1}{N+2} + M(M-1)\frac{(n+1)(n+2)}{(N+2)(N+3)},\end{aligned} \tag{18.78}$$

which gives the variance

$$V \equiv \overline{M^2} - \overline{m}^2 = M\left[\frac{N+M+2}{N+3}\right]\left[\frac{n+1}{N+2}\right]\left[n - \frac{n+1}{N+2}\right]; \tag{18.79}$$

while, from (18.75), the frequency theory gives

$$\overline{m_0} = M_p \tag{18.80}$$

$$V_0 \equiv \left[\overline{m^2} - \overline{m}^2\right]_0 = M_p(1-p). \tag{18.81}$$

If the frequentist takes the center of the confidence interval (18.68) as his 'best' estimate of p, then he will take $p = (n+1)/(N+2)$ in these equations. So, we both obtain the same estimate, but the variance (18.79) is greater by the amount

$$V - V_0 = \frac{M-1}{N+3}M_p(1-p). \tag{18.82}$$

Why this difference? Why is it that the Laplace theory seems to determine the value of m less precisely then the frequency theory? Well, appearances are deceptive here. The fact is that the Laplace theory determines the value of m *more* precisely than the frequency theory; the variance (18.81) is not the entire measure of the uncertainty as to m on the frequency theory, because there is still the uncertainty as to the 'true' value of p. According to (18.81), p is uncertain by about $\pm\sqrt{2p(1-p)/N}$, so the mean value (18.80) is uncertain by about

$$\pm\sqrt{\frac{2p(1-p)}{N}} \tag{18.83}$$

in addition to the uncertainty represented by (18.81). If we suppose that the uncertainties (18.81) and (18.83) are independent, the total mean-square uncertainty as to the value of m on the frequency theory would be represented by the sum of (18.81) and

$$M^2\frac{2p(1-p)}{N}, \tag{18.84}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

which more than wipes out the difference (18.82). The factor of 2 in (18.84) would of course be changed somewhat by adopting a different confidence level, but no reasonable choice can change it very much.

In the frequency theory, the two uncertainties (18.81) and (18.84) appear as entirely separate effects which are determined by applying two different principles: one by conventional probability theory, the other by confidence intervals. In the Laplace theory no such distinction exists; both are given automatically by a *single* calculation. We found exactly this same situation in our particle-counter problem in Chapter 6, when we compared our robot's procedure with that of the orthodox statistician.

The mechanism by which the Laplace theory is able to do this is very interesting. It is just the difference already noted; in the derivation of (18.74), we are continually taking into account additional evidence accumulated in the new experiment, such as F^1 in (18.73). In the frequency theory, the uncertainty (18.83) in p arises because only a finite amount of data was provided by the preliminary experiment given N_n. It is just for that reason that new evidence, such as F^1, is still relevant. In giving a consistent treatment of *all* the evidence, the Laplace theory automatically includes the effect of the finiteness of the preliminary data, which the frequency theory is able to do only crudely by the introduction of confidence intervals. In the Laplace theory there is no need to decide on any arbitrary 'confidence level' because probability theory, when consistently applied to the *whole* problem, already tells us what weight should be given to the preliminary data N_n. What we get in return for this is not merely a more unified treatment; in yielding a smaller net uncertainity in m, the Laplace theory shows that the two sources of uncertainity (18.81) and (18.84) of the frequency theory are *not* independent: they have a small negative correlation, so that they tend to compensate each other. That is the reason for Laplace's smaller probable error. If you think about this very hard, you will be able to see intuitively why this negative correlation has to be there – I won't deprive you of the pleasure of figuring it out for yourself. All this subtlety is completely lost in the frequency theory.

'But,' someone will object, 'you are ignoring a very practical consideration which was the original reason for introducing confidence intervals. While I grant that *in principle* it is better to treat the whole problem in a single calculation, *in practice* we usually have to break it up into two different ones. After all, the preliminary data N_n were obtained by one group of people, who had to communicate their results to another group, who then carried out the second calculation applying these data. It is a practical necessity that the first group be able to state their conclusions in a way that tells honestly what they found, *and how reliable it was.* Their data can also be used in many other ways than in your second calculation, and the introduction of confidence intervals thus fills a very important practical need for communication between different workers.'

Of course, if you have followed everything so far, you know the answer to this. The memory storage problem was our original point of departure, and the problem just discussed is a specific example of just what we pointed out more abstractly in Eq. (18.16). You see from (18.23), and also in our derivation of (18.79), that the only property of the preliminary data which we needed in order to analyze the whole problem was the A_p distribution $(A_p|N_n)$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

that resulted from the preliminary experiment. The principle of confidence intervals was introduced to fill a very practical need. But there was no need to introduce any new principle for this purpose; it is already contained in probability thoery, which shows that the *exact* way of communicating what you have learned is not by specifying confidence intervals, but by specifying your final A_p distribution.

As a further point of comparison, note that in the Laplace theory there was no need to introduce any 'statistical assumption' about independence of events in successive slabs of thickness Δ. In fact, the theory told us, as in (18.73), that these probabilities are *not* independent when we have only a finite amount of preliminary data; and it was just this fact that enabled the Laplace theory to take account of the uncertainty which the frequency theory describes by means of confidence intervals.

This brings up a very fundamental point about probability theory, which the frequency theory fails to recognize, but which is essential for applications to both communication theory and statistical mechanics, as we will show later. What do we mean by saying that two events are 'independent'?

In the frequency theory, the only kind of independence recognized is *causal* independence; i.e. the fact that one event occurred does not in itself exert any physical influence on the occurrence of the other. Thus, in the coin-tossing example discussed in Chapter 6, the fact that the coin comes up heads on one toss of course doesn't *physically* affect the result of the next toss, and so on the frequency theory one would call the coin-tossing experiment a typical case of 'independent repetitions of a random experiment'; the probability of a heads at both tosses *must* be the product of separate probabilities. But then we lose any way of describing the difference between the reasoning of Mr A and Mr B!

In Laplace's theory, 'independence' means something entirely different, which we see from a glance at the product rule: $P(AB|C) = P(B|C)P(A|BC)$. Independence means that $P(A|BC) = P(A|C)$; i.e. *knowledge* that B is true does not affect the probability we assign to A. Thus, independence means not mere *causal* independence, but *logical* independence. Even though heads at one toss does not physically predispose the coin to give heads at the next, the *knowledge* that we got heads may have a very great influence on our predictions as to the next toss.

The importance of this is that the various limit theorems, which we will say more about later, require independence in their derivations. Consequently, even though there may be strict *causal* independence, if there is not also *logical* independence, these limit theorems will not hold. Writers of the frequency school of thought, who deny that probability theory has anything to do with inductive reasoning, recognize the existence only of *causal* connections, and, as a consequence, they have long been applying these limit theorems to physical and communications processes where, we claim, they are incorrect and completely misleading. This was noted long ago by Keynes (1921), who stressed exactly this same point.

I think these comparisons make it very clear that, at least in this kind of problem, the Laplace theory *does* provide the 'better understanding and more efficient formalism' that my colleague asked for.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

18.16 The de Finetti theorem

So far we have considered the notion of an A_p distribution and derived a certain class of probability distributions from it, under the restriction that the *same* A_p distribution is to be used for all trials. Intuitively, this means that we have assumed the underlying 'mechanism' as constant, but unknown. It is clear that this is a very restrictive assumption, and the question arises: How general is the class of probability functions that we can obtain in this way? In order to state the problem clearly, let us define

$$x_n \equiv \begin{cases} 1 & \text{if } A \text{ is true on the } n\text{th trial} \\ 0 & \text{if } A \text{ is false on the } n\text{th trial.} \end{cases} \tag{18.85}$$

Then a state of knowledge about N trials is described in the most general way by a probability function $P(x_1 \dots x_N|N)$, which could, in principle, be defined arbitrarily (except for normalization) at each of the 2^N points.

We now ask: What is a necessary and sufficient condition on $P(x_1 \dots x_N|N)$ for it to be derivable from an A_p distribution? What test could we apply to a given distribution $P(x_1 \dots x_N|N)$ to tell whether it is included in our theory as given above? A necessary condition is clear from our previous equations: any distribution obtainable from an A_p distribution necessarily has the property that the probability that A is true in *n specified* trials, and false in the remaining $(N-n)$ trials, depends only on the numbers n and N; i.e. not on *which* trials in $1 \le n \le N$ were specified. If this is so, we say that $P(x_1 \dots x_N|N)$ defines an *exchangeable sequence.*

An important theorem of de Finetti (1937) asserts that the converse is also true: *any exchangeable probability function $P(x_1 \dots x_N|N)$ can be generated by an A_p distribution.* Thus there is a function $(A_p|X) = g(p)$ such that $g(p) \ge 0$, $\int_0^1 \mathrm{d}p g(p) = 1$, and the probability that in N trials A is true in n specified trials and false in the remaining $(N-n)$ trials, is given by

$$P(n|N) = \int_0^1 \mathrm{d}p p^n (1-p)^{N-n} g(p). \tag{18.86}$$

This can be proved as follows. Note that $p^n(1-p)^{N-n}$ is a polynomial of degree N:

$$\begin{aligned} p^n(1-p)^{N-n} &= p^n \sum_{m=0}^{N-n} \binom{N-m}{m} (-p)^m \\ &= \sum_{k=0}^{N} \alpha_k(N,n) p^k, \end{aligned} \tag{18.87}$$

which defines $\alpha_k(N,n)$. Therefore, *if* (18.86) holds, we would have

$$P(n|N) = \sum_{k=0}^{N} \alpha_k(N,n) \beta_k, \tag{18.88}$$

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

where

$$\beta_k = \int_0^1 \mathrm{d}p p^n g(p) \tag{18.89}$$

is the nth moment of $g(p)$. Thus, specifying $\beta_0, \ldots, \beta_N$ is equivalent to specifying all the $P(n|N)$ for $n = 0, \ldots, N$. Conversely, for given N, specifying $P(n|N)$, $0 \leq n \leq N$, is equivalent to specifying $\{\beta_0, \ldots, \beta_N\}$. In fact, β_N is the probability that $x_1 = x_2 = \cdots = x_N = 1$, regardless of what happens in later trials, and its relation to $P(n|N)$ can be established directly without reference to any function $g(p)$.

So, the problem reduces to this: if the numbers $\beta_0, \ldots, \beta_N$ are specified, under what conditions does a function $g(p) \geq 0$ exist such that (18.89) holds? This is just the well-known *Hausdorff moment problem,* whose solution can be found in many places; for example Widder (1941, Chap. 3). Translated into our notation, the main theorem is as follows. A necessary and sufficient condition that a function $g(p) \geq 0$ exists satisfying (18.89) (and therefore also (18.86)) is that there exists a number B such that

$$\sum_{n=0}^{N} \binom{N}{n} P(n|N) \leq B, \qquad N = 0, 1, \ldots. \tag{18.90}$$

But, from the interpretation of $P(n|N)$ as probabilities, we see that the equality sign always holds in (18.90) with $B = 1$, and the proof is completed.

Here is another way of looking at it, which might be made into a proof with a little more work, and perhaps discloses more clearly the intuitive reason for the de Finetti theorem, as well as showing immediately just how much we have said about $g(p)$ when we specify the $P(n|N)$. Imagine $g(p)$ expanded in the form

$$g(p) = \sum_{n=0}^{\infty} a_n \phi_n(p), \tag{18.91}$$

where $\phi_n(p)$ are the complete orthonormal set of polynomials in $0 \leq p \leq 1$, essentially the Legendre functions:

$$\begin{aligned} \phi_n(p) &= \frac{\sqrt{2n+1}}{n!} \frac{\mathrm{d}^n}{\mathrm{d}p^n} [p(1-p)]^n \\ &= (-1)^n \sqrt{2n+1}\, P_n(2p-1), \end{aligned} \tag{18.92}$$

where $\phi_n(p)$ is a polynomial of degree n, and satisfies

$$\int_0^1 \mathrm{d}p \phi_m(p) \phi_n(p) = \delta_{mn}. \tag{18.93}$$

If we substitute (18.93) into (18.86), only a finite number of terms will survive, because $\phi_k(p)$ is orthogonal to all polynomials of degree $N < k$. Then, it is easily seen that, for given N, specifying the values of $P(n|N)$, $0 \leq n \leq N$, is equivalent to specifying the first $(n + 1)$ expansion coefficients $\{a_0, \ldots, a_N\}$. Thus, as $N \to \infty$, a function $g(p)$, defined by (18.91), becomes uniquely determined to the same extent that a Fourier series uniquely

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020

determines its generating function; i.e. 'almost everywhere'. The main trouble with this argument is that the condition $g(p) \geq 0$ is not so easily established from (18.91).

18.17 Comments

The de Finetti theorem is very important to us because it shows that the connection between probability and frequency which we have found in this chapter holds for a fairly wide class of probability functions $P(x_1, \ldots, x_N|N)$, namely the class of all exchangeable sequences. These results, of course, generalize immediately to the case where there are more than two possible outcomes at each trial.

Possibly even more important, however, is the light which the de Finetti theorem sheds on one of the oldest controversies in probability theory – Laplace's first derivation of the rule of succession. The idea of an A_p distribution is not, needless to say, our own invention. The way we have introduced it here is only our attempt to translate into modern language what we think Laplace was trying to say in that famous passage, 'When the probability of a simple event is unknown, we may suppose all possible values of this probability between 0 and 1 as equally likely.' This statement, which we interpret as saying that, with no prior evidence, $(A_p|X) = \text{const.}$, has been rejected as utter nonsense by virtually everyone who has written on probability theory in this century. And, of course, on any frequency definition of probability, Laplace's statement would have no justification at all. But on any theory it is conceptually difficult, since it seems to involve the idea of a 'probability of a probability', and the use of an A_p distribution in calculations has been largely avoided since the time of Laplace.

The de Finetti theorem puts some much more solid ground under these methods. Independently of all conceptual problems, it is a *mathematical theorem* that whenever you talk about a situation where the probability of a certain sequence of results depends only on the number of successes, not on the particular trials at which they occur, all your probability distributions can be generated from a single function $g(p)$, in just the way we have done here. The use of this generating function is, moreover, a very powerful technique mathematically, as you will quickly discover if you try to repeat some of the above derivations (for example, Eq. (18.24)) without using an A_p distribution. So, it doesn't matter what we might think about the A_p distribution conceptually; its validity as a mathematical tool for dealing with exchangeable sequences is a proven fact, standing beyond the reach of philosophical objections.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 10 Jul 2018 at 07:26:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511790423.020