CHAPTER 18

THE A_p DISTRIBUTION AND RULE OF SUCCESSION

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 X that involves everything it has ever learned). It must scan its entire memory storage banks 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 conclusions rather than isolated facts.

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, convince yourself that it's an honest coin, has a 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 that, 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. My state of belief has a great instability in the case of Mars, but there's something which makes it very stable in the case of the penny.†

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 1950's, 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's 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

† 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.

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:

$$A_p \equiv \left\{ \begin{array}{l} \text{“Regardless of anything else you may have been told,} \\ \text{the probability of } A \text{ is } p.\text{”} \end{array} \right\}$$

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 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,

$$(A_p|X) = 1, \quad 0 \le p \le 1. \tag{18–2}$$

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

$$(A_p|EX) = (A_p|X)\,\frac{P(E|A_pX)}{P(E|X)} = \frac{P(E|A_p)}{P(E|X)} \tag{18–3}$$

Now,

$$P(A|E) = \int_0^1 (AA_p|E)dp\,. \tag{18–4}$$

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 (18–4) by the product rule:

$$P(A|E) = \int_0^1 P(A|A_pE)(A_p|E)dp \tag{18–5}$$

But from the definition (18–1) of A_p, the first factor is just p, and so

$$P(A|E) = \int_0^1 (A_p|E)\, p\, dp\,. \tag{18–6}$$

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.

Relevance

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

$$P(A|E) = P(A|E_a E_b) = P(A|E_a) \tag{18–7}$$

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

$$P(E_b|AE_a) = P(E_b|E_a)\frac{P(A|E_b E_a)}{P(A|E_a)} = P(E_b|E_a) \tag{18–8}$$

Let's call this property "weak irrelevance." Now does this imply that E_b is irrelevant to A_p? Evidently not, for (18–7) says only that the first moments of $(A_p|E_a)$ and $(A_p|E_a E_b)$ are the same. But suppose that for a given E_b, (18–7) holds independently of what E_a might be; call this "strong irrelevance." Then we have

$$P(A|E) = \int_0^1 (A_p|E_a E_b)\, p\, dp = \int_0^1 (A_p|E_a)\, p\, dp. \tag{18–9}$$

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

$$(A_p|E_a E_b) = (A_p|E_a) \tag{18–10}$$

and from Bayes' theorem it follows as in (18–8) that A_p is irrelevant to E_b:

$$p(E_b|A_p E_a) = p(E_b|E_a) \tag{18–11}$$

for all E_a (according to our rules of notation, Appendix B, we may use either p or P for these probability symbols).

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:

$$p(A|EF) = \int_0^1 (A_p|EF)\, p\, dp = \int_0^1 (A_p|E)\, \frac{p(F|A_p E)}{p(F|E)}\, p\, dp. \tag{18–12}$$

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

$$\frac{P(F|A_p E_a E_b)}{p(F|E_a E_b)} = \frac{p(F|A_p E_a)\left[\frac{p(E_b|F A_p E_a)}{p(E_b|A_p E_a)}\right]}{p(F|E_a)\left[\frac{p(E_b|F E_a)}{p(E_b|E_a)}\right]} = \frac{p(F|A_p E_a)}{p(F|E_a)} \tag{18–13}$$

where we have used (18–11).

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: $(F|A_p E_{aa}) = (F|A_p)$. And so we have (18–12) reduced to

$$p(A|EF) = \frac{1}{p(F|E_{aa})} \int_0^1 (A_p|E)\, p(F|A_p)\, p\, dp\,. \tag{18–14}$$

The weak point in this argument is that we haven't 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 apply to the case where the prior information is "completely resolvable." We don't know whether it is the most general case; but we do know that it is not an empty one.

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

$$(A|EF) = \frac{p(A|EF)}{p(\overline{A}|EF)} = \frac{\int_0^1 (A_p|E) p(F|A_p)\, p\, dp}{\int_0^1 (A_p|E) p(F|A_p)\,(1-p)\, dp} \tag{18–15}$$

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 *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 has ever happened to the robot which is relevant to this 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 past experience 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 doesn't mean that the robot is able to throw away entirely all of its past experience, because there's 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 that's exactly 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 $d \sin x/dx = \cos x$?

Let's look once more at Equation (18–14). 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 Figure (18.2a). In the case of previous life on Mars, my state of knowledge is described by an $(A_p|E)$ density something like Figure (18.2b), 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 working some of this out, a newspaper story appeared (Associated Press, December 14, 1957) 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.

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.†

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' is misleading, 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 levels of hormone secretion and rate of heartbeat), 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 many ways similar to Bertrand Russell's 'theory of types', which he introduced as a means of

† 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.

avoiding some paradoxes that arose in the first formulation of his *Principia Mathematica*. There may be a relation between them; but these efforts at what Peano and Poincaré called "logistic" made in the early 20'th Century are now seen as so flawed and confused – with no recognition of the concept of information and the unlimited proliferation of weird and self–contradictory definitions – that it seems safest to scrap this 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.

But we are here treading on the boundaries of present knowledge, so the material of this Chapter 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 mathematical theory. With these cautions in mind, let us examine some concrete examples suggested by the above line of thought.

An Application.

Now let' 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:

$X \equiv$ "For each trial we admit two prior hypotheses: A true, and A 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 n'th 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 true n times in N trials in the past."

$M_m \equiv$ "A true m times in M 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 you have not defined the prior information precisely until you have given, not just verbal statements, but equations, which show how you have 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

$$(A_p|X) = 1, \quad 0 \leq p \leq 1 \tag{18–16}$$

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 $p(M_m|N_n)$. First, note that by many repetitions of our product and sum rules in the same way that we found Equation (9–30), we have the binomial distributions

$$\begin{aligned} p(N_n|A_p) &= \binom{N}{n} p^n (1-p)^{N-n} \\ p(M_m|A_p) &= \binom{M}{m} p^m (1-p)^{M-m} \end{aligned} \tag{18–17}$$

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 posses 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 you see in current textbooks are all like the two above; whenever p appears as a *given* number, there's an A_p hiding invisibly in the right–hand of the probability symbols.

Mathematically, the only functional differences between what we're 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 it; and (2) we are not afraid to use Bayes' theorem to work any propositions 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 *as 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 indispensible for manipulating 'inner' probabilities.

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

$$p(N_n|X) = \int_0^1 (N_n A_p|X)dp = \int_0^1 p(N_n|A_p)(A_p|X)dp = \binom{N}{n}\int_0^1 p^n(1-p)^{N-n}dp.$$

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

$$\int_0^1 x^r(1-x)^s dx = \frac{r!s!}{(r+s+1)!} \tag{18–18}$$

Thus, we have

$$p(N_n|X) = \left\{ \begin{array}{ll} \dfrac{1}{N+1}, & 0 \le n \le N \\ 0, & N < n \end{array} \right\}, \tag{18–19}$$

i.e., just the uniform distribution of maximum entropy; $p(M_m|X)$ is found similarly. Now we can turn (18–17) around by Bayes' theorem:

$$(A_p|N_n) = (A_p|X)\frac{p(N_n|A_p)}{p(N_p|X)} = (N+1)P(N_n|A_p) \tag{18–20}$$

and so finally the desired probability is

$$p(M_m|N_n) = \int_0^1 (M_m A_p|N_n)dp = \int_0^1 p(M_m|A_pN_n)(A_p|N_n)dp. \tag{18–21}$$

Since $p(M_m|A_pN_n) = p(M_m|A_p)$ by the definition of A_p, we have worked out everything in the integrand. Substituting into (18–21), we have again an Eulerian integral, and our result is

$$p(M_m|N_n) = \frac{\dbinom{n+m}{n}\dbinom{N+M-n-m}{N-n}}{\dbinom{N+M+1}{M}} \tag{18–22}$$

Note that this is not the same as the hypergeometric distribution (3–18) of sampling theory. Let's look at this result first in the special case $M = m = 1$. It will then reduce 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

$$p(A|N_n) = \frac{n+1}{N+2}\,. \tag{18–23}$$

We recognize Laplace's rule of succession, which we found before and discussed briefly in terms of Urn sampling in (6–28) – (6–43). Now we need to discuss it more carefully, in a wider context.

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 you'll find this rule 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 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 5,000 years.† One gets a rather large factor (odds of $5000 \times 365.2426 + 1 = 1826214 : 1$) 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 Jeffreys (1939) and Keynes (1921) 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 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. First, 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 was never intended to be applied, because there is additional prior information which the rule of succession does not taken into account.

If you go back and read Laplace (1819) 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 get an entirely different answer. Probability theory gives the results of consistent plausible reasoning on the basis of the information *which was put into it.*

† Some passages in the Bible led early theologians to conclude that the age of the world is about 5,000 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 5,000 years.

But it has to be admitted that, in mentioning the sunrise at all, Laplace made a very unfortunate choice of something to use as 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 conclusion that Laplace had committed a ludicrous error. His friend Keynes was just as quick as his enemy Venn to jump to this conclusion.

Here are some famous examples of the kind of objections to the rule of succession which you will find 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 10 years old today. According to the rule of succession, he has the probability 11/12 of living one more year. His grandfather is 70; and so 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.

What do we mean when we say that there's 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 got 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.*

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.

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're going to consider in that problem. That is part of the "boundary conditions" which must be specified before we have a uniquely defined mathematical problem. If you say, "I don't know what the possible propositions are," that is mathematically equivalent to saying, "I don't know what problem I what to solve" and the only thing the robot can answer in return is: "Come back and ask me again after you have thought about it some more."

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. In the first place, it is true that, on this evidence $P(\text{all have beaks}) \approx 1/1000$ according to Laplace's rule. But also $P(\text{all but one have beaks}) \approx 1/1000$, $P(\text{all but two have beaks}) \approx 1/1000, \cdots$ *etc.* More specifically, if there are N feathered animals, of which we have seen r all with beaks, then rewriting (18–22) in this notation we see that $P(\text{all have beaks}) = P_0 = (r+1)/(N+1) \approx 1/1000$, while $P(\text{all but } n \text{ have beaks})$ is

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

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

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

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.

But in the second place, 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 (18–22) says, we need to consider a problem where our prior information corresponds better to that supposed in its derivation.

Bass or Carp?

A guide of unquestioned veracity assures us that a certain lake contains only two species of fish: Bass and Carp. We catch ten and find them all 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 ten percent Bass, then in ten catches we had a reasonably good chance of finding one; so our state of belief drops off rapidly above ten percent. 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%).

What, then, does Laplace's rule say? Well, it says the "best" estimate of the Bass fraction f, by the criterion of minimum mean–square error, is $\langle f \rangle = 1/12$, or 8.3 percent. The posterior cumulative *pdf* for f is $p(f < f_0|DX) = 1 - (1-f_0)^{11}$; thus the median value is

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

or 6.1 percent; it is an even bet that the Bass population is less than this. The data yield a probability of 2/3, or odds of 2:1, that the lake contains less than 9.5 percent Bass, odds of 5:1 that it is less than 15%, and odds of 10:1 that it is less than 19.6%. 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.

Suppose now that our eleventh 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 population were less than 5%, then we would be unlikely to find one in only eleven 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 percent, and the posterior density is $P(df|DX) = 132 f(1-f)^{10} df$. This yields a median value of 13.6 percent, raised very considerably because the new datum has effectively eliminated the possibility that the Bass population might be below about three percent, 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 was not used in the derivation of the rule of succession.

So Where Does The Rule Stand?

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 (18–2)? 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 ambiguity 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 so much that the solution is useless.

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 a 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 you can, of course, generalize it to any number of hypotheses, as follows.

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 $p(m_1 \ldots m_K|n_1 \ldots n_K)$ that answers this, define the prior knowledge by a K–dimensional uniform prior A_p–density:

$$(A_{p_1} \cdots {}_{p_K}|X) = C\delta(p_1 + \cdots + p_K - 1), \quad p_i \geq 0 \tag{18–24}$$

To find the normalization constant C, we set

$$\int_0^\infty dp_1 \cdots \int_0^\infty dp_K (A_{p_1} \cdots {}_{p_k}|X) = 1 = CI(1) \tag{18–25}$$

where

$$I(r) \equiv \int_0^\infty dp_1 \cdots \int_0^\infty dp_k \delta(p_1 + \cdots + p_K - r) \tag{18–26}$$

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. First, take the Laplace transform of (18–26)

$$\int_0^\infty e^{-\alpha r} I(r) dr = \int_0^\infty dp_1 \cdots \int_0^\infty dp_K e^{-\alpha(p_1 + \cdots + p_K)} = \frac{1}{\alpha^K} \tag{18–27}$$

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

$$I(r) = \frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} \frac{e^{\alpha r}}{\alpha^K} \, d\alpha = \frac{1}{(K-1)!} \frac{d^{K-1}}{d\alpha^{K-1}} e^{\alpha r}\Big|_{\alpha=0} = \frac{r^{K-1}}{(K-1)!} \tag{18–28}$$

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,

$$C = \frac{1}{I(1)} = (K-1)! \tag{18–29}$$

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 (18–26) directly. The prior $p(n_1 \ldots n_K|X)$ is then, using the same trick,

$$\begin{aligned} p(n_1 \ldots n_K|X) &= \frac{N!}{n_1! \ldots n_K!} \int_0^\infty dp_1 \cdots \int_0^\infty dp_K \, p_1^{n_1} \cdots p_K^{n_K} (A_{p_1} \ldots A_{p_K}|X) \\ &= \frac{N_1! \, (K-1)!}{n_1! \ldots n_K!} J(1) \end{aligned} \tag{18–30}$$

where

$$J(r) \equiv \int_0^\infty dp_1 \cdots \int_0^\infty dp_K \, p_1^{n_1} \cdots p_K^{n_K} \; \delta(p_1 + \cdots + p_k - r) \tag{18–31}$$

which we evaluate as before by taking the Laplace transform:

$$\int_0^\infty \mathrm{e}^{-\alpha r} J(r) dr = \int_0^\infty dp_1 \cdots \int_0^\infty dp_K \, p_1^{n_1} \cdots p_K^{n_K} \; e^{-\alpha(p_1 + \cdots + p_K)} = \prod_{i=1}^{K} \frac{n_i!}{\alpha^{n_i+1}} \tag{18–32}$$

So, as in (18–28), we have

$$J(r) = \frac{n_1! \cdots n_K!}{2\pi i} \int_{-i\infty}^{+i\infty} d\alpha \, \frac{e^{\alpha r}}{\alpha^{N+K}} = \frac{n_1! \cdots n_K!}{(N+K-1)!} \, r^{N+K-1} \tag{18–33}$$

and

$$p(n_1 \cdots n_k | X) = \frac{N!\,(K-1)!}{(N+K-1)!}, \qquad n_i \ge 0, \; n_1 + \cdots + n_K = N \tag{18–34}$$

Therefore, by Bayes' theorem

$$\begin{aligned} (A_{p_1 \cdots p_K} | n_1 \cdots n_K) &= (A_{p_1 \cdots p_K} | X) \, \frac{p(n_1 \cdots n_K | A_{p_1 \cdots p_K})}{p(n_1 \cdots n_K | X)} \\ &= \frac{(N+K-1)!}{n_1! \cdots n_K!} \, p_1^{n_1} \cdots p_K^{n_K} \; \delta(p_1 + \cdots + p_K - 1) \end{aligned} \tag{18–35}$$

and finally

$$p(m_1 \ldots m_K | n_1 \ldots n_K) = \int_0^\infty dp_1 \cdots \int_0^\infty dp_K \, p(m_1 \ldots m_K | A_{p_1 \cdots p_K}) \, (A_{p_1 \cdots p_K} | n_1 \cdots n_K)$$

$$= \frac{M!}{m_1! \cdots m_K!} \, \frac{(N+K-1)!}{n_1! \cdots n_K!} \int_0^\infty dp_1 \cdots \int_0^\infty dp_K \, p_1^{n_1+m_1} \cdots p_K^{n_K+m_K} \, \delta(p_1 + \cdots + p_K - 1) \tag{18–36}$$

The integral is the same as $J(1)$ except for the replacement $n_i \to n_i + m_i$. So, from (18–33),

$$p(m_1 \cdots m_K | n_1 \cdots n_K) = \frac{M!}{m_1! \cdots m_K!} \, \frac{(N+K-1)!}{n_1! \cdots n_K!} \, \frac{(n_1+m_1)! \cdots (n_K+m_K)!}{(N+M+K-1)!} \tag{18–37}$$

or, reorganizing into binomial coefficients, the generalization of (18–22) is

$$p(m_1 \ldots m_K | n_1 \ldots n_K) = \frac{\binom{n_1+m_1}{n_1} \cdots \binom{n_K+m_K}{n_K}}{\binom{N+M+K-1}{M}}. \tag{18–38}$$

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 of succession:

$$p(A_1 | n_1, N, K) = \frac{n_1 + 1}{N + K}. \tag{18–39}$$

You 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 Carnap in 1945, which he denotes $c^*(h, e)$ in his "Continuum of Inductive Methods."

Now, 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 get 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, get 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're developing here.

The real reason for introducing the rule of succession lies in the cases where we *do* get 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 (18–38). You see from (18–39) that as soon as the number of observations N is large compared to 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 (18–39) supplies an obviously needed refinement of this, small correction terms when the number of observations is not large compared to the number of propositions.

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 prior evidence and F is new evidence, then

$$p(A|EF) = \int_0^1 (A_p|EF)p\,dp = \frac{\int_0^1 (A_p|F)(A_p|E)p\,dp}{\int_0^1 (A_p|F)(A_p|E)\,dp} \tag{18–40}$$

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;

$$p(A|EF) = p(A|E) \tag{18–41}$$

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. (18–20), gives

$$(A_p|N_n) = \frac{(N+1)!}{n!(N-n)!} p^n (1-p)^{N-n} \approx (\text{constant}) \cdot \exp\left\{ - \left[\frac{(p-f)^2}{2\sigma^2} \right] \right\} \tag{18–42}$$

where

$$\sigma^2 = \frac{f(1-f)}{n} \tag{18–43}$$

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 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'll call weight of evidence. Consider A_p, given two different pieces of evidence, E and F,

$$(A_p|EF) = (\text{constant}) \times (A_p|E)(A_p|F). \tag{18–44}$$

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 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. Now seeing this, we can get a little more insight into the principle of indifference and also make contact between this theory and Carnap's methods of inductive reasoning.

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 have to be satisfied: (1) we have to 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 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 and 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 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.

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 $p(A|N_nX)$ 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

$$p(A|N_nX) = \frac{2n+\lambda}{2N+2\lambda} = \frac{(n+r)+1}{(N+2r)+2}. \qquad (18\text{–}45)$$

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

In our coin–tossing example, the gentleman form 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 Equation (18–45), you 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 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.

But 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 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

$$(A_p|\ )dp = (\text{constant}) \cdot \frac{dp}{p(1-p)} \qquad (18\text{–}46)$$

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 (18–46) already in Chapter 6, Eq. (6–46).

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 (18–46).

How different are the numerical results which the pre–prior assignment (18–46) gives us? Repeating the derivation of (18–20) with this pre–prior assignment we find that, provided n is not zero or N,

$$(A_p|N_n) = \frac{(N-1)!}{(n-1)!(N-n-1)!} p^{n-1}(1-p)^{N-n-1} \qquad (18\text{–}47)$$

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

$$p(A|N_n) = \int_0^1 (A_p|N_n)p\,dp = \frac{n}{N} \qquad (18\text{–}48)$$

equal to the observed frequency, and identical with the maximum–likelihood estimate of p. Likewise, provided $0 < n < N$, we find instead of (18–22) the formula

$$p(M_m|N_n) = \frac{\binom{m+n-1}{m}\binom{M-m+N-n-1}{M-m}}{\binom{N+M-1}{M}} \tag{18–49}$$

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

The de Finetti Representation Theorem

****************************** MORE COMING! ***************************