CHAPTER 5

QUEER USES FOR PROBABILITY THEORY

"I cannot conceal the fact here that in the specific application of these rules, I foresee many things happening which can cause one to be badly mistaken if he does not proceed cautiously . . . "

— James Bernoulli (1713); Part 4, Chapter III

I. J. Good (1950) has shown how we can use probability theory backwards to measure our own strengths of belief about propositions. For example, how strongly do you believe in extrasensory perception?

Extrasensory Perception

What probability would you assign to the hypothesis that Mr. Smith has perfect extrasensory perception? He can guess right every time which number you have written down. To say zero is too dogmatic. According to our theory, this means that we are never going to allow the robot's mind to be changed by any amount of evidence, and we don't really want that. But where *is* our strength of belief in a proposition like this?

Our brains work pretty much the way this robot works, but we have an intuitive feeling for plausibility only when it's not too far from 0 *db*. We get fairly definite feelings that something is more than likely to be so or less than likely to be so. So the trick is to imagine an experiment. How much evidence would it take to bring your state of belief up to the place where you felt very perplexed and unsure about it? Not to the place where you believed it – that would overshoot the mark, and again we'd lose our resolving power. How much evidence would it take to bring you just up to the point where you were beginning to consider the possibility seriously?

We take this man who says he has extrasensory perception, and we will write down some numbers from 1 to 10 on a piece of paper and ask him to guess which numbers we've written down. We'll take the usual precautions to make sure against other ways of finding out. If he guesses the first number correctly, of course we will all say "you're a very lucky person, but I don't believe it." And if he guesses two numbers correctly, we'll still say "you're a very lucky person, but I don't believe it." By the time he's guessed four numbers correctly – well, I still wouldn't believe it. So my state of belief is certainly lower than $-40\,db$.

How many numbers would he have to guess correctly before you would really seriously consider the hypothesis that he has extrasensory perception? In my own case, I think somewhere around 10. My personal state of belief is, therefore, about $-100\,db$. You could talk me into a ± 10 change, and perhaps as much as ± 30, but not much more than that.

But on further thought we see that, although this result is correct, it is far from the whole story. In fact, if he guessed 1000 numbers correctly, I still would not believe that he has ESP, for an extension of the same reason that we noted in Chapter 4 when we first encountered the phenomenon of resurrection of dead hypotheses. An hypothesis A that starts out down at $-100\,db$ can hardly ever come to be believed whatever the data, because there are almost sure to be alternative hypotheses (B_1, $B_2, \ldots$) above it, perhaps down at -60 *db*. Then when we get astonishing data that might have resurrected A, the alternatives will be resurrected instead. Let us illustrate this by two famous examples, involving telepathy and the discovery of Neptune. Also we note some interesting variants of this. Some are potentially useful, some are instructive case histories of probability theory gone wrong, in the way Bernoulli warned us about.

Mrs. Stewart's Telepathic Powers

Before venturing into this weird area, the writer must issue a disclaimer. I was not there, and am not in a position to affirm that the experiment to be discussed actually took place; or if it did, that the data were actually obtained in a valid way. Indeed, that is just the problem that you and I always face when someone tries to persuade us of the reality of ESP or some other marvelous thing – such things never happen to us or in our presence. All we are able to affirm is that the experiment and data have been reported in a real, verifiable reference (Soal and Bateman, 1954). This is the circumstance that we want to analyze now by probability theory. Lindley (1957) and Bernardo (1980) have also taken note of it from the standpoint of probability theory, and Boring (1955) discusses it from the standpoint of psychology.

In the reported experiment, from the experimental design the probability of guessing a card correctly should have been $p = 0.2$, independently in each trial. Let H_p be the "null hypothesis" which states this, and supposes that only "pure chance" is operating (whatever that means). According to the binomial distribution (3–74) H_p predicts that if a subject has no ESP, the number r of successful guesses in n trials should be about (mean $\pm$ standard deviation):

$$(r)_{est} = np \pm \sqrt{np(1-p)}\,. \qquad (5\text{–}1)$$

For $n = 37100$ trials, this is 7420 ± 77.

But according to the report, Mrs. Gloria Stewart guessed correctly $r = 9410$ times in 37100 trials, for a fractional success rate of $f = 0.2536$. These numbers constitute our data D. At first glance, they may not look very sensational; note, however, that her score was

$$\frac{9410 - 7420}{77} = 25.8 \qquad (5\text{–}2)$$

standard deviations away from the chance expectation.

The probability of getting these data, on hypothesis H_p is then the binomial

$$P(D|H_p) = \binom{n}{r} p^r (1-p)^{n-r} \quad . \qquad (5\text{–}3)$$

But the numbers n, r are so large that we need the Stirling approximation to the binomial, derived in Appendix E:

$$P(D|H_p) = A e^{nH(f,p)} \qquad (5\text{–}4)$$

where

$$H(f,p) = f\ \log(p/f) + (1-f)\ \log[(1-p)/(1-f)] = -0.008452 \qquad (5\text{–}5)$$

is the entropy of the observed distribution $(f, 1-f) = (0.2536,\ 0.7464)$ relative to the expected one $(p,\ 1-p) = (0.2000,\ 0.8000)$, and

$$A = \left[\frac{n}{2\pi\, r(n-r)}\right]^{1/2} = 0.00476\,. \qquad (5\text{–}6)$$

Then we may take as the likelihood L_p of H_p, the sampling probability

$$L_p = P(D|H_p) = 0.00476 \exp(-313.6) = 3.15 \times 10^{-139}\,. \qquad (5\text{–}7)$$

This looks fantastically small; but before jumping to conclusions the robot should ask: are the data also fantastically improbable on the hypothesis that Mrs. Stewart has telepathic powers? If they are, then (5–7) may not be so significant after all.

Consider the Bernoulli class of alternative hypotheses H_q $(0 \le q \le 1)$, which suppose that the trials are independent, but that assign different probabilities of success q to Mrs. Stewart ($q > 0.2$ if the hypothesis considers her to be telepathic). Out of this class, the hypothesis H_f that assigns $q = f = 0.2536$ yields the greatest $P(D|H_q)$ that can be attained in the Bernoulli class, and for this the entropy (5–5) is zero, yielding a maximum likelihood of

$$L_f = P(D|H_f) = A = 0.00476 \,. \tag{5–8}$$

So if the robot knew for a fact that Mrs. Stewart is telepathic to the extent of $q = 0.2536$, then the probability that she could generate the observed data would not be particularly small. Therefore, the smallness of (5–7) is indeed highly significant; for then the likelihood ratio for the two hypotheses must be fantastically small. The relative likelihood depends only on the entropy factor:

$$\begin{aligned} \frac{L_p}{L_f} &= \frac{P(D|H_p)}{P(D|H_f)} = e^{nH} = \exp(-313.6) \\ &= 6.61 \times \; 10^{-137} \end{aligned} \quad . \tag{5–9}$$

and the robot would report: "the data do indeed support H_f over H_p by an enormous factor".

Digression on the Normal Approximation

Note, in passing, that in this calculation large errors could be made by unthinking use of the normal approximation to the binomial, also derived in Appendix E (or compare with (4–62)):

$$P(D|H_p, X) \simeq (\text{const.}) \times \exp\left[\frac{-n(f-p)^2}{2p(1-p)}\right] \tag{5–10}$$

To use it here instead of the entropy approximation (5–4), amounts to replacing the entropy $H(f, p)$ by the first term of its power series expansion about the peak. Then we would have found instead a likelihood ratio $\exp(-333.1)$. Thus the normal approximation would have made Mrs. Stewart appear even more marvelous than the data indicate, by an additional odds ratio factor of

$$\exp(19.5) = 2.94 \times 10^8 \,. \tag{5–11}$$

This should warn us that, quite generally, normal approximations cannot be trusted far out in the tails of a distribution. In this case, we are 25.8 standard deviations out, and the normal approximation is in error by over eight orders of magnitude.

Unfortunately, this is just the approximation used by the Chi–squared test discussed later, which can therefore lead us to wildly misleading conclusions when the "null hypothesis" being tested fits the data very poorly. Those who use the Chi–squared test to support their claims of marvels are usually helping themselves by factors such as (5–11). In practice, as discussed in Appendix E, the entropy calculation (5–5) is just as easy and far more trustworthy (although they amount to the same thing within one or two standard deviations of the peak).

Back to Mrs. Stewart

In any event, our present numbers are indeed fantastic; on the basis of such a result, ESP researchers would proclaim a virtual certainty that ESP is real. If we compare H_p and H_f by probability theory, the posterior probability that Mrs. Stewart has ESP to the extent of $q = f = 0.2536$, is

$$P(H_f|D, X) = P(H_f|X)\,\frac{P(D|H_f, X)}{P(D|X)} = \frac{P_f L_f}{P_f L_f \;+\; P_p L_p} \tag{5–12}$$

where P_p, P_f are the prior probabilities of H_p, H_f. But because of (5–9), it hardly matters what these prior probabilities are; in the view of an ESP researcher who does not consider the prior probability $P_f = P(H_f|X)$ particularly small, $P(H_f|D,X)$ is so close to unity that its decimal expression starts with over a hundred 9's.

He will then react with anger and dismay when, in spite of what he considers this overwhelming evidence, we persist in not believing in ESP. Why are we, as he sees it, so perversely illogical and unscientific?

The trouble is that the above calculations (5–9) and (5–12) represent a very naïve application of probability theory, in that they consider only H_p and H_f; and no other hypotheses. If we really knew that H_p and H_f were the only possible ways the data (or more precisely, the observable report of the experiment and data) could be generated, then the conclusions that follow from (5–9) and (5–12) would be perfectly all right. But in the real world, our intuition is taking into account some additional possibilities that they ignore.

Probability theory gives us the results of consistent plausible reasoning from the information *that was actually used* in our calculation. It can lead us wildly astray, as Bernoulli noted in our opening quotation, if we fail to use all the information that our common sense tells us is relevant to the question we are asking. When we are dealing with some extremely implausible hypothesis, recognition of a seemingly trivial alternative possibility can make orders of magnitude difference in the conclusions. Taking note of this, let us show how a more sophisticated application of probability theory explains and justifies our intuitive doubts.

Let H_p, H_f, and L_p, L_f, P_p, P_f be as above; but now we introduce some new hypotheses about how this report of the experiment and data might have come about, which will surely be entertained by the readers of the report even if they are discounted by its writers.

These new hypotheses $(H_1, H_2 \ldots H_k)$ range all the way from innocent possibilities such as unintentional error in the record keeping, through frivolous ones (perhaps Mrs. Stewart was having fun with those foolish people, with the aid of a little mirror that they did not notice), to less innocent possibilities such as selection of the data (not reporting the days when Mrs. Stewart was not at her best), to deliberate falsification of the whole experiment for wholly reprehensible motives. Let us call them all, simply, "deception". For our purposes it does not matter whether it is we or the researchers who are being deceived, or whether the deception was accidental or deliberate. Let the deception hypotheses have likelihoods and prior probabilities L_i, P_i, $i = (1, 2, ..., k)$.

There are, perhaps, 100 different deception hypotheses that we could think of and are not too far–fetched to consider, although a single one would suffice to make our point.

In this new logical environment, what is the posterior probability of the hypothesis H_f that was supported so overwhelmingly before? Probability theory now tells us that

$$P(H_f|D,X) = \frac{P_f L_f}{P_f L_f + P_p L_p + \Sigma P_i L_i} \,. \tag{5–13}$$

Introduction of the deception hypotheses has changed the calculation greatly; in order for $P(H_f|DX)$ to come anywhere near unity it is now necessary that

$$P_p L_p + \Sigma_i P_i L_i << P_f L_f \,. \tag{5–14}$$

Let us suppose that the deception hypotheses have likelihoods L_i of the same order as L_f in (5–8); *i.e.*, a deception mechanism could produce the reported data about as easily as could a truly telepathic Mrs. Stewart. From (5–7), $P_p L_p$ is completely negligible, so (5–14) is not greatly different from

$$\sum P_i << P_f \,. \tag{5–15}$$

But each of the deception hypotheses is, in my judgment, more likely than H_f, so there is not the remotest possibility that the inequality (5–15) could ever be satisfied.

Therefore, this kind of experiment can never convince me of the reality of Mrs. Stewart's ESP; not because I assert $P_f = 0$ dogmatically at the start, but because the verifiable facts can be accounted for by many alternative hypotheses, every one of which I consider inherently more plausible than H_f, and none of which is ruled out by the information available to me.

Indeed, the very evidence which the ESP'ers throw at us to convince us, has the opposite effect on our state of belief; issuing reports of sensational data defeats its own purpose. For if the prior probability of deception is greater than that of ESP, then the more improbable the alleged data are on the null hypothesis of no deception and no ESP, the more strongly we are led to believe, not in ESP, but in deception. For this reason, the advocates of ESP (or any other marvel) will never succeed in persuading scientists that their phenomenon is real, until they learn how to eliminate the possibility of deception in the mind of the reader. As (5–15) shows, the reader's total prior probability of deception by all mechanisms must be pushed down below that of ESP.

It is interesting that Laplace perceived this phenomenon long ago. His *Essai Philosophique sur les probabilités* (1819) has a long chapter on the "Probabilities of Testimonies", in which he calls attention to "*the immense weight of testimonies necessary to admit a suspension of natural laws*". He notes that those who make recitals of miracles, "*decrease rather than augment the belief which they wish to inspire; for then those recitals render very probable the error or the falsehood of their authors. But that which diminishes the belief of educated men often increases that of the uneducated, always avid for the marvelous.*"

We observe the same phenomenon at work today, not only in the ESP enthusiast, but in the astrologer, reincarnationist, exorcist, fundamentalist preacher or cultist of any sort, who attracts a loyal following among the uneducated by claiming all kinds of miracles; but has zero success in converting educated people to his teachings. Educated people, taught to believe that a cause–effect relation requires a physical mechanism to bring it about, are scornful of arguments which invoke miracles; but the uneducated seem actually to prefer them.

Note that we can recognize the clear truth of this psychological phenomenon without taking any stand about the truth of the miracle; it is possible that the educated people are wrong. For example, in Laplace's youth educated persons did not believe in meteorites, but dismissed them as ignorant folklore because they are so rarely observed. For one familiar with the laws of mechanics the notion that "stones fall from the sky" seemed preposterous, while those without any conception of mechanical law saw no difficulty in the idea. But the fall at Laigle in 1803, which left fragments studied by Biot and other French scientists, changed the opinions of the educated – including Laplace himself. In this case the uneducated, avid for the marvelous, happened to be right: *c'est la vie*.

Indeed, in the course of writing this Chapter, the writer found himself a victim of this phenomenon. In the Ph. D. thesis of G. L. Bretthorst (1987), reported in Chapter 21 below and more fully in Bretthorst (1988), we applied Bayesian analysis to estimation of frequencies of nonstationary sinusoidal signals, such as exponential decay in nuclear magnetic resonance (NMR) data, or chirp in oceanographic waves. We found – as was expected on theoretical grounds – an improved resolution over previous fourier transform methods.

If we had claimed a 50% improvement, we would have been believed at once, and other researchers would have adopted this method eagerly. But in fact we found orders of magnitude improvement in resolution. It was, in retrospect, foolish of us to mention this at the outset, for in the minds of others the prior probability that we were irresponsible charlatans was greater than the prior probability that a new method could possibly be that good; and we were not at first believed.

Fortunately, we were able, by presenting many numerical analyses of data and distributing free computer programs so that doubters could check our claims for themselves on whatever data they chose, to eliminate the possibility of deception in the minds of our audience, and the method did find acceptance after all. The Bayesian analyses of free decay NMR signals now permits expermentalists to extract much more information from their data than was possible by taking fourier transforms.

However, the reader should be warned that our probability analysis (5–13) of Mrs. Stewart's performance is still rather naïve in that it neglects correlations; having seen a persistent deviation from the chance expectation $p = 0.2$ in the first few hundred trials, common sense would lead us to form the hypothesis that some unknown systematic cause is at work, and we would come to expect the same deviation in the future. This would alter the numerical values given above, but not enough to change our general conclusions. More sophisticated probability models which are able to take such things into account are given in our discussions of advanced applications below; relevant topics are Dirichlet priors, exchangeable sequences, and autoregressive models.

Now let us return to that original device of I. J. Good, which started this train of thought. After all this analysis, why do we still hold that naïve first answer of -100 *db* for my prior probability of ESP, as recorded above, to be correct? Because Jack Good's imaginary device can be applied to whatever state of knowledge we choose to imagine; it need not be the real one. If I knew that true ESP and pure chance were the only possibilities, then the device would apply and my assignment of -100 *db* would hold. But knowing that there are other possibilities in the real world does not change my state of belief about ESP; so the figure of -100 *db* still holds.

Therefore, in the present state of development of probability theory, the device of imaginary results is usable and useful in a very wide variety of situations, where we might not at first think it applicable. We shall find it helpful in many cases where our prior information seems at first too vague to lead to any definite prior probabilities; it stimulates our thinking and tells us how to assign them after all. Perhaps in the future we shall have more formal principles that make it unnecessary.

Exercise 5.1. By applying the device of imaginary results, find your own strength of belief in any three of the following propositions: (1) Julius Caesar is a real historical person (*i.e.,* not a myth invented by later writers); (2) Achilles is a real historical person; (3) The Earth is more than a million years old; (4) Dinosaurs did not die out; they are still living in remote places; (5) Owls can see in total darkness; (6) The configuration of the planets influences our destiny; (7) Automobile seat belts do more harm than good; (8) High interest rates combat inflation; (9) High interest rates cause inflation. [*Hint*: Try to imagine a situation in which the proposition H_0 being tested, and a single alternative H_1, would be the only possibilities, and you receive new "data" D consistent with H_0: $P(D|H_0) \simeq 1$. The imaginary alternative and data are to be such that you can calculate the probability $P(D|H_1)$. Always use an H_0 that you are inclined not to believe; if the proposition as stated seems highly plausible to you, then for H_0 choose its denial.]

Much more has been written about the Soal experiments in ESP. The deception hypothesis, already strongly indicated by our probability analysis, is supported by additional evidence (Hansel, 1980; Kurtz, 1985). Altogether, an appalling amount of effort has been expended on this incident, and it might appear that the only result was to provide a pedagogical example of the use of probability theory with very unlikely hypotheses. Can anything more useful be salvaged from it?

We think that this incident has some lasting value both for psychology and for probability theory, because it has made us aware of an important general phenomenon which has nothing to do with ESP; a person may tell the truth and not be believed, even though the disbelievers are reasoning in a rational, consistent way. To the best of our knowledge it has not been noted before that probability theory as logic *automatically* reproduces and explains this phenomenon. This leads us to conjecture that it may generalize to other more complex and puzzling psychological phenomena.

Converging and Diverging Views

Suppose that two people, Mr. A and Mr. B have differing views (due to their differing prior information) about some issue, say the truth or falsity of some controversial proposition S. Now we give them both a number of new pieces of information or "data", $D_1, D_2, \ldots D_n$, some favorable to S, some unfavorable. As n increases, the totality of their information comes to be more nearly the same, therefore we might expect that their opinions about S will converge toward a common agreement. Indeed, some authors consider this so obvious that they see no need to demonstrate it explicitly, while Howson & Urbach (1989, p. 290) claim to have demonstrated it.

Nevertheless, let us see for ourselves whether probability theory can reproduce such phenomena. Denote their prior information by I_A, I_B respectively; and let Mr. A be initially a believer, Mr. B a doubter:

$$P(S|I_A) \simeq 1 , \qquad\qquad P(S|I_B) \simeq 0 \tag{5–16}$$

and after receiving data D, their posterior probabilities are changed to

$$\begin{aligned} P(S|DI_A) &= P(S|I_A)\,\frac{P(D|SI_A)}{P(D|I_A)} \\ P(S|DI_B) &= P(S|I_B)\,\frac{P(D|SI_B)}{P(D|I_B)} \end{aligned} \tag{5–17}$$

If D supports S, then since Mr. A already considers S almost certainly true, we have $P(D|SI_A) \simeq P(D|I_A)$, and so

$$P(S|DI_A) \;\simeq\; P(S|I_A) \tag{5–18}$$

Data D have no appreciable effect on Mr. A's opinion (at least, on the probability scale; it might have on the evidence scale). But now one would think that if Mr. B reasons soundly, he must recognize that $P(D|SI_B) > P(D|I_B)$, and thus

$$P(S|DI_B) \;>\; P(S|I_B) \tag{5–19}$$

Mr. B's opinion should be changed in the direction of Mr. A's. Likewise, if D had tended to refute S, one would expect that Mr. B's opinions are little changed by it, while Mr. A's will move in the direction of Mr. B's.

From this we might conjecture that, whatever the new information D, it should tend to bring different people into closer agreement with each other, in the sense that.

$$|P(S|DI_A)) - P(S|DI_B)| < |P(S|I_A) - P(S|I_B)| \tag{5–20}$$

But although this can be verified in special cases, it is not true in general.

Is there some other measure of "closeness of agreement" such as $\log[P(S|DI_A)/P(S|DI_B)]$, for which this converging of opinions can be proved as a general theorem? Not even this is possible; the failure of probability theory to give this expected result tells us that convergence of views is not a general phenomenon. For robots and humans who reason according to the desiderata of Chapter 1, something more subtle and sophisticated is at work.

Indeed, in practice we find that this convergence of opinions usually happens for small children; for adults it happens sometimes but not always. For example, new experimental evidence does cause scientists to come into closer agreement with each other about the explanation of a phenomenon.

Then it might be thought (and for some it is an article of faith in democracy) that open discussion of public issues would tend to bring about a general concensus on them. On the contrary, we observe repeatedly that when some controversial issue has been discussed vigorously for a few years, society becomes polarized into opposite extreme camps; it is almost impossible to find anyone who retains a moderate view. The Dreyfus affair in France, which tore the nation apart for 20 years, is one of the most thoroughly documented examples of this (Bredin, 1986). Today, such issues as nuclear power, abortion, criminal justice, *etc.* are following the same course. New information given simultaneously to different people may cause a convergence of views; but it may equally well cause a divergence.

This divergence phenomenon is observed also in relatively well–controlled psychological experiments. Some have concluded that people reason in a basically irrational way; prejudices seem to be strengthened by new information which ought to have the opposite effect. Kahneman & Tversky (1972) draw the opposite conclusion from such psychological tests, and consider them an argument against Bayesian methods.

But now, in view of the above ESP example, we wonder whether probability theory might also account for this divergence and indicate that people may be, after all, thinking in a reasonably rational, Bayesian way (*i.e.,* in a way consistent with their prior information and prior beliefs). The key to the ESP example is that our new information was not

$$S \equiv \text{“Fully adequate precautions against error or deception were taken,} \\ \text{and Mrs. Stewart did in fact deliver that phenomenal performance.”} \tag{5–21}$$

It was that some ESP researcher has *claimed* that S is true. But if our prior probability for S is lower than our prior probability that we are being deceived, hearing this claim has the opposite effect on our state of belief from what the claimant intended.

But the same is true in science and politics; the new information a scientist gets is not that an experiment did in fact yield this result, with adequate protection against error. It is that some colleague has *claimed* that it did. The information we get from the TV evening news is not that a certain event actually happened in a certain way; it is that some news reporter has *claimed* that it did. Even seeing the event on our screens can no longer convince us, after recent revelations that all major U.S. networks had faked some videotapes of alleged news events.

Scientists can reach agreement quickly because we trust our experimental colleagues to have high standards of intellectual honesty and sharp perception to detect possible sources of error. And this belief is justified because, after all, hundreds of new experiments are reported every month, but only about once in a decade is an experiment reported that turns out later to have been wrong. So our prior probability of deception is very low; like trusting children, we believe what experimentalists tell us.

In politics, we have a very different situation. Not only do we doubt a politician's promises, few people believe that news reporters deal truthfully and objectively with economic, social, or political topics. We are convinced that virtually all news reporting is selective and distorted, designed not to report the facts, but to indoctrinate us in the reporter's socio–political views. And this belief is justified abundantly by the internal evidence in the reporter's own product – every choice of words and inflection of voice shifting the bias invariably in the same direction.

Not only in political speeches and news reporting, but wherever we seek for information on political matters, we run up against this same obstacle; we cannot trust anyone to tell us the truth, because we perceive that everyone who wants to talk about it is motivated either by self–interest or by ideology. In political matters, whatever the source of information, our prior probability of deception is always very high. However, it is not obvious whether this alone can prevent us from coming to agreement.

With this in mind, let us reexamine the equations of probability theory. To compare the reasoning of Mr. A and Mr. B we could write Bayes' theorem (5–17) in the logarithmic form

$$\log\left[\frac{P(S|DI_A)}{P(S|DI_B)}\right] = \log\left[\frac{P(S|I_A)}{P(S|I_B)}\right] + \log\left[\frac{P(D|SI_A)\,P(D|I_B)}{P(D|I_A)\,P(D|SI_B)}\right] \tag{5–22}$$

which might be described by a simple hand–waving mnemonic like

"log posterior = log prior + log likelihood"

Note, however, that (5–22) differs from our log–odds equations of Chapter 4, which might be described by the same mnemonic. There we compared different hypotheses, given the same prior information, and some factors $P(D|I)$ cancelled out. Here we are considering a fixed hypothesis S, in the light of different prior information and they do not cancel, so the "likelihood" term is different.

In the above we supposed Mr. A to be the believer, so log (prior) > 0. Then it is clear that on the log scale their views will converge as expected, the left–hand side of (5–22) tending to zero monotonically (*i.e.*, Mr A will remain a stronger believer than Mr. B) if

$$-\text{log prior} < \text{log likelihood} < 0 \ ,$$

and they will diverge monotonically if

$$\text{log likelihood} > 0 \ .$$

But they will converge with reversal (Mr. B becomes a stronger believer than Mr. A) if

$$-2\,\text{log prior} < \text{log likelihood} < -\text{log prior} \ ,$$

and they will diverge with reversal if

$$\text{log likelihood} < -2\,\text{log prior} \ .$$

Thus probability theory appears to allow, in principle, that a single piece of new information D could have every conceivable effect on their relative states of belief.

But perhaps there are additional restrictions, not yet noted, which make some of these outcomes impossible; can we produce specific and realistic examples of all four types of behavior? Let us examine only the monotonic convergence and divergence by the following scenario, leaving it as an exercise for the reader to make a similar examination of the reversal phenomena.

The new information D is: "Mr. N has gone on TV with a sensational claim that a commonly used drug is unsafe" and three viewers, Mr. A, Mr. B, and Mr. C, see this. Their prior probabilities $P(S|I)$ that the drug is safe are (0.9, 0.1, 0.9) respectively; *i.e.*, initially, Mr. A and Mr. C were believers in the safety of the drug, Mr. B a disbeliever.

But they interpret the information D very differently, because they have different views about the reliability of Mr. N. They all agree that, if the drug had really been proved unsafe, Mr. N would be right there shouting it: that is, their probabilities $P(D|\overline{S}I)$ are (1, 1, 1); but Mr. A trusts his honesty while Mr. C does not; their probabilities $P(D|SI)$ that, if the drug is safe, Mr. N would say that it is unsafe, are (0.01, 0.3, 0.99) respectively.

Applying Bayes' theorem $P(S|DI) = P(S|I)\,P(D|SI)/P(D|I)$ and expanding the denominator by the product and sum rules: $P(D|I) = P(S|I)\,P(D|SI) + P(\overline{S}|I)\,P(D|\overline{S}I)$, we find their posterior probabilities that the drug is safe to be (.083, .032, .899) respectively. Put verbally, they have reasoned as follows:

A: "Mr. N is a fine fellow, doing a notable public service. I had thought the drug to be safe from other evidence, but he would not knowingly misrepresent the facts; therefore hearing his report leads me to change my mind and think that the drug is unsafe after all. My belief in safety is lowered by 20.0 db, so I will not buy any more."

B: "Mr. N is an erratic fellow, inclined to accept adverse evidence too quickly. I was already convinced that the drug is unsafe; but even if it is safe he might be carried away into saying otherwise. So hearing his claim does strengthen my opinion, but only by 5.3 db. I would never under any circumstances use the drug."

C: "Mr. N is an unscrupulous rascal, who does everything in his power to stir up trouble by sensational publicity. The drug is probably safe, but he would almost certainly claim it is unsafe whatever the facts. So hearing his claim has practically no effect (only .005 db) on my confidence that the drug is safe. I will continue to buy it and use it."

The opinions of Mr. A and Mr. B converge in about the way we conjectured in (5–20) because both are willing to trust Mr. N's veracity to some extent. But Mr. A and Mr. C diverge because their prior probabilities of deception are entirely different. So one cause of divergence is, not merely that prior probabilities of deception are large, but that they are greatly different for different people.

However, this is not the only cause of divergence; to show this we introduce Mr. X and Mr. Y, who agree in their judgment of Mr. N:

$$P(D|SI_X) = P(D|SI_Y) = a \ , \qquad P(D|\overline{S}I_X) = P(D|\overline{S}I_Y) = b \tag{5–23}$$

If $a < b$, then they consider him to be more likely to be telling the truth than lying. But they have different prior probabilities for the safety of the drug:

$$P(S|I_X) = x \ , \qquad P(S|I_Y) = y \ . \tag{5–24}$$

Their posterior probabilities are then

$$P(S|DI_X) = \frac{ax}{ax + b(1-x)} \ , \qquad P(S|DI_Y) = \frac{ay}{ay + b(1-y)} \tag{5–25}$$

from which we see that not only are their opinions always changed in the same direction, on the evidence scale they are always changed by the same amount, $\log(a/b)$:

$$\begin{aligned} \log \frac{P(S|DI_X)}{P(\overline{S}|DI_X)} &= \log \frac{x}{1-x} + \log \frac{a}{b} \\ \log \frac{P(S|DI_Y)}{P(\overline{S}|DI_Y)} &= \log \frac{y}{1-y} + \log \frac{a}{b} \end{aligned} \tag{5–26}$$

But this means that on the probability scale, they can either converge or diverge (Exercise 5.2). These equations correspond closely to those in our sequential widget test in Chapter 4, but have now a different interpretation. If $a = b$, then they consider Mr. N totally unreliable and their views are unchanged by his testimony. If $a > b$, they distrust Mr. N so much that their opinions are driven in the opposite direction from what he intended. Indeed, if $b \to 0$, then $\log a/b \to \infty$; they consider it certain that he is lying, and so they are both driven to complete belief in the safety of the drug: $P(S|DI_X) = P(S|DI_Y) = 1$, independently of their prior probabilities.

Exercise 5.2. From these equations, find the exact conditions on (x, y, a, b) for divergence on the probability scale; that is, $|P(S|DI_X) - P(S|DI_Y)| > |P(S|I_X) - P(S|I_Y)|$.

Exercise 5.3. It is evident from (5–26) that Mr. X and Mr. Y can never experience a reversal of viewpoint; that is, if initially Mr. X believes more strongly than Mr. Y in the safety of the drug, this will remain true whatever the values of a, b. Therefore, a necessary condition for reversal must be that they have different opinions about Mr. N; $a_x \neq a_y$ and/or $b_x \neq b_y$. But this does not prove that reversal is actually possible, so more analysis is needed. If reversal is possible, find a sufficient condition on $(x, y, a_x, a_y, b_x, b_y)$ for this to take place, and illustrate it by a verbal scenario like the above. If it is not possible, prove this and explain the intuitive reason why reversal cannot happen.

We see that divergence of opinions is readily explained by probability theory as logic, and that it is to be expected when persons have widely different prior information. But where was the error in the reasoning that led us to conjecture (5–20)? We committed a subtle form of the Mind Projection Fallacy by supposing that the relation: "D supports S" is an absolute property of the propositions D and S. We need to recognize the relativity of it; whether D does or does not support S depends on our prior information. The same D that supports S for one person may refute it for another. As soon as we recognize this, then we no longer expect anything like (5–20) to hold in general. This error is very common; we shall see another example of it in "Paradoxes of Intuition" below.

Kahneman & Tversky claimed that we are not Bayesians, because in psychological tests people often commit violations of Bayesian principles. However, this claim is seen differently in view of what we have just noted. We suggest that people are reasoning according to a more sophisticated version of Bayesian inference than they had in mind.

This conclusion is strengthened by noting that similar things are found even in deductive logic. Wason & Johnson–Laird (1972) report psychological experiments in which subjects erred systematically in simple tests which amounted to applying a single syllogism. It seems that when asked to test the hypothesis "A implies B", they had a very strong tendency to consider it equivalent to "B implies A" instead of "not–B implies not–A". Even professional logicians could err in this way.*

Strangely enough, the nature of this error suggests a tendency toward Bayesianity, the opposite of the Kahneman–Tversky conclusion. For, if A supports B in the sense that for some X, $P(B|AX) > P(B|X)$, then Bayes' theorem states that B supports A in the same sense: $P(A|BX) > P(A|X)$. But it also states that $P(\overline{A}|\overline{B}X) > P(\overline{A}|X)$, corresponding to the syllogism, and in the limit $P(B|AX) \to 1$, Bayes' theorem does not give $P(A|BX) \to 1$, but gives $P(\overline{A}|\overline{B}X) \to 1$, in agreement with the syllogism, as we noted in Chapter 2.

Errors made in staged psychological tests may indicate only that the subjects were pursuing different goals than the psychologists; they saw the tests as basically foolish, and did not think it worth making any mental effort before replying to the questions – or perhaps even thought that the psychologists would be more pleased to see them answer wrongly. Had they been faced with logically equivalent situations where their interests were strongly involved (for example, avoiding a serious accidental injury), they might have reasoned better. Indeed, there are stronger grounds – Natural Selection – for expecting that we would reason in a basically Bayesian way.

Visual Perception – Evolution into Bayesianity?

Another class of psychological experiments fits nicely into this discussion. In the early 20'th Century, Adelbert Ames Jr. was Professor of Physiological Optics at Dartmouth College. He devised

* A possible complication of these tests – semantic confusion – readily suggests itself. We noted in Chapter 1 that the word "implication" has a different meaning in formal logic than it has in ordinary language; "A implies B" does not have the usual colloquial meaning that B is logically deducible from A, as the subjects may have supposed.

ingenious experiments which fool one into 'seeing' something very different from the reality – one misjudges the size, shape, distance of objects. Some dismissed this as idle optical illusioning, but others who saw these demonstrations – notably including Alfred North Whitehead and Albert Einstein – saw their true importance as revealing surprising things about the mechanism of visual perception.† His work was carried on by Professor Hadley Cantrell of Princeton University, who discussed these phenomena in his book "*The Why of Man's Experience*" and produced movie demonstrations of them.

The brain develops already in infancy certain assumptions about the world based on all the sensory information it receives. For example, nearer objects appear larger, have greater parallax, and occlude distant objects in the same line of sight; a straight line appears straight from whatever direction it is viewed, *etc.* These assumptions are incorporated into the artist's rules of perspective and in 3–d computer graphics programs. We hold tenaciously onto them because they have been successful in correlating many different experiences. We will not relinquish successful hypotheses as long as they work; the only way to make one change these assumptions is to put one in a situation where they don't work. For example, in that Ames room where perceived size and distance correlate in the wrong way; a child in walking across the room doubles in height.

The general conclusion from all these experiments is less surprising to our relativist generation than it was to the absolutist generation which made the discoveries. Seeing is not a direct apprehension of reality, as we often like to pretend. Quite the contrary: *Seeing is Inference from Incomplete Information*, no different in nature from the inference that we are studying here. The information that reaches us through our eyes is grossly inadequate to determine what is "really there" before us. The failures of perception revealed by the experiments of Ames and Cantrell are not mechanical failures in the lens, retina, or optic nerve; they are the reactions of the subsequent inference process in the brain *when it receives new data that are inconsistent with its prior information*. These are just the situations where one is obliged to resurrect some alternative hypothesis; and that is what we "see." We expect that detailed analysis of these cases would show an excellent correspondence with Bayesian inference, in much the same way as in our ESP and diverging opinions examples.

Active study of visual perception has continued, and volumes of new knowledge have accumulated, but we still have almost no conception of how this is accomplished at the level of the neurons. Workers note the seeming absence of any organizing principle; we wonder whether the principles of Bayesian inference might serve as a start. We would expect Natural Selection to produce such a result; after all, any reasoning format whose results conflict with Bayesian inference will place a creature at a decided survival disadvantage. Indeed, as we noted long ago (Jaynes, 1957b), to deny that we reason in a Bayesian way is to assert that we reason in a deliberately inconsistent way; we find this very hard to believe. Presumably, a dozen other examples of human and animal perception would be found to obey a Bayesian reasoning format as its "high level" organizing principle, for the same reason.

With this in mind, let us examine a famous case history.

The Discovery of Neptune

Another potential application for probability theory, which has been discussed vigorously by philosophers for over a century, concerns the reasoning process of a scientist, by which he accepts or rejects his theories in the light of the observed facts. We noted in Chapter 1 that this consists largely of the use of two forms of syllogism,

---

† One of his most impressive demonstrations has been recreated at the *Exploratorium* in San Francisco, the full–sized "Ames room" into which visitors can look to see these phenomena at first hand.

$$\text{one strong}: \quad \left\{ \begin{array}{c} \text{If } A, \text{ then } B \\ B \text{ false} \\ \hline A \text{ false} \end{array} \right\} \qquad \text{and one weak}: \quad \left\{ \begin{array}{c} \text{If } A, \text{ then } B \\ B \text{ true} \\ \hline A \text{ more plausible} \end{array} \right\} \tag{5–27}$$

In Chapter 2 we noted that these correspond to the use of Bayes' theorem in the forms

$$P(A|\overline{B}X) = P(A|X)\,\frac{P(\overline{B}|AX)}{P(\overline{B}|X)}\,, \qquad P(A|BX) = P(A|X)\,\frac{P(B|AX)}{P(B|X)} \tag{5–28}$$

respectively and that these forms do agree qualitatively with the syllogisms.

Interest here centers on the question whether the second form of Bayes' theorem gives a satisfactory quantitative version of the weak syllogism, as scientists use it in practice. Let us consider a specific example given by Pólya (1954; Vol. II, pp. 130–132). This will give us a more useful example of the resurrection of alternative hypotheses.

The planet Uranus was discovered by Wm. Herschel in 1781. Within a few decades (*i.e.,* by the time Uranus had traversed about one third of its orbit), it was clear that it was not following exactly the path prescribed for it by the Newtonian theory (laws of mechanics and gravitation). At this point, a naïve application of the strong syllogism might lead one to conclude that the Newtonian theory was demolished.

However, its many other successes had established the Newtonian theory so firmly that in the minds of astronomers the probability of the hypothesis: "Newton's theory is false" was already down at perhaps -50 *db*. Therefore, for the French astronomer Urbain Jean Joseph Leverrier (1811–1877) and the English scholar John Couch Adams (1819–1892) at St. John's College, Cambridge, an alternative hypothesis down at perhaps -20 *db* was resurrected: there must be still another planet beyond Uranus, whose gravitational pull is causing the discrepancy.

Working unknown to each other and backwards, Leverrier and Adams computed the mass and orbit of a planet which could produce the observed deviation and predicted where the new planet would be found, with nearly the same results. The Berlin observatory received Leverrier's prediction on September 23, 1846, and on the evening of the same day, the astronomer Johann Gottfried Galle (1812–1910) found the new planet (Neptune) within about one degree of the predicted position. For many more details, see Smart (1947) or Grosser (1979).

Instinctively, we feel that the plausibility of the Newtonian theory was increased by this little drama. The question is, how much? The attempt to apply probability theory to this problem will give us a good example of the complexity of actual situations faced by scientists, and also of the caution one needs in reading the rather confused literature on these problems.

Following Pólya's notation, let T stand for the Newtonian theory, N for the part of Leverrier's prediction that was verified. Then probability theory gives for the posterior probability of T,

$$P(T|NX) = P(T|X)\,\frac{P(N|TX)}{P(N|X)}\,. \tag{5–29}$$

Suppose we try to evaluate $P(N|X)$. This is the prior probability of N, regardless of whether T is true or not. As usual, denote the denial of T by $\overline{T}$. Since $N = N(T+\overline{T}) = NT + N\overline{T}$, we have, by applying the sum and product rules

$$\begin{aligned} P(N|X) &= P(NT + N\overline{T}|X) = P(NT|X) + P(N\overline{T}|X) \\ &= P(N|TX)\,P(T|X) + P(N|\overline{T}X)\,P(\overline{T}|X) \end{aligned} \tag{5–30}$$

and $P(N|\overline{T}X)$ has intruded itself into the problem. But in the problem as stated this quantity is not defined; the statement $\overline{T} \equiv$ "Newton's theory is false" has no definite implications until we specify what alternative we have to put in place of Newton's theory.

For example, if there were only a single possible alternative according to which there could be no planets beyond Uranus, then $P(N|\overline{T}X) = 0$, and probability theory would again reduce to deductive reasoning, giving $P(T|NX) = 1$, independently of the prior probability $P(T|X)$.

On the other hand, if Einstein's theory were the only possible alternative, its predictions do not differ appreciably from those of Newton's theory for this phenomenon, and we would have $P(N|\overline{T}X) = P(N|TX)$, whereupon $P(T|NX) = P(T|X)$.

Thus, verification of the Leverrier–Adams prediction might elevate the Newtonian theory to certainty, or it might have no effect at all on its plausibility. It depends entirely on this: *Against which specific alternatives are we testing Newton's theory*?

Now to a scientist who is judging his theories, this conclusion is the most obvious exercise of common sense. We have seen the mathematics of this in some detail in Chapter 4, but all scientists see the same thing intuitively without any mathematics.

For example, if you ask a scientist, "How well did the Zilch experiment support the Wilson theory?" you may get an answer like this: "Well, if you had asked me last week I would have said that it supports the Wilson theory very handsomely; Zilch's experimental points lie much closer to Wilson's predictions than to Watson's. But just yesterday I learned that this fellow Woffson has a new theory based on more plausible assumptions, and his curve goes right through the experimental points. So now I'm afraid I have to say that the Zilch experiment pretty well demolishes the Wilson theory."

Digression on Alternative Hypotheses

In view of this, working scientists will note with dismay that statisticians have developed *ad hoc* criteria for accepting or rejecting theories (Chi–squared test, etc.) which make no reference to any alternatives. A practical difficulty of this was pointed out by Jeffreys (1939); there is not the slightest use in rejecting any hypothesis H_0 unless we can do it in favor of some definite alternative H_1 which better fits the facts.

Of course, we are concerned here with hypotheses which are not themselves statements of observable fact. If the hypothesis H_0 is merely that $x < y$, then a direct, error–free measurement of x and y which confirms this inequality constitutes positive proof of the correctness of the hypothesis, independently of any alternatives. We are considering hypotheses which might be called 'scientific theories' in that they are suppositions about what is not observable directly; only some of their consequences – logical or causal – can be observed by us.

For such hypotheses, Bayes' theorem tells us this: *Unless the observed facts are absolutely impossible on hypothesis H_0, it is meaningless to ask how much those facts tend "in themselves" to confirm or refute H_0.* Not only the mathematics, but also our innate common sense (if we think about it for a moment) tells us that we have not asked any definite, well–posed question until we specify the possible alternatives to H_0. Then as we saw in Chapter 4, probability theory can tell us how our hypothesis fares *relative to the alternatives that we have specified*; it does not have the creative imagination to invent new hypotheses for us.

Of course, as the observed facts approach impossibility on hypothesis H_0, we are led to worry more and more about H_0; but mere improbability, however great, cannot in itself be the reason for doubting H_0. We almost noted this after Eq. (5–7); now we are laying stress on it because it will be essential for our later general formulation of significance tests.

Early attempts to devise such tests foundered on the point we are making. John Arbuthnot (1710) noted that in 82 years of demographic data more boys than girls were born in every year.

On the "null hypothesis" H_0 that the probability of a boy is 1/2, he considered the probability of this result to be $2^{-82} = 10^{-24.7}$ [in our measure, $-247\,db$], so small as to make H_0 seem to him virtually impossible, and saw in this evidence for "Divine Providence". He was, apparently, the first person to reject a statistical hypothesis on the grounds that it renders the data improbable. However, we can criticize his reasoning on several grounds.

Firstly, the alternative hypothesis $H_1 \equiv$ "Divine Providence" does not seem usable in a probability calculation because it is not specific. That is, it does not make any definite predictions known to us, and so we cannot assign any probability for the data $P(D|H_1)$ conditional on H_1. [For this same reason, the mere logical denial $H_1 \equiv \overline{H}_0$ is unusable as an alternative.] In fact, it is far from clear why Divine Providence would wish to generate more boys than girls; indeed, if the number of boys and girls were exactly equal every year in a large population, that would seem to us much stronger evidence that some supernatural control mechanism must be at work.

Secondly, Arbuthnot's data told him not only the number N of years with more boys than girls, but also in *which years* each possibility occurred. So whatever the observed N, on the null hypothesis the probability $P(D|H_0)$ of finding the observed sequence would have been just as small, so by his reasoning the hypothesis would have been rejected whatever the data. Furthermore, had he calculated the probability of the actual data instead of merely aggregating them into two bins labelled "more boys than girls" and "more girls than boys", he would have found a probability very much smaller still, whatever the actual data, so the mere smallness of the probability is not in itself the significant thing.

As a simple, but numerically stronger example illustrating this, if we toss a coin 1000 times, then no matter what the result is, the specific observed sequence of heads and tails has a probability of only 2^{-1000}, or $-3010\ db$, on the hypothesis that the coin is honest. If, after having tossed it 1000 times, we still believe that the coin is honest, it can be only because the observed sequence is even more improbable on any alternative hypothesis that we are willing to consider seriously.

Without having the probability $P(D|H_1)$ of the data on the alternative hypothesis *and* the prior probabilities of the hypotheses, there is no well–posed problem and just no rational basis for passing judgment. However, it is mathematically trivial to see that those who fail to introduce prior probabilities at all are led thereby, automatically, to the results that would follow from assigning equal prior probabilities. Failure to mention prior probabilities is so common that we overlook it and suppose that the author intended equal prior probabilities (if he should later deny this, then he can be charged with giving an arbitrary solution to an undefined problem).

Finally, having observed more boys than girls for ten consecutive years, rational inference might have led Arbuthnot to anticipate it for the eleventh year. Thus his hypothesis H_0 was not only the numerical value $p = 1/2$; there was also an implicit assumption of logical independence of different years, of which he was probably unaware. On an hypothesis that allows for positive correlations, for example H_{ex} which assigns an exchangeable sampling distribution, the probability $P(D|H_{ex})$ of the aggregated data could be very much greater than 2^{-82}. Thus Arbuthnot took one step in the right direction, but to get a usable significance test required a conceptual understanding of probability theory on a considerably higher level, as achieved by Laplace some 100 years later.

Another example occurred when Daniel Bernoulli won a French Academy prize of 1734 with an essay on the orbits of planets, in which he represented the orientation of each orbit by its polar point on the unit sphere and found them so close together as to make it very unlikely that the present distribution could result by chance. Although he too failed to state a specific alternative, we are inclined to accept his conclusion today because there seems to be a very clearly implied null hypothesis H_0 of "chance" according to which the points should appear spread all over the sphere with no tendency to cluster together; and H_1 of "attraction", which would make them tend to coincide; the evidence rather clearly supported H_1 over H_0.

Laplace (1812) did a similar analysis on comets, found their polar points much more scattered than those of the planets, and concluded that comets are not "regular members" of the solar system like the planets. Here we finally had two fairly well–defined hypotheses being compared by a correct application of probability theory. (It is one of the tragedies of history that Cournot (1843), failing to comprehend Laplace's rationale, attacked it and reinstated the errors of Arbuthnot, thereby dealing scientific inference a setback from which it is not yet fully recovered).

Such tests need not be quantitative. Even when the application is only qualitative, probability theory is still useful to us in a normative sense; it is the means by which we can detect inconsistencies in our own qualitative reasoning. It tells us immediately what has not been intuitively obvious to all workers: that alternatives are needed before we have any rational criterion for testing hypotheses.

This means that if any significance test is to be acceptable to a scientist, we shall need to examine its rationale to see whether it has, like Daniel Bernoulli's test, some implied if unstated alternative hypotheses. Only when such hypotheses are identified are we in a position to say what the test accomplishes; *i.e.* what it is testing. But not to keep the reader in suspense: a statisticians' formal significance test can always be interpreted as a test of a specified hypothesis H_0 against a specified *class* of alternatives, and thus it is only a mathematical generalization of our treatment of multiple hypothesis tests in Chapter 4, Equations (4–28) – (4–44). However, the standard orthodox literature, which dealt with composite hypotheses by applying arbitrary *ad hockeries* instead of probability theory, never perceived this.

Back to Newton

Now we want to get a quantitative result about Newton's theory. In Pólya's discussion of the feat of Leverrier and Adams, once again no specific alternative to Newton's theory is stated; but from the numerical values used (*loc. cit*, p. 131) we can infer that he had in mind a single possible alternative H_1 according to which it was known that one more planet existed beyond Uranus, but all directions on the celestial sphere were considered equally likely. Then, since a cone of angle 1 degree fills in the sky a solid angle of about $\pi/(57.3)^2 = 10^{-3}$ steradians, $P(N|H_1X) \simeq 10^{-3}/4\pi = 1/13,000$ is the probability that Neptune would have been within one degree of the predicted position.

Unfortunately, in the calculation no distinction was made between $P(N|X)$ and $P(N|\overline{T}X)$; that is, instead of the calculation (5–18) indicated by probability theory, the likelihood ratio actually calculated by Pólya was, in our notation,

$$\frac{P(N|TX)}{P(N|\overline{T}X)} = \frac{P(N|TX)}{P(N|H_1X)} \tag{5–31}$$

Therefore, according to the analysis in Chapter 4, what Pólya obtained was not the ratio of posterior to prior probabilities, but the ratio of posterior to prior odds:

$$\frac{O(N|TX)}{O(N|X)} = \frac{P(N|TX)}{P(N|\overline{T}X)} = 13,000. \tag{5–32}$$

The conclusions are much more satisfactory when we notice this. Whatever prior probability $P(T|X)$ we assign to Newton's theory, if H_1 is the only alternative considered, then verification of the prediction increased the *evidence* for Newton's theory by $10\log_{10}(13,000) = 41$ decibels.

Actually, if there were a new planet it would be reasonable, in view of the aforementioned investigations of Daniel Bernoulli and Laplace, to adopt a different alternative hypothesis H_2, according to which its orbit would lie in the plane of the ecliptic, as Pólya again notes by implication rather than explicit statement. If, on hypothesis H_2, all values of longitude are considered equally likely, we might reduce this to about $10\log_{10}(180) = 23$ decibels. In view of the great uncertainty as to just what the alternative is (*i.e.*, in view of the fact that the problem has not been defined unambiguously), any value between these extremes seems more or less reasonable.

There was a difficulty which bothered Pólya: if the *probability* of Newton's theory were increased by a factor of 13,000, then the prior probability was necessarily lower than (1/13,000); but this contradicts common sense, because Newton's theory was already very well established before Leverrier was born. Pólya interprets this in his book as revealing an inconsistency in Bayes' theorem, and the danger of trying to apply it numerically. Recognition that we are, in the above numbers, dealing with odds rather than probabilities, removes this objection and makes Bayes' theorem appear quite satisfactory in describing the inferences of a scientist.

This is a good example of the way in which objections to the Bayes–Laplace methods which you find in the literature, disappear when you look at the problem more carefully. By an unfortunate slip in the calculation, Pólya was led to a misunderstanding of how Bayes' theorem operates. But I am glad to be able to close the discussion of this incident with a happier personal reminiscence.

In 1956, two years after the appearance of Pólya's work, I gave a series of lectures on these matters at Stanford University, and George Pólya attended them, sitting in the first row and paying the most strict attention to everything that was said. By then he understood this point very well – indeed, whenever a question was raised from the audience, Pólya would turn around and give the correct answer, before I could. It was very pleasant to have that kind of support, and I miss his presence today (George Pólya died, at the age of 97, in September 1985).

But the example also shows clearly that in practice the situation faced by the scientist is so complicated that there is little hope of applying Bayes' theorem to give quantitative results about the relative status of theories. Also there is no need to do this, because the real difficulty of the scientist is not in the reasoning process itself; his common sense is quite adequate for that. The real difficulty is in learning how to formulate new alternatives which better fit the facts. Usually, when one succeeds in doing this, the evidence for the new theory soon becomes so overwhelming that nobody needs probability theory to tell him what conclusions to draw.

At present we have no formal theory at all on the process of "optimal hypothesis formulation" and we are dependent entirely on the creative imagination of individual persons like Newton, Mendel, Einstein, Wegener, Crick. So, we would say that *in principle* the application of Bayes' theorem in the above way is perfectly legitimate; but *in practice* it is of very little use to a scientist.

However, we should not presume to give quick, glib answers to deep questions. The question of exactly how scientists do, in practice, pass judgment on their theories, remains complex and not well analyzed. Further comments on the validity of Newton's theory are offered at the end of this Chapter.

Exercise 5.4. Our story has a curious sequel. In turn it was noticed that Neptune was not following exactly its proper course, and so one naturally assumed that there is still another planet causing this. Percival Lowell, by a similar calculation, predicted its orbit and Clyde Tombaugh proceeded to find the new planet (Pluto), although not so close to the predicted position. But now the story changes: modern data on the motion of Pluto's moon indicate that the mass of Pluto is too small to have caused the perturbation of Neptune which motivated Lowell's calculation. Today the discrepancies in the motions of Neptune and Pluto are still unaccounted for (We are indebted to Dr. Brad Schaefer for this information). Try to extend our probability analysis to take this new circumstance into account; now where does Newton's theory stand? For more background information, see Whyte (1980) or Hoyt (1980).

Horseracing and Weather Forecasting

The above examples noted two different features common in problems of inference; (a) As in the ESP and psychological cases, the information we receive is often not a direct proposition like S in (5-21); it is an indirect claim that S is true, from some "noisy" source that is itself not wholly reliable; and (b) As in the example of Neptune, there is a long tradition of writers who have misapplied Bayes' theorem and concluded that Bayes' theorem is at fault. Both features are present simultaneously in a work of the Princeton philosopher Richard C. Jeffrey (1983), hereafter denoted by RCJ to avoid confusion with the Cambridge scholar Sir Harold Jeffreys.

RCJ considers the following problem. With only prior information I, we assign a probability $P(A|I)$ to A. Then we get new information B, and it changes as usual via Bayes' theorem to

$$P(A|BI) = P(A|I)\, P(B|AI)/P(B|I)\,. \tag{5–33}$$

But then he decides that Bayes' theorem is not sufficiently general, because we often receive new information that is not certain; perhaps the probability of B is not unity but, say, q. To this we would reply: "If you do not accept B as true, then why are you using it in Bayes' theorem this way?" But RCJ follows that long tradition and concludes, not that it is a misapplication of Bayes' theorem to use uncertain information as in (5–33), but that Bayes' theorem is itself faulty, and it needs to be generalized to take the uncertainty of new information into account.

His proposed generalization (denoting the denial of B by $\overline{B}$) is that the updated probability of A should be taken as a weighted average:

$$P(A)_J = q\, P(A|BI) + (1-q)\, P(A|\overline{B}I) \tag{5–34}$$

But this is an adhockery that does not follow from the rules of probability theory unless we take q to be the *prior* probability $P(B|I)$, just the case that RCJ excludes [for then $P(A)_J = P(A|I)$ and there is no updating].

Since (5–34) conflicts with the rules of probability theory, we know that it necessarily violates one of the desiderata that we discussed in Chapters 1 and 2. The source of the trouble is easy to find, because those desiderata tell us where to look. The proposed 'generalization' (5–34) cannot hold generally because we could learn many different things, all of which indicate the same probability q for B; but which have different implications for A. Thus (5–34) violates desideratum (1–23b); it cannot take into account all of the new information, only the part of it that involves (*i.e.*, is relevant to) B.

The analysis of Chapter 2 tells us that, if we are to salvage things and recover a well–posed problem with a defensible solution, *we must not depart in any way from Bayes' theorem.* Instead, we need to recognize the same thing that we stressed in the ESP example; if B is not known with certainty to be true, then B could not have been the new information; the actual information received must have been some proposition C such that $P(B|CI) = q$. But then, of course, we should be considering Bayes' theorem conditional on C, rather than B:

$$P(A|CI) = P(A|I)\, P(C|AI)/P(C|I) \tag{5–35}$$

If we apply it properly, Bayes' theorem automatically takes the uncertainty of new information into account. This result can be written, using the product and sum rules of probability theory, as

$$\begin{aligned} P(A|CI) &= P(AB|CI) + P(A\overline{B}|CI) \\ &= P(A|BCI)\, P(B|CI) + P(A|\overline{B}CI)\, P(\overline{B}|CI) \end{aligned}$$

and if we define $q \equiv P(B|CI)$ to be the updated probability of B, this can be written in the form

$$P(A|CI) = q\, P(A|BCI) + (1-q)\, P(A|\overline{B}CI) \tag{5–36}$$

which resembles (5–34) but is not in general equal to it, unless we add the restriction that the probabilities $P(A|BCI)$ and $P(A|\overline{B}CI)$ are to be independent of C. Intuitively, this would mean that the logic flows thus:

$$(C \rightarrow B \rightarrow A) \qquad \text{rather than} \qquad (C \rightarrow A)\,, \tag{5–37}$$

That is, C is relevant to A only through its intermediate relevance to B (C is relevant to B and B is relevant to A).

RCJ shows by example that this logic flow may be present in a real problem, but fails to note that his proposed solution (5–34) is then the same as the Bayesian result. Without that logic flow, (5–34) will be unacceptable in general because it does not take into account all of the new information. The information which is lost is indicated by the lack of an arrow going directly ($C \rightarrow A$) in the logic flow diagram; information in C which is directly relevant to A, whether or not B is true.

If we think of the logic flow as something like the flow of light, we might visualize it thus: at night we receive sunlight only through its intermediate reflection from the moon; this corresponds to the RCJ solution. But in the daytime we receive light directly from the sun whether or not the moon is there; this is what the RCJ solution has missed. (In fact, when we study the maximum entropy formalism in statistical mechanics and the phenomenon of "generalized scattering", we shall find that this is more than a loose analogy; the process of conditional information flow is in almost exact mathematical correspondence with the Huygens principle of optics.)

Exercise 5.5. We might expect intuitively that when $q \rightarrow 1$ this difference would disappear; *i.e.*, $P(A|BI) \rightarrow P(A|CI)$. Determine whether this is or is not generally true. If it is, indicate how small $1 - q$ must be in order to make the difference practically negligible. If it is not, illustrate by a verbal scenario the circumstances which can prevent this agreement.

We can illustrate this in a more down–to–earth way by one of RCJ's own scenarios:

A = "My horse will win the race tomorrow",
B = "The track will be muddy",
I = "Whatever I know about my horse and jockey in particular, and about horses, jockeys, races, and life in general."

and the probability $P(A|I)$ gets updated as a result of receiving a weather forecast. Then some proposition C such as:

C = "The TV weather forecaster showed us today's weather map, quoted some of the current meteorological data, and then by means unexplained assigned probability q' to rain tomorrow."

is clearly present, but it is not recognized and stated by RCJ. Indeed, to do so would introduce much new detail, far beyond the ambit of propositions (A, B) of interest to horse racers.

If we recognize proposition C explicitly, then we must recall everything we know about the process of weather forecasting, what were the particular meteorological data leading to that forecast, how reliable weather forecasts are in the presence of such data, how the officially announced probability q' is related to what the forecaster really believes (*i.e.*, what we think the forecaster perceives his own interest to be), *etc., etc.*

If the above–defined C is the new information, then we must consider also, in the light of all our prior information, how C might affect the prospects for the race A through other circumstances than the muddiness B of the track; perhaps the jockey is blinded by bright sunlight, perhaps the rival horse runs poorly on cloudy days, whether or not the track is wet. These would be logical relations of the form ($C \rightarrow A$) that (5–34) cannot take into account.

Therefore the full solution must be vastly more complicated than (5–34); but this is, of course, as it should be. Bayes' theorem, as always, is only telling us what common sense does; in general the updated probability of A must depend on far more than just the updated probability q of B.

Discussion

This example illustrates what we have noted before in Chapter 1; that familiar problems of everyday life may be more complicated than scientific problems, where we are often reasoning about carefully controlled situations. The most familiar problems may be so complicated – just because the result depends on so many unknown and uncontrolled factors – that a full Bayesian analysis, although correct in principle, is out of the question in practice. The cost of the computation is far more than we could hope to win on the horse.

Then we are necessarily in the realm of approximation techniques; but since we cannot apply Bayes' theorem exactly, need we still consider it at all? Yes, because Bayes' theorem remains the normative principle telling us what we should aim for. Without it, we have nothing to guide our choices and no criterion for judging their success.

It also illustrates what we shall find repeatedly in later Chapters; generations of workers in this field have not comprehended the fact that Bayes' theorem is *a valid theorem*, required by elementary desiderata of rationality and consistency, and have made unbelievably persistent attempts to replace it by all kinds of intuitive *ad hockeries*. Of course, we expect that any sincere intuitive effort will capture bits of the truth; yet all of these dozens of attempts have proved on analysis to be satisfactory only in those cases where they agree with Bayes' theorem after all.

But we are at a loss to understand what motivates these anti–Bayesian efforts, because we can see nothing unsatisfactory about Bayes' theorem, either in its theoretical foundations, its intuitive rationale, or its pragmatic results. The writer has devoted some 40 years to the analysis of thousands of separate problems by Bayes' theorem, and is still being impressed by the beautiful and important results it gives us, often in a few lines, and far beyond what those *ad hockeries* can produce. We have yet to find a case where it yields an unsatisfactory result (although the result is often surprising at first glance and it requires some meditation to educate our intuition and see that it is correct after all).

Needless to say, the cases where we are at first surprised are just the ones where Bayes' theorem is most valuable to us; because those are the cases where intuitive *ad hockeries* would never have found the result. Comparing Bayesian analysis with the *ad hoc* methods which saturate the literature, whenever there is any disagreement in the final conclusions, we have found it easy to exhibit the defect of the *ad hockery*, just as the analysis of Chapter 2 led us to expect and as we saw in the above example.

In the past, many man–years of effort were wasted in futile attempts to square the circle; had Lindemann's theorem (that π is transcendental) been known and its implications recognized, all of this might have been averted. Likewise, had Cox's theorems been known, and their implications recognized, 100 years ago, many wasted careers might have been turned instead to constructive activity. This is our answer to those who have suggested that Cox's theorems are unimportant, because they only confirm what James Bernoulli and Laplace had conjectured long before.

Today, we have five decades of experience confirming what Cox's theorems tell us. It is clear that, not only is the quantitative use of the rules of probability theory as extended logic the only sound way to conduct inference; it is the *failure* to follow those rules strictly that has for many years been leading to unnecessary errors, paradoxes, and controversies.

Paradoxes of Intuition

A famous example of this situation, known as Hempel's paradox, starts with the premise: "A case of an hypothesis supports the hypothesis". Then it observes: "Now the hypothesis that all crows are black is logically equivalent to the statement that all non–black things are non–crows, and this is supported by the observation of a white shoe". An incredible amount has been written about this seemingly innocent argument, which leads to an intolerable conclusion.

But the error in the argument is apparent at once when one examines the equations of probability theory applied to it: the premise, which was not derived from any logical analysis, is not generally true, and he prevents himself from discovering that fact by trying to judge support of an hypothesis without considering any alternatives.

I. J. Good (1967), in a note entitled "The White Shoe is a Red Herring", demonstrated the error in the premise by a simple counterexample: In World 1 there are one million birds, of which 100 are crows, all black. In World 2 there are two million birds, of which 200,000 are black crows and 1,800,000 are white crows. We observe one bird, which proves to be a black crow. Which world are we in?

Evidently, observation of a black crow gives evidence of

$$10\ \log_{10}\left(\frac{200,000/2,000,000}{100/1,000,000}\right) = 30\ db$$

or an odds ratio of 1000:1, *against* the hypothesis that all crows are black; that is, for World 2 against World 1. Whether an "instance of an hypothesis" does or does not support the hypothesis depends on the alternatives being considered and on the prior information. We learned this in finding the error in the reasoning leading to (5–20). But incredibly, Hempel (1967) proceeded to reject Good's clear and compelling argument on the grounds that it was unfair to introduce that background information about Worlds 1 and 2.

In the literature there are perhaps a hundred "paradoxes" and controversies which are like this, in that they arise from faulty intuition rather than faulty mathematics. Someone asserts a general principle that seems to him intuitively right. Then when probability analysis reveals the error, instead of taking this opportunity to educate his intuition, he reacts by rejecting the probability analysis. We shall see several more examples of this in later Chapters.

As a colleague of the writer once remarked, "Philosophers are free to do whatever they please, because they don't have to do anything right". But a responsible scientist does not have that freedom; he will not assert the truth of a general principle, and urge others to adopt it, merely on the strength of his own intuition. Some outstanding examples of this error, which are not mere philosophers' toys like the RCJ tampering with Bayes' theorem and the Hempel paradox, but have been actively harmful to Science and Society, are discussed in Chapters 15 and 17.

Bayesian Jurisprudence

It is interesting to apply probability theory in various situations in which we can't always reduce it to numbers very well, but still it shows automatically what kind of information would be relevant to help us do plausible reasoning. Suppose someone in New York City has committed a murder, and you don't know at first who it is, but you know that there are 10 million people in New York City. On the basis of no knowledge but this, $e(Guilty|X) = -70\ db$ is the plausibility that any particular person is the guilty one.

How much positive evidence for guilt is necessary before we decide that some man should be put away? Perhaps $+40\ db$, although your reaction may be that this is not safe enough, and the number ought to be higher. If we raise this number we give increased protection to the innocent, but at the cost of making it more difficult to convict the guilty; and at some point the interests of society as a whole cannot be ignored.

For example, if a thousand guilty men are set free, we know from only too much experience that two or three hundred of them will proceed immediately to inflict still more crimes upon society, and their escaping justice will encourage a hundred more to take up crime. So it is clear that the damage to society as a whole caused by allowing a thousand guilty men to go free, is far greater than that caused by falsely convicting one innocent man.

If you have an emotional reaction against this statement, I ask you to think: if you were a judge, would you rather face one man whom you had convicted falsely; or a hundred victims of crimes that you could have prevented? Setting the threshold at +40 *db* will mean, crudely, that on the average not more than one conviction in ten thousand will be in error; a judge who required juries to follow this rule would probably not make one false conviction in a working lifetime on the bench.

In any event, if we took +40 *db* starting out from −70, this means that in order to get conviction you would have to produce about 110 *db* of evidence for the guilt of this particular person. Suppose now we learn that this person had a motive. What does that do to the plausibility of his guilt? probability theory says

$$\begin{aligned} e(Guilty|Motive) &= e(Guilty|X) + 10\log_{10}\frac{P(Motive|Guilty)}{P(Motive|Not\ Guilty)} \\ &\simeq -70 - 10\log_{10} P(Motive|Not\ Guilty) \end{aligned} \tag{5–38}$$

since $P(Motive|Guilty) \simeq 1$, *i.e.,* we consider it quite unlikely that the crime had no motive at all. Thus, the significance of learning that the person had a motive depends almost entirely on the probability $P(Motive|Not\ Guilty)$ that an innocent person would also have a motive.

This evidently agrees with our common sense, if we ponder it for a moment. If the deceased were kind and loved by all, hardly anyone would have a motive to do him in. Learning that, nevertheless, our suspect *did* have a motive, would then be very significant information. If the victim had been an unsavory character, who took great delight in all sorts of foul deeds, then a great many people would have a motive, and learning that our suspect was one of them, is not so significant. The point of this is that we don't know what to make of the information that our suspect had a motive, unless we also know something about the character of the deceased. But how many members of juries would realize that, unless it was pointed out to them?

Suppose that a very enlightened judge, with powers not given to judges under present law, had perceived this fact and, when testimony about the motive was introduced, he directed his assistants to determine for the jury the *number* of people in New York City who had a motive. This number was N_m. Then

$$P(Motive|Not\ Guilty) = \frac{N_m - 1}{(Number\ of\ people\ in\ New York) - 1} \simeq 10^{-7}(N_m - 1)$$

and equation (5–38) reduces, for all practical purposes, to

$$e(Guilty|Motive) \simeq -10\log(N_m - 1) \tag{5–39}$$

You see that the population of New York has cancelled out of the equation; as soon as we know the number of people who had a motive, then it doesn't matter any more how large the city was. Note that (5–39) continues to say the right thing even when N_m is only 1 or 2.

You can go on this way for a long time, and we think you will find it both enlightening and entertaining to do so. For example, we now learn that the suspect was seen near the scene of the crime shortly before. From Bayes' theorem, the significance of this depends almost entirely on how many innocent persons were also in the vicinity. If you have ever been told not to trust Bayes' theorem, you should follow a few examples like this a good deal further, and see how infallibly it tells you what information would be relevant, what irrelevant, in plausible reasoning.†

Even in situations where we would be quite unable to say that numerical values should be used, Bayes' theorem still reproduces qualitatively just what your common sense (after perhaps some meditation) tells you. This is the fact that George Pólya demonstrated in such exhaustive detail that the present writer was convinced that the connection must be more than qualitative.

COMMENTS

There has been much more discussion of the status of Newton's theory than we indicated above. For example, it has been suggested by Charles Misner that we cannot apply a theory with full confidence until we know its limits of validity – where it fails.

Thus relativity theory, in showing us the limits of validity of Newtonian mechanics, also confirmed its accuracy within those limits; so it should increase our confidence in Newtonian theory when applied within its proper domain (velocities small compared to that of light). Likewise, the first law of thermodynamics, in showing us the limits of validity of the caloric theory, also confirmed the accuracy of the caloric theory within its proper domain (processes where heat flows but no work is done). At first glance this seems an attractive idea, and perhaps this is the way scientists really should think.

Nevertheless, Misner's principle contrasts strikingly with the way scientists actually do think. We know of no case where anyone has avowed that his confidence in a theory was increased by its being, as we say, "overthrown". Furthermore, we apply the principle of conservation of momentum with full confidence, not because we know its limits of validity, but for just the opposite reason; we do not know of any such limits. Yet scientists believe that the principle of momentum conservation has real content; it is not a mere tautology.

Not knowing the answer to this riddle, we pursue it only one step further, with the observation that if we are trying to judge the validity of Newtonian mechanics, we cannot be sure that relativity theory showed us all its limitations. It is conceivable, for example, that it may fail not only in the limit of high velocities, but also in that of high accelerations. Indeed, there are theoretical reasons for expecting this; for Newton's $F = ma$ and Einstein's $E = mc^2$ can be combined into a perhaps more fundamental statement:

$$F = (E/c^2)\, a\ . \qquad (5\text{–}40)$$

Why should the force required to accelerate a bundle of energy E depend on the velocity of light?

We see a plausible reason at once, if we adopt the – almost surely true – hypothesis that our allegedly "elementary" particles cannot occupy mere mathematical points in space, but are extended structures of some kind. Then the velocity of light determines how rapidly different parts of the structure can "communicate" with each other. The more quickly all parts can learn that a

† Note that in these cases we are trying to decide, from scraps of incomplete information, on the truth of an Aristotelian proposition; whether the defendant did or did not commit some well–defined action. This is the situation – an issue of fact – for which probability theory as logic is designed. But there are other legal situations quite different; for example, in a medical malpractice suit it may be that all parties are agreed on the facts as to what the defendant actually did; the issue is whether he did or did not exercise reasonable judgment. Since there is no official, precise definition of "reasonable judgment", the issue is not the truth of an Aristotelian proposition (however, if it were established that he wilfully violated one of our Chapter 1 desiderata of rationality, we think that most juries would convict him). It has been claimed that probability theory is basically inapplicable to such situations, and we are concerned with the partial truth of a non–Aristotelian proposition. We suggest, however, that in such cases we are not concerned with an issue of truth at all; rather, what is wanted is a value judgment. We shall return to this topic later (Chapters 13, 18; Appendix I).

force is being applied, the more quickly they can all respond to it. We leave it as an exercise for the reader to show that one can actually derive Eq. (5–40) from this premise (Hint: the force is proportional to the deformation that the particle must suffer before all parts of it start to move).

But this embryonic theory makes further predictions immediately. We would expect that when a force is applied suddenly, a short transient response time would be required for the acceleration to reach its Newtonian value. If so, then Newton's $F = ma$ is not an exact relation, only a final steady state condition, approached after the time required for light to cross the structure. It is conceivable that such a prediction could be tested experimentally.

Thus the issue of our confidence in Newtonian theory is vastly more subtle and complex than merely citing its past predictive successes and its relation to relativity theory; it depends also on our whole theoretical outlook.

It appears to us that actual scientific practice is guided by instincts that have not yet been fully recognized, much less analyzed and justified. We must take into account not only the logic of science, but also the sociology of science (perhaps also its soteriology). But this is so complicated that we are not even sure whether the extremely skeptical conservatism with which new ideas are invariably received, is in the long run a beneficial stabilizing influence, or a harmful obstacle to progress.

What is Queer? In this Chapter we have examined some applications of probability theory that seem "queer" to us today, in the sense of being "off the beaten track". Any completely new application must presumably pass through such an exploratory phase of queerness. But in many cases, particularly the Bayesian jurisprudence and psychological tests with a more serious purpose than ESP, we think that queer applications of today may become respectable and useful applications of tomorrow. Further thought and experience will make us more aware of the proper formulation of a problem – better connected to reality – and then future generations will come to regard Bayesian analysis as indispensable for discussing it. Now we return to the many applications that are already advanced beyond the stage of queerness, into that of respectability and usefulness.