CHAPTER 13

DECISION THEORY – HISTORICAL BACKGROUND

"Your act was unwise," I exclaimed "as you see
by the outcome." He solemnly eyed me.
"When choosing the course of my action," said he,
"I had not the outcome to guide me."

— Ambrose Bierce

In several previous discussions we inserted parenthetic remarks to the effect that "there is still an essential point missing here, which will be supplied when we take up decision theory." However, in postponing the topic until now we have not deprived the reader of a needed technical tool, because the solution of the decision problem was, from our viewpoint, so immediate and intuitive that we did not need to invoke any underlying formal theory.

Inference vs. Decision

This situation arose as soon as we started applying probability theory to our first problem. When we illustrated the use of Bayes' theorem by sequential testing in Chapter 4, we noted that there is nothing in probability theory *per se* which could tell us where to put the critical levels at which the robot changes its decision: whether to accept the batch, reject it, or make another test. The location of these critical levels obviously depends in some way on value judgments as well as on probabilities; what are the consequences of making wrong decisions, and what are the costs of making further tests?

The same situation occurred in Chapter 6 when the robot was faced with the job of estimating a parameter. Probability theory determined only the robot's state of knowledge about the parameter; it did not tell us what estimate it should in fact make. We noted at that time that taking the mean value over the posterior *pdf* was the same as making that decision which minimizes the expected square of the error; but we noted also that in some cases we should really prefer the median.

Qualitatively and intuitively, these considerations are clear enough; but before we can claim to have a really complete design for our robot, we must clean up the logic of this, and show that our procedures were not just intuitive *ad hockeries*, but were optimal by some clearly defined criterion of optimality. Wald's decision theory aims to accomplish this.

A common feature of all the problems considered thus far was: probability theory alone can solve only the inference problem; *i.e.,* it can give us only a probability distribution which represents the robot's final state of knowledge with all the available prior information and data taken into account; but in practice its job does not end at that point. *An essential thing which is still missing in our design of this robot is the rule by which it converts its final probability assignment into a definite course of action.* But for us, formal decision theory will only legitimize – not change – what our intuition has already told us to do.

Decision theory has for us a different kind of importance, in the light that it sheds on the centuries–old controversies about the foundations of probability theory. The procedures of decision theory can be derived equally well from either of two diametrically opposed views about what probability theory is, and it therefore forms a kind of bridge between them and suggests that decision theory might help to resolve the controversy. Therefore, we dwell here on the historical background and relation of the two approaches to the decision problem.

Daniel Bernoulli's Suggestion

As one might expect from the way this situation appeared in the most elementary applications of probability theory, this problem is by no means new. It was clearly recognized, and a definite solution offered for a certain class of problems, by Daniel Bernoulli (1738). In a crude form, the same principle had been seen even earlier, at the time when probability theory was concerned almost exclusively with problems of gambling. Although today it seems very hard to understand, the historical record shows clearly and repeatedly that the notion of "expectation of profit" was very intuitive to the first workers in probability theory; even more intuitive than that of probability.

Consider each possibility, $i = 1, 2, ..., n$, assign probabilities p_i to them, and also assign numbers M_i which represent the "profit" we would obtain if the i'th possibility should in fact turn out to be true. Then the expectation of profit is, in either of our standard notations,

$$E(M) = \langle M \rangle = \sum_{i=1}^{n} p_i \, M_i \tag{13–1}$$

The prosperous merchants in 17'th Century Amsterdam bought and sold expectations as if they were tangible goods. It seemed obvious to many that a person acting in pure self–interest should always behave in such a way as to maximize his expected profit. This, however, led to some paradoxes (particularly that of the famous St. Petersburg problem) which led Bernoulli to recognize that simple expectation of profit is not always a sensible criterion of action.

For example, suppose that your information leads you to assign probability 0.51 to heads in a certain slightly biased coin. Now you are given the choice of two actions: (1) to bet every cent you have at even money, on heads for the next toss of this coin; (2) not to bet at all. According to the criterion of expectation of profit, you should always choose to gamble when faced with this choice. Your expectation of profit, if you do not gamble, is zero; but if you do gamble, it is

$$\langle M \rangle = 0.51 \, M_0 + 0.49 \, (-M_0) = 0.02 \, M_0 > 0 \tag{13–2}$$

where M_0 is the amount you have now. Nevertheless it seemed obvious to Bernoulli, as it doubtless does also to the reader, that nobody in his right mind would really choose the first alternative. This means that our common sense, in some cases, rejects the criterion of maximizing expected profit.

Suppose that you are offered the following opportunity. You can bet any amount you want on the basis that, with probability $(1 - 10^{-6})$, you will lose your money; but with probability 10^{-6}, you will win 1,000,001 times the amount you had wagered. Again, the criterion of maximizing expected profit says that you should bet all the money you have. Common sense rejects this solution even more forcefully.

Daniel Bernoulli proposed to resolve these paradoxes by recognition that the true value to a person, of receiving a certain amount of money, is not measured simply by the amount received; it depends also upon how much he has already. In other words, Bernoulli said that we should recognize that the mathematical expectation of profit is not the same thing as its "moral expectation." A modern economist is expressing the same idea when he speaks of the "diminishing marginal utility of money."

In the St. Petersburg game we toss an honest coin until it comes up heads for the first time. The game is then terminated. If heads occurs for the first time at the n'th throw, the player receives 2^n dollars. The question is: what is a "fair" entrance fee for him to pay, for the privilege of playing this game? If we use the criterion that a fair game is one where the entrance fee is equal to the expectation of profit, you see what happens. This expectation is infinite:

$$\sum_{k=1}^{\infty}(2^{-k})(2^{k}) = \sum_{k=1}^{\infty} 1 = \infty. \tag{13–3}$$

Nevertheless it is clear again that no sane person would be willing to risk more than a very small amount on this game. We quote Laplace (1819) at this point:

> "Indeed, it is apparent that one franc has much greater value for him who possesses only 100 than for a millionaire. We ought then to distinguish the absolute value of the hoped–for benefit from its relative value. The latter is determined by the motives which make it desirable, whereas the first is independent of them. The general principle for assigning this relative value cannot be given, but here is one proposed by Daniel Bernoulli which will serve in many cases: The relative value of an infinitely small sum is equal to its absolute value divided by the total fortune of the person interested."

In other words, Bernoulli proposed that the "moral value," or what the modern economist would call the "utility" of an amount M of money should be taken proportional to $\log M$. Laplace, in discussing the St. Petersburg problem and this criterion, reports the following result without giving the calculation: a person whose total fortune is 200 francs ought not reasonably to stake more than 9 francs on the play of this game. Let us, 180 years later, check Laplace's calculation.

For a person whose initial "fortune" is m francs, the fair fee $f(m)$ is determined by equating his present utility with his expected utility if he pays the fee and plays the game; *i.e.,* $f(m)$ is the root of

$$\log m = \sum_{n=1}^{\infty} \frac{1}{2^n} \log(m - f + 2^n). \tag{13–4}$$

Computer evaluation gives $f(200) = 8.7204$; Laplace, without a computer, did his calculation very well. Likewise, $f(10^3) = 10.95$, $f(10^4) = 14.24$, $f(10^6) = 20.87$. Even a millionaire should not risk more than 21 francs on this dubious game.

It seems to us that this kind of numerical result is entirely reasonable. However the logarithmic assignment of utility is not to be taken literally either in the case of extremely small fortunes (as Laplace points out), or in the case of extremely large ones, as the following example of Savage (1954) shows:

Suppose your present fortune is \$1,000,000; then if your utility for money is proportional to the logarithm of the amount, you should be as willing as not to accept a wager in which, with probability one–half, you'll be left with only \$1,000, and with probability one–half, you will be left with \$1,000,000,000. Most of us would consider such a bet to be distinctly disadvantageous to a person with that initial fortune. This shows that our intuitive "utility" for money must increase even less rapidly than the logarithm for extremely large values. Chernoff and Moses (1959) claim that it is bounded; this appears to us plausible theoretically but not really demonstrated in the real world (why do billionaires continue to work harder than anybody else to earn still more money?).

The gist of Daniel Bernoulli's suggestion was therefore that, in the gambler's problem of decision making under uncertainty, one should act so as to maximize the expected value, not necessarily of the profit itself, but of some function of the profit which he called the "moral value". In more modern terminology the optimist will call this "maximizing expected utility," while the pessimist will speak instead of "minimizing expected loss", the loss function being taken as the negative of the utility function.

The Rationale of Insurance

Let us illustrate some of the above remarks briefly with the example of insurance, which is in some ways like the St. Petersburg game. The following scenario is oversimplified in obvious ways; nevertheless it makes some valid and important points. Insurance premiums are always set high enough to guarantee the insurance company a positive expectation of profit over all the contingencies covered in the contract, and every dollar the company makes is a dollar lost by a customer. Then why should anyone ever want to buy insurance?

The point is that the individual customer has a utility function for money that may be strongly curved over ranges of \$1000; but the insurance company is so much larger that its utility for money is accurately linear over ranges of millions of dollars. Thus, let P be the premium for some proposed insurance contract, let $i = 1, 2, \ldots n$ enumerate the contingencies covered, the i'th having probability w_i and cost to the insurance company, if it happens, of L_i. Let the prospective customer have Daniel Bernoulli's logarithmic utility for money and an initial amount M. Of course, by M we should understand his so–called "net worth", not merely the amount of cash he has on hand. Then the expected utility for the insurance company and for the customer, if he does or does not buy the insurance, will be:

	BUY	*DON'T BUY*
Company	$P - \sum w_i L_i$	0
Customer	$\log(M - P)$	$\sum w_i \log(M - L_i)$

So if $\langle L \rangle < P$, the company wants to sell the insurance, and if $\langle \log(M - L) \rangle < \log(M - P)$ the customer wants to buy it. If the premium is in the range

$$\langle L \rangle < P < [M - \exp\langle \log(M - L) \rangle] \tag{13–5}$$

it will be advantageous for both to do business.

We leave it as an exercise for the reader to show from (13–5) that a poor man should buy insurance, but a rich man should not unless his assessment of expected loss $\langle L \rangle$ is much greater than the insurance company's. Indeed, if your present fortune is much greater than any likely loss, then your utility for money is nearly as linear as the insurance company's, in the region where it matters; and you may as well be your own insurance company.

Further insight into the rich man's psychology is had by noting that if $M >> \langle L \rangle$ we may expand in powers of M^{-1};

$$M - \exp\langle \log(M - L) \rangle = \langle L \rangle + \frac{\mathrm{var}(L)}{2M} + \ldots \tag{13–6}$$

where $\mathrm{var}(L) = \langle L^2 \rangle - \langle L \rangle^2$. Thus a moderately rich man might be willing to buy insurance even if the premium is slightly larger than his expected loss, because this removes the uncertainty $\mathrm{var}(L)$ about the actual loss which he would otherwise have to live with; we have an aversion not only to risk, but to uncertainty about it.

Further insight into the poor man's psychology is had by writing the RHS of (13–5) as

$$M - \exp\langle \log(M - L) \rangle = M - \Pi_i \exp[w_i \log(M - L_i)] \tag{13–7}$$

Let the L_i be enumerated so that $L_1 \geq L_2 \geq L_3 \ldots$. Then this expression does not make sense unless $M > L_1$; but presumably it is not possible to have $M < L_1$, for one cannot lose more than he has. But if M approaches L_1, the last term becomes singular [$\exp(-\infty)$] and drops out.

Equation (13–5) then reduces to $\langle L \rangle < P < M$; it appears that this unfortunate person should always buy insurance if he can, even if this leaves him as poor as if the worst possible contingency had happened to him!

Of course, this only illustrates that the logarithmic utility assignment is unrealistic for very small amounts. In fact, the utility is clearly bounded in that region also; he who possesses only one penny does not consider it a calamity to lose it. We may correct this, as many have done, by replacing $\log M$ by $\log(M + b)$, where b is an amount so small that we consider it practically worthless. This modifies our conclusion from (13–7) in a way that we leave for the reader to work out, and which may suggest a good choice for b.

Entropy and Utility

The logarithmic assignment of utility is reasonable for many purposes, as long as it is not pushed to extremes. It is also, incidentally, closely connected with the notion of entropy, as shown by Bellman and Kalaba (1956). A gambler, who receives partially reliable advance tips on a game, acts (*i.e.,* decides on which side and how much to bet) so as to maximize the expected logarithm of his fortune. They show that (1) one can never go broke following this strategy, in contrast to the strategy of maximizing expected profit, where it is easily seen that with probability one this will happen eventually (the classical "gambler's ruin" situation), and (2) the amount one can reasonably expect to win on any one game is clearly proportional to the amount M_0 he has to begin with, so after n games, one could hope to have an amount $M = M_0 e^{\alpha n}$. Evidently, to use the logarithmic utility function means that one acts so as to maximize the expectation of α.

Exercise 13.1. Show that the maximum attainable $\langle \alpha \rangle$ is just $(H_0 - H)$, where H is the entropy which describes the gambler's uncertainty as to the truth of his tips, and H_0 is the maximum possible entropy, if the tips were completely uninformative.

A similar result is derived below. This suggests that, with more development of the theory, entropy might have an important place in guiding the strategy of a businessman or stock market investor.

There is a more subtle use of these considerations; the possibility not only of maximizing our own utility, but of manipulating the utility considerations of others so as to induce them to behave as we wish. Competent administrators know, instinctively but qualitatively, how to offer rewards and punishments so as to keep their organizations running smoothly and on course. A much oversimplified but quantitative example of this follows.

The Honest Weatherman

The weatherman's prior information and data yield a probability $p = P(Rain|Data, I)$ that it will rain tomorrow. Then what probability q will he announce publicly, in his evening TV forecast? This depends on his perceived utility function. We suspect that weather forecasters systematically overstate the probability of bad weather, *i.e.,* announce a value $q > p$, in the belief that they will incur more criticism from failing to predict a storm that arrives than from predicting one that fails to arrive.†

Nevertheless, we would prefer to be told the value p actually indicated by all the data at hand; indeed, if we were sure that we were being told this, we could not reasonably criticize the weatherman for his failures. Is it possible to give the weatherman a utility environment that will induce him always to tell the truth?

† Evidence for this is seen in the fact that, in St. Louis, we experience a predicted nonstorm almost every other week; but a nonpredicted storm is so rare that it is a major news item.

Suppose we write the weatherman's employment contract to stipulate that he will never be fired for making too many wrong predictions; but that each day, when he announces a probability q of rain, his pay for that day will be $B\log(2q)$ if it actually rains the next day, and $B\log 2(1-q)$ if it does not, where B is a base rate that does not matter for our present considerations, as long as it is high enough to make him want the job. Then the weatherman's expected pay for today, if he announces probability q, is

$$B[p\log 2q + (1-p)\log 2(1-q)] = B[\log 2 + p\log q + (1-p)\log(1-q)]\,. \tag{13–8}$$

Taking the first and second derivatives, we find that this is a maximum when $q = p$.

Now any continuous utility function appears linear if we examine only a small segment of it. Thus, if the weatherman considers a single days' pay small enough so that his utility for it is linear in the amount, it will always be to his advantage to tell the truth. There exist combinations of rewards and utility functions for which, quite literally, honesty is the best policy.

More generally, let there be n possible events $(A_1 \dots A_n)$ for which the available prior information and data indicate the probabilities $(p_1 \dots p_n)$. But a predictor chooses to announce instead the probabilities $(q_1 \dots q_n)$. Let him be paid $B\log(nq_i)$ if the event A_i subsequently occurs; he is rewarded for placing a high probability on the true event. Then his expectation of pay is

$$B[\log n - I(q;p)] \tag{13–9}$$

where $I(q;p) \equiv \sum p_i \log q_i$ is essentially (to within an additive constant) the relative entropy of the distributions [today commonly called the Kullback–Leibler Information, although its fundamental properties were proved and exploited already by Gibbs (1902, Chap. 11)]. Then it will be to his advantage to announce always $q_i = p_i$, and his maximum expectation of pay is

$$B[\log n - H(p_1 \dots p_n)] \tag{13–10}$$

where $H(p_i)$ is the entropy that measures his uncertainty about the A_i. It is not only to his advantage to tell the truth; it is to his advantage to acquire the maximum possible amount of information so as to decrease that entropy.

So, with an appropriate system of rewards, not only is honesty the best policy; industry is also encouraged. We see from this that a person who acts in his own self–interest is not necessarily acting counter to the interests of the rest of Society. Socio–economic activity is not a zero–sum game; it is possible, at least theoretically, to organize things so that the individual's self–interest runs parallel to that of Society as a whole (but we do not know how well this utopian situation is approximated by present Societies).

Reactions to Daniel Bernoulli and Laplace

The mathematically elementary – yet evidently important – nature of these results, might make one think that such things must have been not only perceived by many, but put to good use immediately, as soon as Daniel Bernoulli and Laplace had started this train of thought. Indeed, it seems in retrospect surprising that the notion of entropy was not discovered in this way, 100 years before Gibbs.

But the actual course of history has been very different; for most of the 20'th Century the "frequentist" school of thought either ignored the above line of reasoning or condemned it as metaphysical nonsense. In one of the best known books on probability theory (Feller, 1950; p. 199), Daniel Bernoulli's resolution of the St. Petersburg paradox is rejected without even being described, except to assure the reader that he "tried in vain to solve it by the concept of moral expectation." Warren M. Hirsch, in a review of the book, amplified this as follows:

"Various mystifying 'explanations' of this paradox had been offered in the past, involving, for example, the concept of moral expectation. These explanations are hardly understandable to the modern student of probability. Feller gives a straightforward mathematical argument which leads to the determination of finite entrance fee with which the St. Petersburg game has all the properties of a fair game."

We have just seen how 'vain' and 'hardly understandable' Daniel Bernoulli's efforts were. Reading Feller, one finds that he 'resolved' the paradox merely by defining and analyzing a different game. He undertakes to explain the rationale of insurance in the same way; but since he rejects Daniel Bernoulli's concept of a curved utility function, he concludes that insurance is always necessarily 'unfair' to the insured. These explanations are hardly understandable to the modern economist (or to us).

In the 1930's and 1940's a form of decision rules, as an adjunct to hypothesis testing, was expounded by J. Neyman and E. S. Pearson. It enjoyed a period of popularity with electrical engineers (Middleton, 1960) and economists (Simon, 1977) but it is now obsolete because it lacks two fundamental features now recognized as essential to the problem. In Chapter 14 we give a simple example of the Neyman–Pearson procedure, which shows how it is related to others.

Then in 1950 Abraham Wald gave a formulation that operates at a more fundamental level which makes it appear likely to have a permanent validity, as far as it goes, and gives a rather fundamental justification to Daniel Bernoulli's intuitive ideas. But these efforts were not appreciated in all quarters. Maurice Kendall (1963) wrote:

"There has been a strong movement in the U.S.A. to regard inference as a branch of decision theory. Fisher would have maintained (and in my opinion rightly) that inference in science is not a matter of decision, and that, in any case, criteria for choice in decision based on pay–offs of one kind or another are not available. This, broadly speaking, is the English as against the American point of view. $\cdots$ I propound the thesis that some such difference of attitude is inevitable between countries where what a man does is more important than what he thinks, and those where what he thinks is more important than what he does."

But we need not rely on second–hand sources for Fisher's attitude toward decision theory; as noted in Chapter 16, he was never at a loss to express himself on anything. In discussing significance tests, he writes (Fisher, 1956; p. 77):

"$\cdots$ recently $\cdots$ a considerable body of doctrine has attempted to explain, or rather to reinterpret, these tests on the basis of quite a different acceptance procedure. The differences between these two situations seem to the author many and wide, and I do not think it would have been possible to overlook them had the authors of this reinterpretation had any real familiarity with work in the natural sciences, or consciousness of those features of an observational record which permit of an improved scientific understanding."

Then he identifies Neyman and Wald as the objects of his criticism.

Apparently, Kendall, appealing to motives usually disavowed by scholars, regarded decision theory as a defect of the American, as opposed to the British character (although neither Neyman nor Wald was born or educated in America – they fled here from Europe). Fisher regarded it as an aberration of minds not versed in natural science (although the procedures were due originally to Daniel Bernoulli and Laplace, whose stature as natural scientists will easily bear comparison with Fisher's).

We agree with Kendall that the approach of Wald does indeed give the impression that inference is only a special case of decision; and we deplore this as much as he did. But we observe that in the original Bernoulli–Laplace formulation (and in ours), the clear distinction between these two functions is maintained, as it should be. But while we perceive this necessary distinction between inference and decision, we perceive also that inference not followed by decision is largely idle, and no natural scientist worthy of the name would undertake the labor of conducting inference unless it served some purpose.

These quotations give an idea of the obstacles which the perfectly natural, and immensely useful, ideas of Daniel Bernoulli and Laplace had to overcome; 200 years later, anyone who suggested such things was still coming under attack from the entrenched 'orthodox' statistical Establishment – and in a way that reflected no credit on the attackers. Let us now examine Wald's theory.

Wald's Decision Theory

Wald's formulation, in its initial stages, had no apparent connection with probability theory. We begin by imagining (*i.e.,* enumerating) a set of possible "states of Nature," $\{\theta_1, \theta_2, \ldots, \theta_N\}$ whose number is always, in practice, finite although it might be a useful limiting approximation to think of them as infinite or even as forming a continuum. In the quality–control example of Chapter 4, the "state of Nature" was the unknown number of defectives in the batch.

There are certain illusions that tend to grow and propagate here. Let us dispel one by noting that, in enumerating the different states of Nature, we are not describing any real (verifiable) property of Nature – for, one and only one of them is in fact true. The enumeration is only a means of describing a *state of knowledge* about the range of possibilities. Two persons, or robots, with different prior information may enumerate the θ_j differently without either being in error or inconsistent. One can only strive to do the best he can with the information he has, and we expect that the one with better information will naturally – and deservedly – make better decisions. This is not a paradox, but a platitude.

The next step in our theory is to make a similar enumeration of the decisions $\{D_1, D_2, \ldots, D_k\}$ that might be made. In the quality–control example, there were three possible decisions at each stage:

$D_1 \equiv$ "Accept the batch"
$D_2 \equiv$ "Reject the batch"
$D_3 \equiv$ "Make another test"

In the particle counter problem of Mr. B in Chapter 6, where we were to estimate the number n_1 of particles passing through the counter in the first second, there were an infinite number of possible decisions:

$D_i \equiv$ "n_1 is estimated as equal to 0, 1, 2, ..."

If we are to estimate the source strength, there are so many possible estimates that we thought of them as forming a continuum of possible decisions, even though in actual fact we can write down only a finite number of decimal digits.

This theory is clearly of no use unless by "making a decision" we mean, "deciding to act as if the decision were correct". It is idle for the robot to "decide" that $n_1 = 150$ is the best estimate unless we are then prepared to act on the assumption that $n_1 = 150$. Thus the enumeration of the D_i that we give the robot is a means of describing our knowledge as to what kinds of actions are *feasible*; it is idle and computationally wasteful to consider any decision which we know in advance corresponds to an impossible course of action.

There is another reason why a particular decision might be eliminated; even though D_1 is easy to carry out, we might know in advance that it would lead to intolerable consequences. An automobile driver can make a sharp turn at any time; but his common sense usually tells him not to. Here we see two more points: (1) there is a continuous gradation – the consequences of an action might be serious without being absolutely intolerable, and (2) the consequences of an action (= decision) will in general depend on what is the true state of Nature – a sudden sharp turn does not always lead to disaster, and it may actually avert disaster.

This suggests a third concept we need – the loss function $L(D_i, \theta_j)$, which is a set of numbers representing our judgment as to the "loss" incurred by making decision D_i if θ_j should turn out to be the true state of Nature. If the D_i and θ_j are both discrete, this is a loss matrix L_{ij}.

Quite a bit can be done with just the θ_j, D_i, L_{ij} and there is a rather extensive literature dealing with criteria for making decisions with no more than this. In the early days of this theory the results were summarized in a very readable and entertaining form by Luce and Raiffa (1957), and in the aforementioned elementary textbook of Chernoff and Moses (1959), which we recommend as still very much worth reading today. This culminated in the more advanced work of Raiffa & Schlaifer (1961), which is still a standard reference work because of its great amount of useful mathematical material (and, perhaps, its absence of rambling philosophy).

For a modern exposition with both the philosophy and the mathematics in more detail than we give here, see James Berger (1985). This is written from a Bayesian viewpoint almost identical to ours and it takes up many technical circumstances important for inference but which are not, in our view, really part of decision theory.

The minimax criterion is: for each D_i find the maximum possible loss $M_i = \max_j(L_{ij})$; then choose that D_i for which M_i is a minimum. This would be a reasonable strategy if we regard Nature as an intelligent adversary who foresees our decision and deliberately chooses the state of Nature so as to cause us the maximum frustration. In the theory of some games, this is not a completely unrealistic way of describing the situation, and consequently minimax strategies are of fundamental importance in game theory (von Neumann and Morgenstern, 1953).

But in the decision problems of the scientist, engineer, or economist we have no intelligent adversary, and the minimax criterion is that of the long–faced pessimist who concentrates all his attention on the worst possible thing that could happen, and thereby misses out on the favorable opportunities.

Equally unreasonable from our standpoint is the starry–eyed optimist who believes that Nature is deliberately trying to help him, and so uses this "minimin" criterion: for each D_i find the minimum possible loss $m_i = \min_j(L_{ij})$ and choose the D_i that makes m_i a minimum.

Evidently, a reasonable decision criterion for the scientist, engineer, or economist is in some sense intermediate between minimax and minimin, expressing our belief that Nature is neutral toward our goals. Many other criteria have been suggested, with such names as maximin utility (Wald), α–optimism–pessimism (Hurwicz), minimax regret (Savage), *etc.* The usual procedure, as described in detail by Luce and Raiffa, has been to analyze any proposed criterion to see whether it satisfies about a dozen qualitative common–sense conditions such as

(1) *Transitivity:* If D_1 is preferred to D_2, and D_2 preferred to D_3, then D_1 should be preferred to D_3

(2) *Strong Domination:* If for all states of Nature θ_j we have $L_{ij} < L_{kj}$, then D_i should always be preferred to D_k.

This kind of analysis, although straightforward, can become tedious. We do not follow it any further, because the final result is that there is only one class of decision criteria which passes all the tests, and this class is obtained more easily by a different line of reasoning.

A full decision theory, of course, cannot concern itself merely with the θ_j, D_i, L_{ij}. We also, in typical problems, have additional evidence E, which we recognize as relevant to the decision problem, and we have to learn how to incorporate E into the theory. In the quality–control example of Chapter 4, E consisted of the results of the previous tests.

At this point the decision theory of Wald takes a long, difficult, and as we now realize, unnecessary mathematical detour. One defines a "strategy" S, which is a set of rules of the form, "If I receive new evidence E_i, then I will make decision D_k." In principle one first enumerates all conceivable strategies (whose number is, however, astronomical even in quite simple problems), and then eliminates the ones considered undesirable by the following criterion. Denote by

$$p(D_k|\theta_j S) = \sum_i p(D_k|E_i\theta_j S)\, p(E_i|\theta_j) \tag{13–11}$$

the sampling probability that, if θ_j is the true state of Nature, strategy S would lead us to make decision D_k, and define the *risk* presented by θ_j with strategy S as the expected loss over this distribution:

$$R_j(S) = \langle L\rangle_j = \sum_k p(D_k|\theta_j S) L_{kj}. \tag{13–12}$$

Then a strategy S is called *admissible* if no other S' exists for which

$$R_j(S') \le R_j(S), \qquad \textit{all } j. \tag{13–13}$$

If an S' exists for which the strict inequality holds for at least one θ_j, then S is termed *inadmissible.* The notions of risk and admissibility are evidently sampling theory criteria, not Bayesian, since they invoke only the sampling distribution. Wald, thinking in sampling theory terms, considered it obvious that the optimal strategy should be sought only within the class of admissible ones.

A principal object of Wald's theory is then to characterize the class of admissible strategies in mathematical terms, so that any such strategy can be found by carrying out a definite procedure. The fundamental theorem bearing on this is Wald's Complete Class Theorem which establishes a result shocking to sampling theorists (including Wald himself). Berger (1985, Chap. 8) discusses this in Wald's terminology. The term "complete class" is defined in a rather awkward way (Berger, *loc. cit.*, pp. 521–522). What Wald really wanted was just the set of all admissible rules, which Berger calls a "minimal complete class". From Wald's viewpoint it is a highly nontrivial mathematical problem to prove that such a class exists, and to find an algorithm by which any rule in the class can be constructed.

However, from our viewpoint these are unnecessary complications, signifying only an inappropriate definition of the term "admissible". We shall return to this issue in Chapter 17 and come to a different conclusion; an 'inadmissible' decision may be overwhelmingly preferable to an 'admissible' one, because the criterion of admissibility ignores prior information – even information so cogent that, for example, in major medical, public health, or airline safety decisions, to ignore it would put lives in jeopardy and support a charge of criminal negligence.

This illustrates the folly of inventing noble–sounding names like 'admissible' and 'unbiased' for principles that are far from noble; and not even fully rational. In the future we should profit from this lesson and take care that we describe technical conditions by names that are ethically and morally neutral, and so do not have false connotations which could mislead others for decades, as these have.

Since in real applications we do not want to – and could not – restrict ourselves to admissible rules anyway, we shall not follow this quite involved argument. We give a different line of reasoning which leads to the rules which are appropriate in the real world, while giving us a better understanding of the reason for them.

What makes a decision process difficult? Well, if we knew which state of Nature was the correct one, there would be no problem at all; if θ_3 is the true state of Nature, then the best decision D_i is the one which renders L_{i3} a minimum. In other words, once the loss function has been specified, our uncertainty as to the best decision arises solely from our uncertainty as to the state of Nature. Whether the decision minimizing L_{i3} is or is not best depends on this: How strongly do we believe that θ_3 is the true state of Nature? How plausible is θ_3?

To our robot it seems a trivial step – really only a rephrasing of the question – to ask next, "Conditional on all the available evidence, what is the *probability* P_3 that θ_3 is the true state of Nature?" Not so to the sampling theorist, who regards the word "probability" as synonomous with "long–run relative frequency in some random experiment". On this definition it is meaningless to speak of the probability of θ_3, because the state of nature is not a "random variable". Thus, if we adhere consistently to the sampling theory view of probability, we shall conclude that probability theory cannot be applied to the decision problem, at least not in this direct way.

It was just this kind of reasoning which led statisticians, in the early part of this century, to relegate problems of parameter estimation and hypothesis testing to a new field, Statistical Inference, which was regarded as distinct from probability theory, and based on entirely different principles. But let us examine a typical problem of this type from the sampling theory viewpoint, and see how introducing the notion of a loss function changes this conclusion.

Parameter Estimation for Minimum Loss

There is some unknown parameter α, and we make n repeated observations of a quantity, obtaining an observed "sample" $x \equiv \{x_1 \ldots x_n\}$. We interpret the symbol x, without subscripts, as standing for a vector in an n–dimensional "sample space" and suppose that the possible results x_i of individual observations are real numbers which we think of as continuously variable in some domain $(a \le x_i \le b)$. From observation of the sample x, what can we say about the unknown parameter α? We have already studied such problems from the Bayesian "Probability Theory as Logic" viewpoint; now we consider them from the sampling theory viewpoint.

To state the problem more drastically, suppose that we are compelled to choose one specific numerical value as our "best" estimate of α, on the basis of the observed sample x, and any other prior information we might have, and then to act as if this estimate were true. This is the decision situation which we all face daily, both in our professional capacity and in everyday life. The driver approaching a blind intersection cannot know with certainty whether he will have enough time to cross it safely; but still he is compelled to make a decision based on what he can see, and act on it.

Now it is clear that in estimating α, the observed sample x is of no use to us unless we can see some kind of logical (not necessarily causal) connection between α and x. In other words, if we knew α, but not x, then the probabilities which we would assign to various observable samples must depend in some way on the value of α. If we consider the different observations as independent, as was almost always done in the sampling theory of parameter estimation, then the sampling density function factors:

$$f(x|\alpha) = f(x_1|\alpha) \ldots f(x_n|\alpha). \tag{13–14}$$

However, this very restrictive assumption is not necessary (and in fact does not lead to any formal simplification) in discussing the general principles of parameter estimation from the decision theory standpoint.

Let $\beta = \beta(x_1 \ldots x_n)$ be an "estimator", *i.e.*, any function of the data values, proposed as an estimate of α. Also, let $L(\alpha, \beta)$ be the "loss" incurred by guessing the value β when α is in fact the true value. Then for any given estimator the risk is the "pre–data" expected loss; *i.e.* the loss *for a person who already knows the true value of* α *but does not know what data will be observed:*

$$R_\alpha \equiv \int L(\alpha, \beta)\, f(x|\alpha)\, dx. \tag{13–15}$$

By $\int(\ )dx$ we mean the n–fold integration

$$\int \cdots \int (\ )\, dx_1 \ldots dx_n. \tag{13–16}$$

We may interpret this notation as including both the continuous and discrete cases; in the latter $f(x|\alpha)$ is a sum of delta–functions.

On the view of one who uses the frequency definition of probability, the above phase, "for a person who already knows the true value of α" is misleading and unwanted. The notion of the probability of sample x *for a person with a certain state of knowledge* is entirely foreign to him; he regards $f(x|\alpha)$ not as a description of a mere state of knowledge about the sample, but as an objective statement of fact, giving the relative frequencies with which different samples would be observed "in the long run".

Unfortunately, to maintain this view strictly and consistently would reduce the legitimate applications of probability theory almost to zero; for one can (and most of us do) work in this field for a lifetime without ever encountering a real problem in which one actually has knowledge of the "true" limiting frequencies for an infinite number of trials; how could one ever acquire such knowledge? Indeed, quite apart from probability theory, no scientist ever has sure knowledge of what is "really true"; the only thing we can ever know with certainty is: *what is our state of knowledge*? To describe this is all that any science could ever have claimed to do.

Then how could one ever assign a probability which he knew was equal to a limiting frequency in the real world? It seems to us that the belief that probabilities are realities existing in Nature is pure Mind Projection Fallacy; and true "Scientific Objectivity" demands that we escape from this delusion and recognize that in conducting inference our equations are not describing reality, only our information about reality.

In any event, the "frequentist" believes that R_α is not merely the "expectation of loss" in the present situation, but is also, with probability 1, the limit of the average of *actual* losses which would be incurred by using the estimator β an indefinitely large number of times; *i.e.*, by drawing a sample of n observations repeatedly with a fixed value of α. Furthermore, the idea of finding the estimator which is "best for the present specific sample" is quite foreign to his outlook; because he regards the notion of probability as referring to a collection of cases rather than a single case, he is forced to speak instead of finding that estimator "which will prove best, on the average, in the long run".

On the frequentist view, therefore, it would appear that the best estimator will be the one that minimizes R_α. Is this a variational problem? A small change $\delta\beta(x)$ in the estimator changes the risk by

$$\delta R_\alpha = \int \frac{\partial L(\alpha,\beta)}{\partial\beta}\, f(x|\alpha)\,\delta\beta(x)\,dx\,. \qquad (13\text{–}17)$$

If we were to require this to vanish for all $\delta\beta(x)$, this would imply

$$\frac{\partial L}{\partial\beta} = 0, \qquad \text{all possible } \beta\,. \qquad (13\text{–}18)$$

Thus the problem as stated has no truly stationary solution except in the trivial – and useless – case where the loss function is independent of the estimated value β; if there is any "best" estimator by the criterion of minimum risk, it cannot be found by variational methods.

Nevertheless, we can get some understanding of what is happening by considering (13–12) for some specific choices of loss function. Suppose we take the quadratic loss function $L(\alpha,\beta) = (\alpha-\beta)^2$. Then (13–12) reduces to

$$R_\alpha = \int (\alpha^2 - 2\alpha\beta + \beta^2)\, f(x|\alpha)\,dx \qquad (13\text{–}19)$$

or,

$$R_\alpha = (\alpha - \langle\beta\rangle)^2 + \mathrm{var}(\beta) \qquad (13\text{–}20)$$

where $\mathrm{var}(\beta) = \langle \beta^2 \rangle - \langle \beta \rangle^2$ is the variance of the sampling *pdf* for β, and

$$\langle \beta^n \rangle \equiv \int [\beta(x)]^n \, f(x|\alpha)\, dx \tag{13–21}$$

is the n'th moment of that *pdf*. The risk (13–20) is the sum of two positive terms, and a good estimator by the criterion of minimum risk has two properties:

(1) $\langle \beta \rangle = \alpha$

(2) $\mathrm{var}(\beta)$ is a minimum

These are just the two conditions which sampling theory has considered most important. An estimator with property (1) is called *unbiased* [more generally, the function $b(\alpha) = \langle \beta \rangle - \alpha$ is called the *bias* of the estimator $\beta(x)$], and one with both properties (1) and (2) was called *efficient* by R. A. Fisher. Nowadays, it is often called an *unbiased minimum variance* (UMV) estimator.

In Chapter 17 we shall examine the relative importance of removing bias and minimizing variance, and derive the Cramér–Rao inequality which places a lower limit on the possible value of $\mathrm{var}(\beta)$. For the present, our concern is only with the failure of (13–17) to provide any optimal estimator for a given loss function. This weakness of the sampling theory approach to parameter estimation, that it does not tell us how to find the best estimator, but only how to compare different guesses, can be overcome as follows: we give a simple substitute for Wald's Complete Class Theorem.

Reformulation of the Problem

It is easy to see why the criterion of minimum risk is bound to get us into trouble and is unable to furnish any general rule for constructing an estimator. The mathematical problem was: for given $L(\alpha, \beta)$ and $f(x|\alpha)$, what function $\beta(x_1 \ldots x_n)$ will minimize R_α?

Although this is not a variational problem, it might have a unique solution; but the more fundamental difficulty is that the solution will still, in general depend on α. Of course, there may be special cases where the α–dependence drops out; but in general the criterion of minimum risk leads to an impossible situation – even if we could solve the mathematical minimization problem and had before us the best estimator $\beta_\alpha(x_1 \ldots x_n)$ for each value of α, we could use that result only if α were already known, in which case we would have no need to estimate. We were looking at the problem backwards!

This makes it clear how to correct the trouble. It is of no use to ask what estimator is 'best' for some particular value of α; the answer to that question is always, obviously, $\beta(x) = \alpha$. But the only reason for using an estimator is that α is unknown. The estimator must therefore be some compromise that allows for all possibilities within some prescribed range of α; within this range it must do the best job of protecting against loss no matter what the true value of α turns out to be.

Thus it is some weighted average of R_α;

$$\langle R \rangle = \int g(\alpha)\, R_\alpha \, d\alpha \tag{13–22}$$

that we should really minimize, where the function $g(\alpha) \geq 0$, measures in some way the relative importance of minimizing R_α for the various possible values that α might turn out to have.

But the mathematical character of the problem is completely changed by adopting (13–22) as our criterion; we now have a solvable variational problem with a unique, well–behaved, and useful solution. The first variation in $\langle R \rangle$ due to an arbitrary variation $\delta\beta(x_1 \ldots x_n)$ in the estimator is

$$\delta \langle R \rangle = \int \cdots \int dx_1 \cdots dx_n \left\{ \int d\alpha \, g(\alpha) \, \frac{\partial L(\alpha, \beta)}{\partial \beta} \, f(x_1 \ldots x_n | \alpha) \right\} \delta\beta(x_1 \ldots x_n) \tag{13–23}$$

which vanishes independently of $\delta\beta$ if

$$\int d\alpha\, g(\alpha)\, \frac{\partial L(\alpha,\beta)}{\partial\beta}\, f(x_1 \ldots x_n|\alpha) = 0 \tag{13–24}$$

for all possible samples $\{x_1 \ldots x_n\}$.

Equation (13–24) is the fundamental integral equation which determines the 'best' estimator by our new criterion. Taking the second variation, we find the condition that (13–24) shall yield a true minimum is

$$\int d\alpha\, g(\alpha)\, \frac{\partial^2 L}{\partial\beta^2}\, f(x_1 \ldots x_n|\alpha) > 0. \tag{13–25}$$

Thus a sufficient condition for a minimum is simply $\partial^2 L/\partial\beta^2 \geq 0$, but this is stronger than necessary.

If we take the quadratic loss function $L(\alpha,\beta) = K(\alpha-\beta)^2$, equation (13–24) reduces to

$$\int d\alpha\, g(\alpha)\, (\alpha-\beta)\, f(x_1 \ldots x_n|\alpha) = 0$$

or, the optimal estimator for quadratic loss is

$$\beta(x_1 \ldots x_n) = \frac{\int d\alpha\, g(\alpha)\, \alpha\, f(x_1 \ldots x_n|\alpha)}{\int d\alpha\, g(\alpha)\, f(x_1 \ldots x_n|\alpha)}. \tag{13–26}$$

But this is just the mean value over the posterior *pdf* for α:

$$f(\alpha|x_1 \ldots x_n I) = \frac{g(\alpha)\, f(x_1 \ldots x_n|\alpha)}{\int d\alpha\, g(\alpha)\, f(x_1 \ldots x_n|\alpha)} \tag{13–27}$$

given by Bayes' theorem, if we interpret $g(\alpha)$ as a prior probability density! This argument shows, perhaps more clearly than any other we have given, why the *mathematical form* of Bayes' theorem intrudes itself inevitably into parameter estimation.

If we take as a loss function the absolute error, $L(\alpha,\beta) = |\alpha-\beta|$, then the integral equation (13–24) becomes

$$\int_{-\infty}^{\beta} d\alpha\, g(\alpha)\, f(x_1 \ldots x_n|\alpha) = \int_{\beta}^{\infty} d\alpha\, g(\alpha)\, f(x_1 \ldots x_n|\alpha) \tag{13–28}$$

which states that $\beta(x_1 \ldots x_n)$ is to be taken as the *median* over the posterior *pdf* for α:

$$\int_{-\infty}^{\beta} d\alpha\, f(\alpha|x_1 \ldots x_n I) = \int_{\beta}^{\infty} d\alpha\, f(\alpha|x_1 \ldots x_n I) = \frac{1}{2} \tag{13–29}$$

Likewise, if we take a loss function $L(\alpha,\beta) = (\alpha-\beta)^4$, equation (13–24) leads to an estimator $\beta(x_1 \ldots x_n)$ which is the real root of

$$f(\beta) = \beta^3 - 3\overline{\alpha}\beta^2 + 3\overline{\alpha^2}\beta - \overline{\alpha^3} = 0 \tag{13–30}$$

where

$$\overline{\alpha^n} = \int d\alpha \, \alpha^n \, f(\alpha|x_1 \ldots x_n I) \tag{13–31}$$

is the n'th moment of the posterior *pdf* for α. [That (13–30) has only one real root is seen on forming the discriminant; the condition $f'(\beta) \geq 0$ for all real β is just $(\overline{\alpha^2} - \overline{\alpha}^2) \geq 0$.]

If we take $L(\alpha, \beta) = |\alpha - \beta|^k$, and pass to the limit $k \to 0$, or if we just take

$$L(\alpha, \beta) = \left\{ \begin{array}{ll} 0\,, & \alpha = \beta \\ 1\,, & \text{otherwise} \end{array} \right\} \tag{13–32}$$

Eq. (13–24) tells us that we should choose $\beta(x_1 \ldots x_n)$ as the "most probable value", or *mode* of the posterior *pdf* $f(\alpha|x_1 \ldots x_n I)$. If $g(\alpha) =$ const in the high–likelihood region, this is just the maximum likelihood estimate advocated by Fisher.

In this result we see finally just what maximum likelihood accomplishes, and under what circumstances it is the appropriate method to use. The maximum likelihood criterion is the one in which we care only about the chance of being exactly right; and if we are wrong, we don't care how wrong we are. This is just the situation we have in shooting at a small target, where "a miss is as good as a mile". But it is clear that there are few other situations where this would be a rational way to behave; almost always, the amount of error is of some concern to us, and so maximum likelihood is not the best estimation criterion.

Note that in all these cases it was the posterior *pdf*, $f(\alpha|x_1 \ldots x_n I)$ that was involved. That this will always be the case is easily seen by noting that our "fundamental integral equation" (13–24) is not so profound after all. It can be written equally well as

$$\frac{\partial}{\partial \beta} \int d\alpha \, g(\alpha) L(\alpha, \beta) f(x_1 \ldots x_n|\alpha) = 0. \tag{13–33}$$

But if we interpret $g(\alpha)$ as a prior probability density, this is just the statement that we are indeed to minimize the expectation of $L(\alpha, \beta)$; but it is not the expectation over the sampling *pdf* for β; it is always the expectation over the Bayesian posterior *pdf* for α!

We have here an interesting case of "chickens coming home to roost". If a sampling theorist will think his estimation problems through to the end, he will find himself obliged to use the Bayesian mathematical algorithm even if his ideology still leads him to reject the Bayesian rationale for it. This mathematical form is required for simple logical reasons, independently of all philosophical hangups over "which quantities are random?" or the "true meaning of probability."

Wald's complete class theorem led him to essentially the same conclusion: if the θ_j are discrete and we agree not to include in our enumeration of states of Nature any θ_j that is known to be impossible, then the class of admissible strategies is just the class of Bayes strategies [*i.e.*, those that minimize expected loss over a posterior *pdf*]. If the possible θ_j form a continuum, the admissible rules are the proper Bayesian ones; *i.e.*, Bayes rules from proper (normalizable) prior probabilities. But few people have ever tried to follow his proof of this; Berger (1985) gives instead a number of isolated special results.

There is a great deal of mathematical nit–picking, also noted by Berger, over the exact situation when one tries to jump into an improper prior in infinite parameter spaces without considering any limit from a proper prior. But for us such questions are of no interest, because the concept of admissibility is itself flawed when stretched to such extreme cases. Because of its refusal to consider any prior information whatsoever, it must consider all points of an infinite domain equivalent; the resulting singular mathematics is only an artifact that corresponds to no singularity in the real problem, where prior information always excludes the region at infinity.

For a given sampling distribution and loss function, we are content to say simply that the defensible decision rules are the Bayes rules characterized by the different proper priors, and their well–behaved limits. This is the conclusion that was shocking to sampling theorists, and was psychologically perhaps the main spark that touched off our present 'Bayesian Revolution' in statistics. But in arriving at these inevitable results, the Bayesian rationale has the advantages that (1) it leads us to this conclusion immediately; (2) it makes it obvious that its range of validity and usefulness is far greater than supposed by the sampling theorist.

However, not all real decision problems are covered by the above theory. For example, the state of Nature tomorrow might be influenced by our decision today (as when one decides to get an education). Recognizing this is a step in the direction of game theory, or dynamic programming. But to treat such problems does not require any departure from the principles of probability theory as logic; only a generalization of what we did above.

Effect of Varying Loss Functions

Since the new feature of the theory being expounded here lies only in the introduction of the loss function, it is important to understand how the final results depend on the loss functions by some numerical examples. Suppose that the prior information I and data D lead to the following posterior *pdf* for a parameter α:

$$f(\alpha|DI) = ke^{-k\alpha}\,, \qquad 0 \le \alpha < \infty. \tag{13–34}$$

The n'th moment of this *pdf* is

$$\langle \alpha^n \rangle = \int_0^\infty \alpha^n\, f(\alpha|DI)\, d\alpha = n!\, k^{-n}\,. \tag{13–35}$$

With loss function $(\alpha-\beta)^2$, the best estimator is the mean value

$$\beta = \langle \alpha \rangle = k^{-1}. \tag{13–36}$$

With the loss function $|\alpha-\beta|$, the best estimator is the median, determined by

$$\frac{1}{2} = \int_0^\alpha f(\alpha|DI)\, d\alpha = 1 - e^{-k\beta} \tag{13–37}$$

or

$$\beta = k^{-1}\log_e(2) = 0.693\langle\alpha\rangle\,. \tag{13–38}$$

To minimize $\langle(\alpha-\beta)^4\rangle$, we should choose β to satisfy (13–30), which becomes $y^3 - 3y^2 + 6y - 6 = 0$ with $y = k\beta$. The real root of this is at $y = 1.59$, so the optimal estimator is

$$\beta = 1.59\,\langle\alpha\rangle. \tag{13–39}$$

For the loss function $(\alpha-\beta)^{s+1}$ with s an odd integer, the fundamental equation (13–33) is

$$\int_0^\infty (\alpha-\beta)^s\, e^{-k\alpha}\, d\alpha = 0 \tag{13–40}$$

which reduces to

$$\sum_{m=0}^{s} \frac{(-k\beta)^m}{m!} = 0 \tag{13–41}$$

The case $s = 3$ leads to (13–39),while in the case $s = 5$, loss function $(\alpha - \beta)^6$, we find

$$\beta = 2.025 \, \langle \alpha \rangle . \tag{13–42}$$

As $s \to \infty$, β also increases without limit. But the maximum-likelihood estimate, which corresponds to the loss function $L(\alpha, \beta) = -\delta(\alpha - \beta)$, or equally well to

$$\lim_{k \to 0} |\alpha - \beta|^k \tag{13–43}$$

is $\beta = 0$. These numerical examples merely illustrate what was already clear intuitively; when the posterior *pdf* is not sharply peaked, the best estimate of α depends very much on which particular loss function we use.

One might suppose that a loss function must always be a monotonically increasing function of the error $|\alpha - \beta|$. In general, of course, this will be the case, but there is nothing in this theory which restricts us to such functions. You can think of some rather frustrating situations in which, if you are going to make an error, you would rather make a large one than a small one. William Tell was in just that fix. If you study our equations for this case, you will see that there is really no very satisfactory decision at all (*i.e.*, no decision has small expected loss); and nothing can be done about it.

Note that the decision rule is invariant under any proper linear transformation of the loss function; *i.e.*, if $L(D_i, \theta_j)$ is one loss function, then the new one

$$L'(D_i, \theta_j) \equiv a + bL(D_i, \theta_j) \qquad \begin{array}{r} -\infty < a < \infty \\ 0 < b < \infty \end{array} \tag{13–44}$$

will lead to the same decision, whatever the prior probabilities and data. Thus, in a binary decision problem, given the loss matrix

$$L_{ij} = \begin{pmatrix} 10 & 19 \\ 100 & 10 \end{pmatrix}$$

we can equally well use

$$L'_{ij} = \begin{pmatrix} 0 & 1 \\ 10 & 0 \end{pmatrix}$$

corresponding to $a = -10/9$, $b = 1/9$. This may simplify the calculation of expected loss quite a bit.

General Decision Theory

In the foregoing, we examined decision theory only in terms of one particular application; parameter estimation. But we really have the whole story already; the criterion (13–33) for constructing the optimal estimator generalizes immediately to the criterion for finding the optimal decision of any kind. The final rules are simple; to solve the problem of inference, there are four steps:

(1) Enumerate the possible states of nature θ_j, discrete or continuous, as the case may be.

(2) Assign prior probabilities $p(\theta_j|I)$ which represent whatever prior information I you have about them.

(3) Assign sampling probabilities $p(E_i|\theta_j)$ which represent your prior knowledge about the mechanism of the measurement process yielding the possible data sets E_i.

(4) Digest any additional evidence $E = E_1 E_2 \cdots$ by application of Bayes' theorem, thus obtaining the posterior probabilities $p(\theta_j|EI)$.

That is the end of the inference problem, and $p(\theta_j|EI)$ expresses all the information about the θ_j that is contained in the prior information and data. To solve the problem of decision there are three more:

(5) Enumerate the possible decisions D_i.

(6) Assign the loss function $L(D_i, \theta_j)$ that tells what you want to accomplish.

(7) Make that decision D_i which minimizes the expected loss over the posterior probabilities for θ_j

After all is said and done, the final rules of calculation to which the theorems of Cox, Wald, and Shannon lead us are just the ones which had been given already by Laplace and Daniel Bernoulli in the 18'th century on intuitive grounds, except that the entropy principle generalizes the principle of indifference in step (2).

Theoretically, these rules are now determined uniquely by elementary qualitative desiderata of rationality and consistency. Some protest that they do not have any prior probability or loss function. The theorem is that rationality and consistency require you to behave *as if* you had them; for every strategy that obeys the desiderata, there is a prior probability and loss function which would have led to that strategy; conversely, if a strategy is derived from a prior probability and loss function, it is guaranteed to obey the desiderata.

Pragmatically, these rules either include, or improve upon, practically all known statistical methods for hypothesis testing and point estimation of parameters. If you have mastered them, then you have just about the entire field at your fingertips. The outstanding thing about them is their intuitive appeal and simplicity – if we sweep aside all the polemics and false starts that have cluttered up this field in the past and consider only the constructive arguments that lead directly to these rules, it is clear that the underlying rationale could be developed fully in a one–semester undergraduate course.

However, in spite of the formal simplicity of the rules themselves, really facile application of them in nontrivial problems involves intricate mathematics, and fine subtleties of concept; so much so that several generations of workers in this field misapplied them and concluded that the rules were all wrong. So, we still need a good deal of leading by the hand in order to develop facility in using this theory. It is like learning how to play a musical instrument – anybody can make noise with it, but to play this instrument well requires years of practice.

COMMENTS

"Objectivity" of Decision Theory

Decision Theory occupies a unique position in discussion of the logical foundations of statistics, because, as we have seen in (13–24) and (13–33), its procedures can be derived from either of two diametrically opposed viewpoints about the nature of probability theory. While there appears to be universal agreement as to the actual procedures that should be followed, there remains a fundamental disagreement as to the underlying reason for them, having its origin in the old issue of frequency vs. non–frequency definitions of probability.

From a pragmatic standpoint, such considerations may seem at first to be unimportant. However, in the attempt to apply decision–theory methods in real problems one learns very quickly that these questions intrude in the initial stage of setting up the problem in mathematical terms. In particular, our judgment as to the generality and range of validity of decision–theory methods depends on how these conceptual problems are resolved. Our aim is to expound the viewpoint according to which these methods have the greatest possible range of application.

Now we find that the main source of controversy here is on the issue of prior probabilities; on the sampling theory viewpoint, if the problem involves use of Bayes' theorem then these methods are just not applicable unless the prior probabilities are known frequencies. But to maintain this position consistently would imply an enormous restriction on the range of legitimate applications; indeed, we doubt whether there has ever been a real problem in which the prior probabilities were, in fact, known frequencies. But can the mathematical form of our final equations shed any light on this issue?

Notice first that only the product $g(\alpha)L(\alpha,\beta)$ is involved in (13–24) or (13–33); thus we could interpret the problem in three different ways:

(1) Prior Probability $g(\alpha)$, loss function $L(\alpha,\beta) = (\alpha-\beta)^2$
(2) Uniform prior probability, loss function $L(\alpha,\beta) = g(\alpha)\,(\alpha-\beta)^2$
(3) Prior probability $h(\alpha)$, loss function $g(\alpha)(\alpha-\beta)^2/h(\alpha)$,

but the optimal decision is just the same. This is equally true for any loss function.

We emphasize this rather trivial mathematical fact because of a curious psychological phenomenon. In expositions of decision theory written from the sampling theory viewpoint [for example, Chernoff and Moses (1959)], the writers are reluctant to introduce the notion of prior probability. They postpone it as long as possible, and finally give in only when the mathematics forces them to recognize that prior probabilities are the only basis for choice among the different admissible decision rules. Even then, they are so unhappy about the use of prior probabilities that they feel it necessary always to invent a situation – often highly artificial – which makes the prior probabilities appear to be frequencies; and they will not use this theory for any problem where they do not see how to do this.

But these same writers do not hesitate to pull a completely arbitrary loss function out of thin air without any basis at all, and proceed with the calculation! The equations show that if the final decision depends strongly on which particular prior probability assignment we use, it is going to depend just as strongly on which particular loss function we use. If one worries about arbitrariness in the prior probabilities, then in order to be consistent, he ought to worry just as much about arbitrariness in the loss functions. If one claims (as sampling theorists did for decades and as some still do) that uncertainty as to the proper choice of prior probabilities invalidates the Laplace–Bayes theory, then in order to be consistent, one must claim also that uncertainty as to the proper choice of loss functions invalidates Wald's decision theory.

The reason for this strange lopsided attitude is closely connected with a certain philosophy variously called behavioristic, or positivistic, which wants us to restrict our statements and concepts to objectively verifiable things. Therefore the observable *decision* is the thing to emphasize, while the process of plausible reasoning and the judgment described by a prior probability must be deprecated and swept under the rug. But we see no need to do this, because it seems to us obvious that rational action can come only as the result of rational thought.

If we refuse to consider the problem of rational thought merely on the grounds that it is not "objective", the result will not be that we obtain a more "objective" theory of inference or decision. The result will be that we have lost the possibility of getting any satisfactory theory at all, because we have denied ourselves any way of describing what is actually going on in the decision process. And, of course, the loss function is just the expression of a purely subjective value judgment, which can in no way be considered any more "objective" than the prior probabilities.

In fact, prior probabilities are usually far more "objective" than loss functions, both in the mathematical theory and in the everyday decision problems of "real life". In the mathematical theory we have general formal principles – maximum entropy, transformation groups, marginalization – that remove the arbitrariness of prior probabilities for a large class of important problems, which includes most of those discussed in textbooks. But we have no such principles for determining loss/utility functions.

This is not to say that the problem has not been discussed. de Groot (1970) notes the very weak abstract conditions (transitivity of preferences, *etc.*) sufficient to guarantee existence of a utility function. Long ago, L. J. Savage considered construction of utility functions by introspection. This is described by Chernoff & Moses (1959); suppose there are two possible rewards r_1 and r_2; then for what reward r_3 would you be indifferent between (r_3 for sure), or (either r_1 or r_2 as decided by the flip of a coin)? Presumably, r_3 is somewhere between r_1 and r_2. If one makes enough such intuitive judgments and manages to correct all intransitivities, a crude utility function emerges. Berger (1985, Chapter 2) gives a scenario in which this happens.

However, this is hardly a practical procedure, much less a formal principle; and the result is just as arbitrary as if one simply drew a curve freehand. Indeed, the latter is much easier and cannot get one into intransitivity difficulties. One can, of course, invent a crude prior in the same way, as L. J. Savage often demonstrated. Such constructions, if one can transfer them into a computer, will be better than nothing; but they are clearly desperation moves in lieu of a really satisfactory formal theory such as we have in the principles of maximum entropy and transformation groups for priors.

Noting that the decision depends only on the product of loss function and prior suggests what seems at first an attractive possibility; could we simplify the foundations of this theory so as to make it obvious that we need only a single function, not two? The writer pondered this for some time, but decided finally that this is not the right direction for future development, because (1) priors and loss functions have very different – almost opposite – roles to play, both in the mathematical theory and in "real life", and (2) the theory of inference involving priors is more fundamental than that of loss functions; the latter would need to be developed much further before it would be fit to join with priors into a single mathematical quantity.

Loss Functions in Human Society

For example, let us note the sharp contrast between the roles of prior probabilities and loss functions in human relations. People with similar prior probabilities get along well together, because they have about the same general view of the world and philosophy of life. People with radically different prior probabilities cannot get along – this has been the root cause of all the religious wars and most of the political repressions throughout history.

Loss functions operate in just the opposite way. People with similar loss functions are after the same thing, and are in contention with each other. People with different loss functions get along well because each is willing to give something that the other wants. Amicable trade or business transactions, advantageous to all, is possible only between parties with very different loss functions. We illustrated this by the example of insurance above.

In "real life" decision problems, each man knows, pretty well, what his prior probabilities are; and because his beliefs are based on all his past experience, they are not easily changed by one more experience, so they are fairly stable. But in the heat of argument he may lose sight of his loss function.

Thus the labor mediator must deal with parties with sharply opposing ideologies; policies considered good by one are considered evil by the other. The successful mediator realizes that mere talk will not alter prior beliefs; and so his role must be to turn the attention of both parties away from this area, and explain clearly to each what his loss function is. In this sense, we can claim that in real life decision problems, the loss function is often far more "subjective" (in the sense of being less well fixed in our minds) than the prior probabilities.

Indeed, failure to judge one's own loss function correctly is one of the major dangers that humans face. Having a little intelligence, one can invent myths out of his own imagination, and come to believe them. Worse, one person may persuade thousands of others to believe his private myths, as the sordid history of religious, political, and military disasters shows.

We think that these considerations have a bearing on other social problems. For example, some psychologists never tire of trying to explain criminal behavior in terms of early childhood experiences. It is conceivable that these may generate a certain general 'propensity' to crime; but the fact that the vast majority of people with the same experiences do not become criminals, shows that a far more important and immediate cause must exist. Perhaps criminal behavior has a much simpler explanation: poor reasoning, leading to a wrongly perceived loss function. Whatever our early childhood experiences, law abiding citizens have just the same motivations as do criminals; all of us have felt the urge to commit robbery, assault, and murder. The difference is that the criminal does not think ahead far enough to appreciate the predictable consequences of his actions.

Inability to perceive one's own loss function can have disastrous personal consequences in other ways. Consider the case of Ramanujan, whom many would consider to be, in one particular area, the greatest mathematical genius who ever lived. His death at age 32 was probably the result of his own ridiculous dietary views. He refused to eat the food served in Hall at Trinity College, Cambridge (although it was undoubtedly more wholesome than any food he had ever eaten before coming to England) and tried to subsist on rotten fruit shipped from India without refrigeration.

A strikingly similar case is that of Kurt Gödel, whom many would consider the greatest – certainly the best known – of all logicians. He died of starvation in a hospital with the finest food facilities, because he became obsessed with the idea that the doctors were trying to poison him. It is curious that the greatest intellectual gifts sometimes carry with them the inability to perceive simple realities that would be obvious to a moron.

A New Look at the Jeffreys Prior

Our noting that the optimal decision depends only on the product of prior probability and loss function sets off several other lines of thought. As we noted in Chapter 12, Jeffreys (1939) proposed that, in the case of a continuous parameter α known to be positive, we should express prior ignorance by assigning, not uniform prior density, but a prior density proportional to $(1/\alpha)$. The theoretical justification of this rule was long unclear; but it yields very sensible–looking results in practice, which led Jeffreys to adopt it as fundamental in his significance tests. We saw in Chapter 12 that, in the case that α is a scale parameter, the Jeffreys prior is uniquely determined by invariance under the scale transformation group; but now we can see a quite different justification for it.

If we use the absolute error loss function $|\beta - \alpha|$ when α is known to be positive, then to assign $g(\alpha) = \text{const.}$ in (13–24) and (13–33) amounts to saying that we demand an estimator which yields, as nearly as possible, a constant absolute accuracy for all values of α in $0 < \alpha < \infty$. That is clearly asking for too much in the case of large α; and we must pay the price in a poor estimate for small α. But the median of Jeffreys' posterior distribution is mathematically the same thing as the optimal estimator for uniform prior and loss function $|\beta - \alpha|/\alpha$; we ask for, as nearly as possible, a constant *percentage* accuracy over all values of α. This is, of course, what we do want in most cases where we know that $0 < \alpha < \infty$. Another reason for the superior performance of Jeffreys' rule is thus made apparent, if we re–interpret it as saying that the $(1/\alpha)$ factor is part of the loss function.

Decision Theory is not Fundamental

What parts of the theories expounded here will be a permanent part of human thinking, what parts may evolve on into different forms in the future? We can only speculate, but it seems clear to the writer that there is something necessary and timeless in the methods of inference developed here; not only their compelling theoretical basis[†] explained in Chapters 1 and 2, but equally well, the beautiful way they work out in practice in all the later Chapters – always giving us the right answer to whatever question we ask of them, while orthodox methods yield sense and nonsense about equally often – convinces us that these methods cannot be altered in any substantive way in the future.

However, views as to the foundation of those methods may change; for example, instead of our desiderata of logical consistency, future workers may prefer desiderata of optimal information processing, as suggested by the work of Zellner (1988). Indeed, many advantages would result from more common recognition that inference has fundamentally nothing to do with 'randomness' or 'chance' but is concerned rather with optimal *processing of information*. We noted at the end of Chapter 2 how Gödel's theorem appears as a platitude rather than a paradox, as soon as we recognize the information processing aspect of mathematics.

But we can feel no such certainty about the decision theory addendum to inference. In the first place, many present applications already require an extension to, or in the direction of, game theory or dynamic programming. The State of Nature may be chosen by another person; or it may be influenced by our decision without the intervention of a conscious second agent. There may be more than two agents involved. They might be either adversaries or helpful friends. Those are more complicated situations than the ones we have considered here. We do not think such extensions appropriate to our present topic of Scientific Inference, because we do not think of ourselves as playing an adversary game against Nature. However, future scientists may find good reasons to consider the more general theory.

For all the reasons noted in this Chapter, it now appears that from a fundamental standpoint loss functions are less firmly grounded than are prior probabilities. This is just the opposite of the view that propelled the Wald–inspired development of decision theory in the 1950's, when priors were regarded as vague and ill–defined, but nobody seemed to notice that loss functions are far more so. For reasons we cannot explain, loss functions appeared to workers at that time to be "real" and definite, although no principles for determining them were ever given, beyond the platitude that any function with a continuous derivative appears linear if we examine a sufficiently small piece of it.

In the meantime, there have been several advances in the technique for assigning priors by logical analysis of the prior information; but to the best of our knowledge we have as yet no formal principles at all for assigning numerical values to loss functions; not even when the criterion is purely economic, because the utility of money remains ill–defined.

Another Dimension?

There is another respect in which loss functions are less firmly grounded than are prior probabilities. We consider it an important aspect of "objectivity" in inference – almost a principle of morality – that we should not allow our opinions to be swayed by our desires; what we believe should be independent of what we want. But the converse need not be true; on introspection, we would probably agree that what we want depends very much on what we know, and we do not feel guilty of any inconsistency or irrationality on that account.

Indeed, it is clear that the act of assigning a loss function is itself only a means of describing certain *prior information* about the phenomena of interest, which now notes not just their plausibilities, but also their consequences. Thus a change in prior information which affects the prior probabilities could very well induce a change in the loss function as well.

But then, having admitted this possibility, it appears that value judgments need not be introduced in the form of loss functions at all. Already at the end of Chapter 1 we noted the possibility

[†] Of course, better proofs than those we were able to give in Chapter 2 will be found.

of future 'multi–dimensional' models of human mental activity. In view of the above considerations, the doors now seem wide open for new developments in that direction; representing a mental state about a proposition or action not by one coordinate (plausibility) as in present probability theory; but by two coordinates (plausibility and value). Thus, while the principles of 'one–dimensional' inference seem permanent, the future can still bring many kinds of change in the representation of value judgments, which need not resemble present decision theory at all. But this in turn reacts back on the question of foundations of probability theory.

Thomas Bayes (1763) thought it necessary to explain the notion of probability in terms of that of expectation;† and this persisted to modern times in the work of both de Finetti (1972, 1974) and Wald (1950). At first glance, it appears that the work of de Finetti on foundations of probability theory could hardly be more different in outlook from Wald's decision theory; yet these two avenues to Bayesianity shared the common premise that value judgments are in some way primary to inference.

de Finetti would base probability theory on the notion of "coherence", which means roughly that in betting one should behave as if he assigned probabilities to the events (dice tosses, *etc.*) being betted on; but those probabilities should be chosen so that he cannot be made a sure loser, whatever the final outcome of those events.

It has always seemed objectionable to some, including this writer, to base probability theory on such vulgar things as betting, expectation of profit, *etc.* Somehow, we think that the principles of logic ought to be on a higher plane. But that was only an aesthetic feeling; now in recognizing the indefinite and provisional nature of loss functions we have a more cogent reason for not basing probability theory on decisions or betting. Any rules which were found to be coherent but not consistent, would be unusuable in practice because a well–posed question would have more than one "right" answer with nothing to choose between them. This is, in our view, still another aspect of the superiority of Richard Cox's approach, which stresses logical consistency instead and, just for that reason, is more likely of a lasting place in probability theory.

† The difficulty of reading Bayes today can be appreciated from the bewildering sentence in which he states this: "The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening."