E. T. JAYNES
Associate Professor of Physics
Washington University

New Engineering Applications of Information Theory

SUMMARY

A problem of decision making in the face of uncertainty is formulated and solved by using the notion of entropy as a criterion for setting up prior probability assignments. The resulting mathematical formalism is identical with that given by Gibbs for Statistical Mechanics, here applied in an entirely different context. This use of probability theory (similar to that of Laplace) would, on the usual viewpoint of statisticians, give rise to many objections and conceptual difficulties. We therefore give a brief survey of statistics, showing how recent developments have vindicated the original methods and viewpoint of Laplace, and made them available for treatment of problems which in most recent textbooks are considered to be outside the field of probability theory.

Offprints From
Proceedings Of The First Symposium On Engineering, etc.,
Edited By Bogdanoff and Kozin
Published By John Wiley & Sons, Inc., 1963

The title of this talk is quite ambiguous. In the first place, there is no common agreement as to the meaning of "Information Theory." What area of activity does it define? In the published literature we can find all kinds of answers to this, ranging from the narrowest possible view that "Information Theory" is synonymous with "Communication Theory" to definitions so broad and vague that they seem to include all human activity.

Taking a stand somewhere between these extremes, I will use the term "Information Theory" as standing for any application of probability theory in which we make use of the notion of entropy as a measure of "amount of uncertainty." I won't try to say just how much of probability theory is thereby staked out, because Information Theory, so defined, is a rapidly growing field.

Entropy is, in a sense, inherent in probability theory, independently of the work of Boltzmann, Gibbs, Szilard, von Neumann, and Shannon. For example, if we write Bayes' theorem in the logarithm-of-odds form [1] which has become popular in recent years (due to Wald's introduction [2] of the probability-ratio test in sequential analysis), we find that the expressions resulting are really conditional entropies. What is new in the past decade is merely the recognition of a situation that has always existed.

The notion of entropy is assuming an ever-increasing importance in statistics generally. Although new results obtained by its use are to date rather modest, it has turned out to have great value as a unifying principle by which we can see old results in a new light. This has been shown particularly by S. Kullback [3], who demonstrates that many of the procedures which had been developed by statisticians in a more or less ad hoc way for treatment of special problems, have a simple interpretation in terms of Information Theory. This new insight enables him to generalize old methods and in some cases (particularly the Chi-squared test) to improve them

At the same time, some of the famous paradoxes and controversies about interpretation (the subjective-objective nature of probability, etc,) which have plagued probability theory for two centuries, tend to disappear when we adopt the viewpoint suggested by Information Theory. It is, in my view, an open question whether "Information Theory," as defined here, may eventually become synonymous with "Probability Theory."

Another ambiguous thing in my title is the word "new." Is this meant in the strong sense that the application itself is new; i. e. the problem could not have been treated at all without Information Theory? Or in the weak sense that what is new is merely the recognition that Information Theory has something relevant to say about the problem? In spite of many attempts, I have not been able to imagine any problem which can be solved by using Information Theory, which would be absolutely impossible to solve without it. Information Theory has only emphasized the truth in Laplace's famous remark, "Probability Theory is nothing but Common Sense Reduced to Calculation." If we have enough common sense, we may find that we don't need any mathematical theory to tell us what to do. Indeed, the various methods of statistical inference were developed in the past in just that way; each of them is a mathematical model chosen so that it reproduces a small part of what we call common sense. So, we have to understand "new" in the weak sense; the criterion is not whether information theory is necessary, but only whether it is helpful.

The engineer is continually faced with the problem of making decisions in the face of uncertainty. Let us define his job broadly, and perhaps facetiously, as the planning of gadgets or procedures which are to work predictably and are to be in some way useful. Obviously, a certain area of knowledge, about the laws of physics and about the ways of people, is essential to him. But an equally essential part of his problem is the inevitable state of ignorance in which he must work. He does not know in advance exactly what quality of materials will be used, how faithfully his designs will be reduced to practice, whether the ultimate user will actually use the gadget in the way he has visualized. If his brainchild should fail, he never knows in advance exactly what the consequences of that failure will be (although he can often make a pretty good guess!) Nevertheless, his job is to do the best he can; to make the best possible use of his positive knowledge in such a way as to minimize, in some sense, the possible bad effects of his ignorance.

It is obvious enough that failure to take into account all the available knowledge relevant to a problem could be diastrous. Perhaps the main point I want to make in this talk is the converse; failure to recognize frankly the full extent of our ignorance, and to take this ignorance explicitly into account, can be just as diastrous. More specifically, I want to show you a mathematical model of the reasoning of a person who is trying to be fair, trying to avoid prejudice and remain noncommittal when he does not know; in other words, trying to avoid drawing conclusions which are not warranted by the available evidence. This model may, I hope, prove useful in engineering problems of a certain intermediate degree of complexity, which we can describe loosely by saying that the problem is sufficiently complicated so that our unaided common sense fails us, but at the same time is sufficiently simple so that the situation can be described by a manageable amount of mathematics.

To fix ideas more clearly, let us look for a moment at a specific, and not too unrealistic, decision problem which might arise. Mr. A is in charge of a Widget factory, which proudly advertises that it can make delivery in 24 hours on any size order. This, of course, is not really true, and Mr. A's job is to protect, as best he can, the Advertising Manager's reputation for veracity. This means that each morning he must decide whether the day's run of 200 Widgets will be painted red, yellow, or green. (For complex technological reasons, not relevant to the present problem, only one color can be produced per day). We follow his problem of decision making through several stages of increasing knowledge.

Stage 1 When he arrives at work, Mr. A's positive knowledge is that he has in stock 100 red Widgets, 150 yellow, and 50 green. His ignorance lies in the fact that he does not know how many orders for each type will come in during the day. Clearly, in this state of ignorance, Mr. A will attach the highest significance to any tiny scrap of information about orders likely to come in today; and if no such scraps are to be had, we do not envy Mr. A his job. Still, if a decision has to be made on no more information than this, his common sense will probably tell him that he had better build up that stock of green widgets.

Stage 2 Mr. A learns that, averaged over the past year, average daily orders have been for 50 red Widgets, 100 yellow, and 10 green. He will, I think, immediately decide to make yellow ones today, and probably red ones tomorrow.

Stage 3 But now Mr. A (who is evidently new on this job) learns that the average individual order is for 75 Widgets if the customer wants red ones, while users of yellow widgets order on the average only 10 at a time, and the average order for green is 20. This new information does not change the expected daily demand; but if Mr. A is very shrewd, I think he may change his mind again, and decide to make red Widgets today and almost certainly yellow ones tomorrow.

Stage 4 Finally, Mr. A gets a phone call, telling him that an emergency order for 40 green widgets is on its way by special messenger. Up to this point, Mr. A's decision problem has been so simple that he needed no mathematics, only ordinary common sense, to solve it. But now, I think he is in a position where some mathematics might be welcome. Let us summarize the various stages of this problem in a table:

	R	Y	G	Decision
1 In stock	100	150	50	G
2 Av. Daily Order Total	50	100	10	Y
3 Av. Individual Order	75	10	20	R
4 Specific Order			40	?

Mr. A has a fair chance of getting through today without trouble. But, no matter what decision he makes today, he is likely to be in trouble tomorrow. Suddenly, he realizes that this job is not as simple as it seemed. In order to think through his problem completely, he must not plan only for today - each morning his decision should be based on what is, at that time, the best estimate he can make of orders for the entire future.

Now it is not obvious how, or even whether, probability theory can be applied to this kind of problem. It is, in fact, so far from obvious that in the late 1940's a general theory of decision making in the face of uncertainty was developed, largely by Wald [4], which in its initial stages had no apparent connection with probability theory. I would like to give you a very brief account of some of the ideas it involved.

We begin by imagining (i. e. enumerating) a set of possible unknown "states of Nature," $\theta_1, \theta_2, \dots \theta_N$, whose number might be finite or infinite. In Mr. A's problem these correspond to the different possible orders which might come in *as far as he knows*. If we consider first the "truncated" problem (already noted as too simplified to be realistic) where decisions are made on a day-to-day basis with no thought of tomorrow, then to each state of nature there corresponds an ordered triple of non-negative integers $\theta = \{n_1, n_2, n_3\}$ giving respectively the total orders for red, yellow and green Widgets that will come in today. We have in this case an infinite, but discrete, set of θ_j. Of course, the θ_j might also form a continuum.

Already at this stage we can see a feature which has not been emphasized in the literature, but which is quite important for the viewpoint I want to develop. In enumerating the different states of nature, we are not describing any objective (measurable) property of nature - for, in fact, one and only one of them will be realized. The enumeration is only a means of describing our state of ignorance. It is meaningless to ask whether one particular enumeration is "correct" without first asking, "what is the prior information that is being described by the set of θ_j?" Two observers with different amounts of prior information may enumerate the θ_j differently without either being inconsistent. The rules of this game are simply that each observer must do the best he can on the basis of the information he has. At this stage, these remarks may seem trite; but bear with me.

The next step in our theory is to make a similar enumeration of the possible decisions $\{D_1 \; D_2 \; \dots \; D_k\}$ that might be made. In Mr. A's truncated problem there are only three possible decisions:

D_1 = "make red widgets today"
D_2 = "make yellow widgets today"
D_3 = "make green widgets today"

Again, the enumeration of the D_i is a means of describing our knowledge as to what kinds of actions are feasible; it is idle 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 left 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 sharp left turn does not always lead to diaster.

This suggests the third concept we need - the loss function $L(D_i, \theta_j)$, which is a set of numbers representing our judgement 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 becomes a loss matrix L_{ij}. In Mr. A's truncated problem L_{ij} has three rows, but an infinite number of columns.

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. The material we need for our purposes has been summarized in a very readable and entertaining form by Luce and Raiffa [5] and more recently in the elementary textbook of Chernoff and Moses [6]. The minimax criterion is this: for each D_i find the maximum possible loss $M_i \equiv \max_j (L_{ij})$: then choose that D_i for which M_i is a minimum. The minimax criterion would be a reasonable one 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 [5]. But in the decision problems of the scientist or engineer 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 fails to take advantage of the favorable possibilities.

Equally unreasonable for us is the opposite extreme of the starry-eyed optimist who uses this "minimin" criterion: for each D_i find the minimum possible loss $m_i \equiv \min_j (\ell_{ij})$ and choose the D_i that makes m_i a minimum.

Evidently, a reasonable decision criterion for the scientist and engineer is, in some sense, intermediate between minimax and minimin. Many other criteria have been suggested which go by the names of maximin utility (Wald), α - optimism-pessimism (Hurwicz), minimax regret (Savage), insufficient reason (Laplace), etc. The usual procedure, as described in detail by Luce and Raiffa [5], has been to analyze any proposed criterion to see whether it satisfies about a dozen qualitative commonsense conditions such as (1) Transitivity: if D_1 is preferred to D_2, and D_2 preferred to D_3, then D_1 must be preferred to D_3, and (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 analysis, although straightforward, can become tedious. I will 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.

What is it that 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 entirely on this: How strongly do you believe that θ_3 is the true state of nature? How plausible is θ_3?

To a physicist or engineer it seems like a very small step - really only a rephrasing of the question - to ask next, "What is the probability P_3 that θ_3 is the true state of nature?" Not so to the statistician of the dominant school of thought, which Savage [7] has labeled "objectivist." To the objectivist, the word "probability" is synomous with, "long-run relative frequency in some random experiment." But then, 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 frequency definition of probability, we will have to 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 (which are really decision problems and as such are included in our general formulation) to a new field, Statistical Inference, which was regarded as distinct from probability theory.

LAPLACE'S THEORY

Laplace had a different conception of probability theory. To him, it was not merely a set of rules for calculating frequencies - it was also the "calculus of inductive reasoning." It was an extension of logic to the intermediate cases where propositions are neither proved or disproved, but the evidence affects their plausibility - a quantitative rendering of what our common sense perceives qualitatively.

More precisely, by "Laplace's theory" I mean the following. Denote various propositions by letters, A, B, C, etc., and read AB as the proposition "both A and B are true," $\bar{A}$ as "A is false." The symbol (A|B), which is a real number in the interval [0, 1], stands for "the probability of A, given B." These symbols are manipulated as follows:

$$(AB|C) = (A|BC)(B|C) \tag{1}$$

$$(A|B) + (\bar{A}|B) = 1 \tag{2}$$

All relations of probability theory can be derived by repeated application of these two fundamental rules. In particular, from the fact that the left-hand side of (1) is symmetric in A and B, we obtain _Bayes' theorem_:

$$(A|BC) = (A|C)\frac{(B|AC)}{(B|C)} \tag{3}$$

which represents the learning process, since it shows how the prior probability (A|C) is changed to the posterior probability (A|BC) as a result of acquiring new information B.

According to Laplace, it is legitimate to assign probabilities to any clearly stated proposition, and the above rules generate an idealized mathematical model of the process of plausible reasoning carried out by human brains. Recently, Polya [8] has demonstrated the complete qualitative correspondence between these rules and human common sense. It is easily seen that they include deductive logic as a special case.

In addition to the above rules, Laplace needed for applications (1) a principle by which initial probabilities are set up in starting a problem, and (2) a principle for converting final probabilities into a definite decision. Both of these were supplied in the early 18th century by James and Daniel Bernoulli respectively, and used by Laplace:

(1) Imagine an initial state of knowledge X where we have enumerated a set of mutually exclusive and exhaustive possible propositions $(A_1 --- A_n)$ about which a decision has to be made, but have not yet incorporated any other evidence. At this stage we assign equal probabilities $(A_i | X) = n^{-1}$ to the A_i. This is the "principle of insufficient reason."

(2) Assign the "utility," U_{ij} of making decision A_i if A_j should turn out to be true, and make that decision A_i which maximizes the expected utility

$$<U>_i = \sum_{j=1}^{n} U_{ij} (A_j | EX) \qquad (4)$$

over the posterior probabilities $(A_j | EX)$, where E is any additional evidence about the A_j. This is Daniel Bernoulli's principle of "moral expectation."

Now, this looks very much like a prescription for solving our decision problem (with only minor changes, such as recognition that the set of possible decisions is not necessarily in 1:1 correspondence with the set of possible states of nature). Laplace used these rules from about 1774, in decision problems of astronomy; given a set of astronomical observations, decide whether discrepancies indicate a new systematic effect worth working on, or whether they are merely experimental errors. This early use of decision theory led him to some of the most important discoveries in celestial mechanics.

OBJECTIVIST STATISTICS

For several decades, it has been fashionable to discredit Laplace's methods and to deny that probability theory has anything to do with inductive reasoning. One can find this attitude in almost all books on probability and statistics written in this century. Indeed, the whole program of Statistical Inference had the explicitly stated aim of avoiding the supposed mistakes of Laplace, by developing entirely new approaches. For parameter estimation many criteria were proposed, and the method of maximum likelihood assumed a position of central importance. This, however, is mathematically identical with application of Bayes' theorem, with the uniform prior probability assigned by insufficient reason; then choosing that alternative with the greatest posterior probability. Thus it is included in the above Laplace rules as the decision criterion corresponding to a utility function $U_{ij} = \delta(A_i, A_j)$; we care only about the chances of being right, and if we are wrong we don't care how wrong we are. (This characterization of maximum likelihood, incidentally, tells us exactly under what circumstances it is the appropriate method to use; i. e. target shooting).

One of the major advances in statistical practice in recent years has been the introduction of Wald's sequential method for quality-control testing. In the original 1947 exposition [2], there is no mention of Bayes' theorem, and in 1950 Feller [9] issued a stern warning against the use of Bayes' theorem in quality-control testing, on just the grounds noted above; the state of nature (here the condition of a particular machine) is not "random." But in that same year, I. J. Good [1] showed that Wald's sequential procedure is also mathematically identical with the application of Bayes' theorem with uniform prior probabilities, then deciding that an hypothesis is true if its posterior probability reaches a certain preassigned level. This is just the way Laplace was handling hypothesis testing problems in the 18th century. The only difference is that Wald's method amounts to calculating a monotonic function of the posterior probability, $f(P) = \log[P/(1-P)]$, instead of P.

Recently, I gave a lecture on sequential testing at a well-known University, and afterward I was approached by a student in a rather dazed condition. It turned out that he was a graduate student in Statistics, and was taking a course in sequential analysis. His instructor had repeatedly warned against the use of the thoroughly discredited methods of Laplace and Bayes, and was enthusiastic about this wonderful new approach that had finally solved the problem. To the student, it was a shattering experience to see that the methods he was taught to use were identical with the methods he had been taught not to use!

How about the modern formulation [4, 6] of decision theory? Here one defines a class of "admissible" decision rules which consists, in simple terms, of all those any sane person would ever consider using; a strategy is admissible if no other exists which is better for all states of nature. After rather long mathematical arguments, we find the following. If we agree that we will not include in our enumeration any state of nature which is known in advance to be impossible, then the class of admissible decision rules is identical with the class that can be found by carrying out the following steps:

(A) Assign a-priori probabilities $p_j = (\theta_j | X)$ to the states of nature θ_j.

(B) Digest any subsequent information E_1, E_2 etc. about the θ_j by repeated application of Bayes' theorem, resulting in the posterior probabilities $P_j = (\theta_j | XE_1E_2---)$.

(C) Make that decision which minimizes the expected loss

$$<L>_i = \sum_j L_{ij}P_j \quad . \tag{5}$$

But, you see, these are precisely the rules Laplace was advocating and using 180 years ago, and which a generation of statisticians has been taught are nonsense!

Different admissible decision rulse correspond to different assignments of prior probabilities and loss functions. Actually, however,. it is only their product $\lambda_{ij} = L_{ij}p_j$ that enters into the final decision, as we see if we substitute (3) into (5):

$$<L>_i = \sum_j L_{ij}(\theta_j | X) \frac{(E_1E_2 ...| \theta_j X)}{(E_1E_2 ... | X)}$$

$$= \sum_j \lambda_{ij} \frac{(E | \theta_j X)}{(E | X)} \quad . \tag{6}$$

Recognition of this is important for several reasons. In the first place, it shows that every admissible decision rule is identical with one arising from Bayes' theorem with uniform prior probabilities; as far as the mathematics is concerned, we can <u>postulate</u> uniform prior probabilities, and characterize the decision rules entirely by the loss functions. Conceptually, however, I think we would all reject this possibility - the prior probabilities do play an important part, and we should keep the freedom to insert prior information in a manner independent of the value judgments implied in a loss function. More important is this. Statisticians of the "objectivist" school of thought are so concerned about arbitrariness in prior probability assignments that they are unwilling to introduce them at all, unless they are also known frequencies. Mathematically, it is trivial to see that refusal to use prior probabilities at all is equivalent to assignment of uniform prior probabilities. But from Eqn. (6) we see a more illuminating fact: *If the final decision depends strongly on which particular prior probability assignment we make, 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 should worry just as much about arbitrariness in the loss function.

We can sum up the foregoing in this way. In spite of a diametrically opposed viewpoint as to the nature of probability, the net result of advances in statistics over the past 40 years is that the statistician has finally returned to the original mathematical procedures of Laplace, from whence he started.

This trend becomes understandable when we recognize the following. Although the results are usually stated as a prediction of what would happen "on the average, " or "in the long run, " every application of probability theory or statistical inference to a *specific* situation is simply a problem of plausible reasoning. The trouble was that we were unwilling to accept Laplace's interpretations of Eqns. (1) and (2) as rules for carrying out plausible reasoning. But suppose now that Laplace was right after all, and Eqns. (1) and (2) are in fact the only consistent set of quantitative rules for plausible reasoning. Then, independently of any philosophy of interpretation, by the time the procedures of statistical inference had been made fully consistent, one would be forced to re-discover Laplace's methods. As soon as this situation was recognized, the distinction between probability theory and statistical inference would collapse.

Evidently, it becomes important to understand in just what sense probability theory may be said to be a "calculus" of plausible reasoning." The following result will be developed more fully elsewhere [10]. If there exists a satisfactory mathematical model of the process of plausible reasoning described by Polya [8], it seems reasonable to require of it three conditions: (A) representation of plausibilities by real numbers, (B) qualitative correspondence with common sense, (C) consistency. In an important contribution, Cox [11] has shown how the conditions of consistency of such a model may be stated in the form of functional equations, whose general solutions can be found. By a slight extension of this analysis, it can be shown that the three conditions above lead uniquely to Eqns. (1) and (2).

We conclude that the principles of plausible reasoning given to us by Laplace could be evaded only by developing a "lattice theory" in which condition (A) was abandoned. However, no such attempt is made in current statistical practice. Although many new concepts have been introduced, such as likelihood, efficiency, confidence level, significance level, etc., all of these are attempts to represent degrees of plausibility by real numbers.

Now, the objectivist viewpoint leads one to consider mathematical problems of such magnitude that the situation could be uncovered only by the herculean efforts of Abraham Wald. Laplace's viewpoint leads to the same final results by arguments so short and elementary that they can be understood by anyone familiar with high-school algebra. In the face of this, does it really make sense to continue saying that Laplace's viewpoint was naive and should be avoided? I think it is clear that a return to Laplace's viewpoint would bring about a great simplification and unification of this field. The greatest advantages would be in pedagogy--in a one-year undergraduate course we would give a young scientist or engineer the basis of all statistical practice, in a form which he could apply at once to his own problems.

The "objectivist" is anxious that his assertions shall be limited to objective statements of fact; hence the emphasis on frequencies rather than "subjective" state of ignorance. It is the first stage of sophistication to want to do this. But there is a second stage of sophistication in which we realize that any such aim is doomed to failure. For, the only thing about which I can ever speak with certainty is not what is "really" true, but only *what is my state of knowledge*. Thus the Laplace viewpoint, far from being naive, is the only one which fully expresses the natural limitations on our search for objectivity. In Laplace's theory a probability assignment is "subjective" in the sense that it describes a state of knowledge, rather than any measurable property of the physical world; but it is completely "objective" in the sense that it is independent of the personality of the user. Two observers, given the same set of propositions to reason about and the same prior evidence about them, must assign the same probabilities to them.

In this connection it is important to notice that, given two propositions A and B, the probability (A|B) has no definite numerical value. Eqns. (1) and (2) determine only relations between different probabilities, and thus do not tell us whether our proposed probability assignments are "correct," but only whether they are mutually consistent. This corresponds exactly to the situation in deductive logic, where a syllogism does not tell us whether our assignment of truth-values is correct, but only whether it is consistent. A numerical value can be assigned to (A|B) only after we have enumerated the possible alternatives, if A should be false. In other words, we must define our "sample space," or "hypothesis space" $\{A_1 \ldots A_n\}$. This is as necessary in Laplace's theory as in the objectivist approach. As has been shown elsewhere [10], one large class of objections to Laplace's viewpoint which one finds in the recent statistical literature can be traced to the author's failure to realize this.

BACK TO INFORMATION THEORY

Laplace's statistical practice was limited by the fact that he had only one principle, insufficient reason, to set up prior probability assignments. He was not handicapped by this, because in his problems calculation soon showed the evidence for or against some hypothesis to be so overwhelming that it made very little difference which prior probability (and correspondingly, which loss function) he used. In refining the application to problems like that of Mr. A, where the evidence is not so clear, we have to be more careful about these questions.

Actually, there are very few problems where the prior information is really of the form required by insufficient reason - there is always some sort of vague prior knowledge which renders some possibilities more likely than others. What this means is that we need more principles, extensions of insufficient reason, if we are to treat a wide variety of problems in a fully consistent way. Every new such principle we can find will open up a new class of applications for probability theory.

I would like to suggest here that the notion of entropy provides one such extension of insufficient reason. It was necessary to take this long detour into statistics generally, because in the usual "objectivist" viewpoint, neither the principle of insufficient reason nor the principle that I want to advocate here would make any sense. If we insist that a probability assignment must stand for a positive assertion about relative frequencies, then there can be no justification for the principle of insufficient reason; the fact that I know nothing about the various possibilities is not enough to make them occur equally often!

It is only in terms of Laplace's viewpoint - that a probability distribution is not primarily an assertion about relative frequencies (although we do not deny that the probabilities may, in some cases, be numerically equal to relative frequencies), but is rather a means of describing a certain state of ignorance, that the principle makes sense. It is essential for my purposes, that we accept this change in interpretation, and I have tried to show that this would also be desirable in other parts of statistics.

Please forgive me if I seem to be belaboring this point; but for several decades the statistical literature has been filled with objections to Laplace's use of insufficient reason, which arise from the author's failure to realize that for Laplace a probability assignment was not an assertion about frequencies. Connections between probability and frequency exist in many forms, analyzed in detail elsewhere [10]. In Laplace's theory, they have nothing to do with the definition of probability. On the contrary, all such connections are deducible as mathematical consequences of probability theory, interpreted as a "calculus of inductive reasoning." For example, prediction of the frequency with which some event will happen "in the long run" is, in Laplace's theory, simply a problem of parameter estimation, not different in principle from any other. One must calculate the probability p(f)df that the frequency f will lie in the range df, whereupon the (mean ± standard deviation) of this distribution will provide what is in most cases a satisfactory estimate of frequency, and a statement about the reliability of the estimate. The results are for all practical purposes identical with what the objectivist statistician would obtain by the method of confidence intervals, which he interprets as giving an estimate of the unknown "true" probability.

In a large class of problems, which includes the case of independent repetitions of a random experiment, the mean value estimate of the frequency of some event is found to be numerically equal to the probability assigned to that event at a single trial; but the reliability of the prediction depends very much on other details. Indeed, if we insist that the probability is the frequency, we leave ourselves no way of describing the fine details of our state of knowledge, which determine the reliability of the prediction. This remark has an important bearing on the problems of phase transitions and turbulence in statistical mechanics [10].

Much of what I have said about the principle of insufficient reason applies also to the notion of entropy. However useful the mathematical expression

$$\Sigma\ p_i \log p_i$$

may be, the concept of entropy is still foreign to the objectivist viewpoint in statistics. Terms such as "information" or "amount of uncertainty" really have no place in his scheme of things, and entropy remains an unwelcome stranger.

It is only in terms of the Laplace viewpoint of probability that the notion of entropy assumes a natural position. Here the formulation as given to us by Laplace has long suffered from a lack of just such a quantity. Our criterion for determining prior probability assignments is, intuitively, just that the prior probabilities should describe our positive knowledge, but should not assume anything beyond that. In other words, prior probabilities should be those with the maximum entropy consistent with our prior knowledge.

The expression for entropy has long been used in statistical mechanics, and the statement that "the entropy of a system is a measure of our degree of ignorance as to its true state" can be traced back to Boltzmann. Ever since then this "subjective" interpretation of entropy has had its advocates among the physicists; but it has been either ignored, or else briefly mentioned and immediately rejected, in every textbook on statistical mechanics ever written. Thus it has never had the status of being the "official viewpoint." In spite of the well-known work by Szilard pointing out connections between entropy and information, most workers in statistical mechanics have remained reluctant to assign any such meaning to entropy, and so it was left for Shannon to show, in a context completely removed from thermodynamics, that the same expression has just the mathematical properties needed to make it a reasonable measure of "amount of ignorance."

I don't believe there is any really rigorous argument that proves that one must use this expression, rather than some other. At least, at present there is no way we could convince a person, who did not want to believe it, that this particular mathematical expression is singled out from all others to play this role. However, we have by now an abundance of heuristic arguments all leading to this conclusion. In this respect we are in exactly the position of Archimedes who found, about 500 B. C., by an ingenious mechanical argument, a formula for the volume of a sphere: $V = 4\pi r^3/3$. Archimedes recognized that his argument was not rigorous, but it was sufficiently convincing that his formula was generally adopted. Actually, it was not until about 1900 years later that a rigorous derivation was given, by Leibniz. I believe we are now in an exactly similar intermediate state, where we know the answer but have not yet found a rigorous proof that it is the answer. Whether such proof will be found tomorrow or a thousand years from now, I do not know; but for the present I propose to go ahead and use what we have.

All right, let us now get down to specific and constructive things. In stage 2 of Mr. A's decision problem, he was able to enumerate the different possible orders which might come in during the day, and he also knew the average daily order for each type of widget. In this state of knowledge, he cannot use insufficient reason to assign probabilities, because the knowledge of average values does give him some reason for preferring some possibilities to others. If he is to avoid jumping to unwarranted conclusions, he should assign probabilities to today's possible orders which incorporate this knowledge of average values but do not assume anything beyond that. If we accept, with Shannon, the expression $S = - \Sigma p_k \log p_k$ as the proper measure of uncertainty represented by any probability assignment, then the probability assignment which has maximum entropy subject to the prescribed average values, is the one which correctly describes Mr. A's state of knowledge at stage 2. This leads to a familiar mathematical problem solved in every textbook on statistical mechanics.

The average values, $<f_1(\theta)>, <f_2(\theta)>, \dots <f_m(\theta)>$ of several functions $f_i(\theta)$ of the state of nature θ are considered known. We assign probabilities p_j to the possible states of nature θ_j, in such a way as to maximize S subject to these constraints. The constraint $\Sigma_j p_j = 1$ is formally included by defining $f_o(\theta) \equiv 1$, and requiring $<f_o> = 1$. We introduce a Lagrange multiplier for each of the $f_j(\theta)$, and obtain the variational problem

$$\delta[S - \mu_o <f_o> - \lambda_1 <f_1> \dots - \lambda_m <f_m>] = 0 \qquad (7)$$

or

$$\delta \sum_j [\, p_j \log p_j + \mu_o p_j + \lambda_1 f_1(\theta_j) pk + \dots$$

$$+ \lambda_m f_m(\theta_j) p_j] = 0$$

The solution is that the probability assigned to the state of nature θ_j is of exponential form:

$$p_j = \exp[-\lambda_o - \lambda_1 f_1(\theta_j) - \lambda_2 f_2(\theta_j) - \dots - \lambda_m f_m(\theta_j)] \qquad (8)$$

where $\lambda_o \equiv 1 + \mu_o$; and in order to fix the values of the Lagrange multipliers λ_i, we must force this to agree with the prescribed average values:

$$\begin{aligned} <1> &= \Sigma\, p_j \\ <f_1> &= \Sigma\, f_1(\theta_j)p_j \\ <f_2> &= \Sigma\, f_2(\theta_j)p_j \\ &\dots\dots \end{aligned} \qquad (9)$$

All these steps are summarized most neatly if we define a function, which is called the partition function:

$$Z(\lambda_1 \dots \lambda_m) \equiv \sum_j \exp[-\lambda_1 f_1(\theta_j) - \dots - \lambda_m f_m(\theta_j)] \qquad (10)$$

In terms of this function, our conditions reduce to:

$$\begin{aligned} \lambda_o &= \log Z(\lambda_1 \dots \lambda_m) \\ <f_1> &= -\frac{\partial}{\partial\lambda_1} \log Z(\lambda_1 \dots \lambda_m) \\ <f_2> &= -\frac{\partial}{\partial\lambda_2} \log Z(\lambda_1 \dots \lambda_m) \dots\dots \end{aligned} \qquad (11)$$

Of course, you recognize that these rules of calculation are identical with the formalism of statistical mechanics, given to us by Gibbs. It was only through the work of Shannon that we could see what these rules meant intuitively, and thus how general was their applicability.

Let us solve Mr. A's truncated problem in stage 2, where the mathematical expectations $<n_1>$, $<n_2>$, $<n_3>$ of orders for red, yellow, and green widgets are given as 50, 100, 10 respectively. With three average values given, we will have three Lagrange multipliers λ_1, λ_2, λ_3, and the partition function is

$$Z(\lambda_1, \lambda_2, \lambda_3) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{n_3=0}^{\infty} \exp(-\lambda_1 n_1 - \lambda_2 n_2 - \lambda_3 n_3)$$

$$= \prod_{i=1}^{3} (1 - e^{-\lambda_i})^{-1} \qquad (12)$$

The λ_i are determined from Eqn. (11):

$$< n_i > = - \frac{\partial}{\partial \lambda_i} \log Z = \frac{1}{e^{\lambda_i} - 1} . \qquad (13)$$

The maximum-entropy probability assignment p_j for the states of nature $\theta_j = \{n_1 \; n_2 \; n_3\}$ factors:

$$p(n_1 \; n_2 \; n_3) = p_1(n_1) p_2(n_2) p_3(n_3) \qquad (14)$$

with

$$p_i(n_i) = (1 - e^{-\lambda_i})e^{-\lambda_i n_i} \quad , \quad n_i = 0,\ 1,\ 2,\ ,,,$$

$$= \frac{1}{< n_i > + 1} \left[\frac{< n_i >}{< n_i > + 1} \right]^{n_i} \qquad (15)$$

Thus in stage 2, Mr. A's state of knowledge about today's orders is given by

$$p_1(n_1) = \frac{1}{51}\left(\frac{50}{51}\right)^{n_1}$$

$$p_2(n_2) = \frac{1}{101}\left(\frac{100}{101}\right)^{n_2}$$

$$p_3(n_3) = \frac{1}{11}\left(\frac{10}{11}\right)^{n_3} \qquad (16)$$

Now in order to proceed we have to introduce a loss function. Mr. A's judgment is that there is no loss if all orders are filled today; otherwise the loss will be equal to the total number of unfilled orders. The present stock of red, yellow, and green widgets is $S_1 = 100$, $S_2 = 150$, $S_3 = 50$ respectively. On decision D_1 (make red widgets) the available stock S_1 will be increased by the day's run of 200 widgets, and the loss will be

$$L(D_1; n_1\, n_2\, n_3) = g(n_1 - S_1 - 200) + g(n_2 - S_2) + g(n_3 - S_3) \qquad (17)$$

where

$$g(x) \equiv \left\{ \begin{array}{ll} x, & x \geq 0 \\ 0, & x \leq 0 \end{array} \right\} \qquad (18)$$

Likewise, on decisions D_2, D_3 the loss will be

$$L(D_2; n_1 n_2 n_3) = g(n_1 - S_1) + g(n_2 - S_2 - 200) + g(n_3 - S_3) \qquad (19)$$

$$L(D_3; n_1 n_2 n_3) = g(n_1 - S_1) + g(n_2 - S_2) + g(n_3 - S_3 - 200) \qquad (20)$$

So, if decision D_1 is made, the expected loss will be

$$\begin{aligned} \langle L \rangle_1 &= \sum_{n_i} p(n_1 n_2 n_3) L(D_1; n_1 n_2 n_3) \\ &= \sum_{n_1} p_1(n_1) g(n_1 - S_1 - 200) + \sum_{n_2} p_2(n_2) g(n_2 - S_2) \\ &\quad + \sum_{n_3} p_3(n_3) g(n_3 - S_3) \end{aligned} \qquad (21)$$

and similarly for D_2, D_3.

The summation are elementary, giving

$$\langle L \rangle_1 = \langle n_1 \rangle e^{-\lambda_1 (S_1 + 200)} + \langle n_2 \rangle e^{-\lambda_2 S_2} + \langle n_3 \rangle e^{-\lambda_3 S_3}$$

$$\langle L \rangle_2 = \langle n_1 \rangle e^{-\lambda_1 S_1} + \langle n_2 \rangle e^{-\lambda_2 (S_2 + 200)} + \langle n_3 \rangle e^{-\lambda_3 S_3}$$

$$\langle L \rangle_3 = \langle n_1 \rangle e^{-\lambda_1 S_1} + \langle n_2 \rangle e^{-\lambda_2 S_2} + \langle n_3 \rangle e^{-\lambda_3 (S_3 + 200)}$$

$$(22)$$

To slide-rule accuracy we have

$$e^{-\lambda_1 S_1} = 0.133 \qquad e^{-200\lambda_1} = 0.175$$
$$e^{-\lambda_2 S_2} = 0.222 \qquad e^{-200\lambda_2} = 0.134 \tag{23}$$
$$e^{-\lambda_3 S_3} = 0.0085 \qquad e^{-200\lambda_3} = 0.5 \times 10^{-8}$$

and

$$\langle L \rangle_1 = 22.4 \quad \text{unfilled orders}$$
$$\langle L \rangle_2 = 9.7 \quad \text{unfilled orders} \tag{24}$$
$$\langle L \rangle_3 = 28.9 \quad \text{unfilled orders}$$

showing a strong preference for decision D_2, as Mr. A's common sense had already anticipated.

You will recognize that Stage 2 of Mr. A's decision problem is mathematically the same as the theory of the harmonic oscillator in quantum statistical mechanics. There is still another engineering application of the harmonic oscillator equations, in some problems of message encoding [12]. I am trying to emphasize the generality of this theory, which is mathematically quite old and well known, but which has been applied in the past only in some specialized problems in physics. This general applicability can be seen only after we are emancipated from the objectivist view that all probability distributions must be justified in the frequency sense. Historically, this made it appear to most workers in statistical mechanics that the methods of Gibbs could be justified only via unproved "ergodic hypotheses" (in spite of the fact that Gibbs himself never mentioned them). I have suggested [13] that we interpret Gibbs' equations not as assertions about frequencies but as examples of inductive reasoning. It is clear that the laws of inductive reasoning do not depend on ergodic theorems or any other aspect of the laws of physics - ergo, the methods of Gibbs can be applied to any problem of inductive reasoning where the given information can be stated in the form of mean values.

In stage 3 of Mr. A's problem we have some additional pieces of information, giving the average individual orders for red, yellow, and green widgets. This new information makes it expedient to set up a more detailed enumeration of the states of nature, in which we take into account not only the total orders for each type, but also the breakdown into individual orders. We could have done this also in stage 2, but since at that stage there was no information available, bearing on this breakdown, it would have added nothing to the problem. However, in the interest of checking the consistency of this theory, you may find it amusing to retrace stage 2 on this basis and see how it would have led to exactly the same results given above.

In stage 3, a possible state of nature can be described as follows. We receive u_1 individual orders for 1 red widget each, u_2 orders for 2 red widgets each, . . . u_r individual orders for r red widgets each. Also, we receive v_y orders for y yellow widgets each, and w_g orders for g green widgets each.Thus a state of nature is specified by an infinite number of non-negative integers:

$$\theta = \{ u_1 \, u_2 \ldots ; v_1 \, v_2 \ldots ; w_1 \, w_2 \ldots \} \tag{25}$$

and conversely every such set of integers represents a conceivable state of nature, to which we assign a probability $p(u_1\ u_2\ \ldots;\ v_1\ v_2\ \ldots;\ w_1\ w_2\ \ldots)$.

Today's total demand for red, yellow and green widgets is, respectively

$$\begin{aligned} n_1 &= \sum_{r=1}^{\infty} r\, u_r \\ n_2 &= \sum_{y=1}^{\infty} y\, v_y \qquad\qquad (26) \\ n_3 &= \sum_{g=1}^{\infty} g\, w_g\ , \end{aligned}$$

the mathematical expectations of which were given before as $\langle n_1 \rangle = 50$, $\langle n_2 \rangle = 100$, $\langle n_3 \rangle = 10$. The total number of individual orders for red, yellow and green widgets are respectively

$$\begin{aligned} m_1 &= \sum_{r=1}^{\infty} u_r \\ m_2 &= \sum_{y=1}^{\infty} u_y \qquad\qquad (27) \\ m_3 &= \sum_{g=1}^{\infty} w_g \end{aligned}$$

And the new feature of stage 3 is that $\langle m_1 \rangle$, $\langle m_2 \rangle$, $\langle m_3 \rangle$ are also known. For example, the statement that the average individual order for red widgets is 75, means that $\langle n_1 \rangle = 75 \langle m_1 \rangle$.

With six average values given, we will have six Lagrange multipliers $\{\lambda_1\ \mu_1;\ \lambda_2\ \mu_2;\ \lambda_3\ \mu_3\}$. The maximum-entropy probability assignment will have the form

$$p(u_1 u_2 \ldots ; v_1 v_2 \ldots ; w_1 w_2 \ldots) = \exp(-\lambda_o - \lambda_1 n_1 - \mu_1 m_1 - \lambda_2 n_2 - \mu_2 m_2 - \lambda_3 n_3 - \mu_3 m_3)$$

which factors:

$$p(u_1 u_2 \ldots ; v_1 v_2 \ldots ; w_1 w_2 \ldots) = p_1(u_1 u_2 \ldots)\, p_2(v_1 v_2 \ldots) \cdot p_3(w_1 w_2 \ldots) \tag{28}$$

The partition function also factors:

$$Z = Z_1(\lambda_1 \mu_1)\, Z_2(\lambda_2 \mu_2)\, Z_3(\lambda_3 \mu_3) \tag{29}$$

with

$$Z_1(\lambda_1 \mu_1) = \sum_{u_1=1}^{\infty} \sum_{u_2=1}^{\infty} \ldots \exp[-\lambda_1(u_1 + 2u_2 + 3u_3 + \ldots) - \mu_1(u_1 + u_2 + u_3 + \ldots)] = \prod_{r=1}^{\infty} \frac{1}{1 - e^{-r\lambda_1 - \mu_1}} \tag{30}$$

with similar expressions for Z_2, Z_3. To find λ_1, μ_1 we apply the general rule, Eqn. (11):

$$<n_1> = \frac{\partial}{\partial \lambda_1} \sum_{r=1}^{\infty} \log(1 - e^{-r\lambda_1 - \mu_1}) = \sum_{r=1}^{\infty} \frac{r}{e^{r\lambda_1 + \mu_1} - 1} \tag{31}$$

$$< m_1 > = \frac{\partial}{\partial \mu_1} \sum_{r=1}^{\infty} \log(1 - e^{-r\lambda_1 - \mu_1}) = \sum_{r=1}^{\infty} \frac{1}{e^{r\lambda_1 + \mu_1} - 1} \tag{32}$$

Comparing with Eqns. (26) and (27), we see that

$$<u_r> = \frac{1}{e^{r\lambda_1 + \mu_1} - 1} \tag{33}$$

and now the secret is out - Stage 3 of Mr. A's decision problem is just the theory of an ideal Bose-Einstein gas in quantum statistical mechanics! The index r corresponds to the r'th single-particle energy level, u_r to the number of particles in the r'th state, λ_1 and μ_1 to the temperature and chemical potential.

In the present problem it is clear that for all r, $<u_r> << 1$, and that $<u_r>$ cannot decrease appreciably below $<u_1>$ until r is of the order of 75, the average individual order. Therefore, μ_1 will be numerically large, and λ_1 numerically small, compared to unity. This means that the series (31), (32) converge very slowly and are useless for numerical work. However, we can transform them into rapidly converging ones as follows:

$$\sum_{r=1}^{\infty} \frac{1}{e^{\lambda r + \mu} - 1} = \sum_{r=1}^{\infty} \sum_{n=1}^{\infty} e^{-n(\lambda r + \mu)}$$

$$= \sum_{n=1}^{\infty} \frac{e^{-n(\mu + \lambda)}}{1 - e^{-n\lambda}} \tag{34}$$

The first term is already an excellent approximation. Similarly,

$$\sum_{r=1}^{\infty} \frac{r}{e^{\lambda r + \mu} - 1} = \sum_{n=1}^{\infty} \frac{e^{-n(\mu + \lambda)}}{(1 - e^{-n\lambda})^{2}} \qquad (35)$$

and so (31) and (32) become

$$< n_1 > \simeq \frac{e^{-\mu_1}}{\lambda_1^{2}} \qquad (36)$$

$$< m_1 > \simeq \frac{e^{-\mu_1}}{\lambda_1} \qquad (37)$$

or

$$\lambda_1 \simeq \frac{< m_1 >}{< n_1 >} = \frac{1}{75} \qquad (38)$$

$$e^{\mu_1} \simeq \frac{< n_1 >}{< m_1 >^{2}} = 112 \qquad (39)$$

$$\mu_1 = 4.72 \qquad (40)$$

Looking back, we see that the error in the approximate formulas (36), (37) is less than 0.5 percent.

The probability that u_r has a particular values is, from (28) or (30)

$$p(u_r) = (1 - e^{r\lambda_1 - \mu_1})e^{-(r\lambda_1 + \mu_1)u_r}$$

which has the mean value (33) and the variance

$$\mathrm{var}(u_r) = \langle u_r{}^2 \rangle - \langle u_r \rangle^2 = \frac{e^{r\lambda_1 + \mu_1}}{(e^{r\lambda_1 + \mu_1} - 1)^2} \tag{41}$$

The total demand for red widgets

$$n_1 = \sum_{r=1}^{\infty} r u_r$$

is expressed as the sum of a large number of independent "random variables." The probability distribution for n_1 will have the mean value (36) and the variance

$$\mathrm{var}(n_1) = \sum_{r=1}^{\infty} r^2 \mathrm{var}(u_r) = \sum_{r=1}^{\infty} \frac{r^2 e^{r\lambda_1 + \mu_1}}{(e^{r\lambda_1 + \mu_1} - 1)^2} \tag{42}$$

which we convert into the rapidly convergent sum

$$\sum_{r,n=1}^{\infty} n r^2 e^{-n(r\lambda+\mu)} = \sum_{n=1}^{\infty} n \frac{e^{-n(\lambda + \mu)} \; e^{-n(2\lambda + \mu)}}{(1 - e^{-n\lambda})^3} \tag{43}$$

or, approximately,

$$\mathrm{var}(n_1) \simeq \frac{2e^{-\mu_1}}{\lambda_1{}^3} = \frac{2}{\lambda_1} \langle n_1 \rangle . \tag{44}$$

By analogy with the central limit theorem, the probability distribution for n_1 will be very nearly gaussian:

$$p(n_1) \simeq \left(\frac{\lambda_1}{4\pi \langle n_1 \rangle}\right)^{1/2} \exp\left\{-\frac{\lambda_1 (n_1 - \langle n_1 \rangle)^2}{4\langle n_1 \rangle}\right\} \qquad (45)$$

for those values of n_1 which can arise in many different ways. For example, the case $n_1 = 2$ can arise in only two ways: $u_1 = 2$, or $u_2 = 1$, all other u_r being zero. On the other hand, the case $n_1 = 150$ can arise in an enormous number of different ways, and the "smoothing" mechanism of the central limit theorem can operate. Thus, Eqn. (45) is a good approximation for the large values of n_1 of interest to us, but it is a very poor approximation for small n_1.

The expected loss on the various decisions is, as we saw in Eqn. (22), the sum of three terms arising from failure to meet orders for red, yellow, or green widgets respectively. If we do not make red widgets today, then the possibility of failing to meet orders for red widgets contributes to the expected loss the amount

$$\sum_{n_1=0}^{\infty} p(n_1) g(n_1 - S_1) \simeq \left[\frac{\lambda_1}{4\pi \langle n_1 \rangle}\right]^{\frac{1}{2}} \int_{S_1}^{\infty} (n_1 - S_1) \exp\left\{-\frac{\lambda_1 (n_1 - \langle n_1 \rangle)^2}{4\langle n_1 \rangle}\right\} dn_1$$

$$= \frac{\langle n_1 \rangle - S_1}{2}\left[1 + \operatorname{erf} \alpha_1 (\langle n_1 \rangle - S_1)\right] + \frac{1}{2\alpha_1 \sqrt{\pi}} \exp\left[-\alpha_1^2 (\langle n_1 \rangle - S_1)^2\right]$$

(46)

where $\alpha_1^2 = \lambda_1 / 4\langle n_1 \rangle$, and erf x is the error function

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dx. \tag{47}$$

If we do decide to make red widgets today, the possibility of failing to meet red orders contributes to the expected loss the above expression (46) with S_1 replaced by $(S_1 + 200)$.

Similar equations hold for yellow and green widgets. Although the approximations we made are not equally good in all cases, let us use Eqn. (46) for the partial losses and apply it three times with the given numerical values

$$S_1 = 100, \quad S_2 = 150, \quad S_3 = 50$$

$$\langle n_1 \rangle = 50, \quad \langle n_2 \rangle = 100, \langle n_3 \rangle = 10$$

$$\alpha_1 = 0.0082, \alpha_2 = 0.016, \quad \alpha_3 = 0.035.$$

Doing the indicated calculations, we find that on the decisions D_1, D_2, D_3, the expected losses are

$$\langle L \rangle_1 = (0) + 2.86 + 0.18 = 3.04 \text{ unfilled orders}$$

$$\langle L \rangle_2 = 14.9 + (0) + 0.18 = 15.1 \text{ unfilled orders}$$

$$\langle L \rangle_3 = 14.9 + 2.86 + (0) = 17.8 \text{ unfilled orders} \tag{48}$$

where (0) stands for a term orders of magnitude smaller than the others. The breakdown indicated is to be read as follows: If Decision D_1 (make red widgets) is made, there is negligible loss from the possibility of failing to meet red orders, while the possibility of failure with yellow orders leads to an expected loss of 2. 86, and only 0. 18 for green.

The results show the great preference for D_1 caused by the additional information about average individual orders, which had the intuitive effect of making the situation with respect to yellow widgets much safer than it seemed in stage 2.

It is in passage from stage 3 to stage 4 (where the new information consists of a specific order for 40 green widgets), that our common sense first fails us. Now both the red and green situations seem rather precarious, and our common sense lacks the "resolving power" to tell which is the more serious. Strangely enough, this new knowledge, which makes the problem so hard for our common sense, causes no difficulty at all in the mathematics. The previous equations still apply, with the sole difference that the stock S_3 of green widgets is reduced from 50 to 10. We now have $(<n_3> - S_3) = 0$ so that (46) reduces to

$$\frac{1}{2\alpha_3\sqrt{\pi}} = 8.08$$

and in place of (48) we have

$$<L>_1 = (0) + 2.86 + 8.08 = 10.94$$

$$<L>_2 = 14.9 + (0) + 8.08 = 23.0$$

$$<L>_3 = 14.9 + 2.86 + (0) = 17.8 \qquad (49)$$

So, Mr. A. should stick to his decision to make red widgets! Our common sense fails just because there is now so little difference between $< L>_1$ and $<L>_3$.

I have tried to show that use of probability theory in the sense of Laplace, with prior probabilities determined by the principle of maximum entropy, leads to a reasonable method of treating decision problems and to results in good correspondence with common sense. Mathematically, our equations are nothing but the Gibbs formalism in statistical mechanics, the only new feature being the recognition that the Gibbs methods are of far more general applicability than had been supposed.

The moral of this is simply that questions about "interpretation of a formalism," which the positivist philosophy tends to reject as meaningless and useless, are on the contrary of central importance in scientific work. It is, of course, true that in an application already established, a different interpretation of the equations cannot lead to any new results. But our judgment as to the range of validity of a formalism can depend entirely on how we interpret it. The interpretation (probability) ≡ (frequency) has led to a great and unnecessary restriction on the kinds of problem where probability theory can be applied. The scientist or engineer today is faced with many problems which require the broader Laplace interpretation.

BIBLIOGRAPHY

1. Good, I. J., "Probability and the Weighing of Evidence," C. Griffin and Sons, London (1950).

2. Wald, A., "Sequential Analysis," J. Wiley and Sons, Inc., N. Y. (1947).

3. Kullback, S., "Information Theory and Statistics," J. Wiley and Sons, Inc., N. Y. (1959).

4. Wald, A., "Statistical Decision Functions," J. Wiley and Sons, Inc., N. Y. 1950).

5. Luce, R. D. and Raiffa, H., "Games and Decisions," J. Wiley and Sons, Inc., N. Y. (1957); Chap. 13.

6. Chernoff, H., and Moses, L., "Elementary Decision Theory," J. Wiley and Sons, Inc., N. Y. (1959).

7. Savage, L. J., "Foundations of Statistics," J. Wiley and Sons, Inc., N. Y. (1954).

8. Polya, G., "Mathematics and Plausible Reasoning," Princeton Univ. Press, (1945); Vol. II.

9. Feller, W., "An Introduction to Probability Theory and Its Applications," J. Wiley and Sons, Inc., N. Y. (1950); p. 85 (2nd edition)(1957); p. 114.

10. Jaynes, E. T., "Probability Theory in Science and Engineering," McGraw-Hill Book Co., N. Y. (in press).

11. Cox, R. T., "Probability, Frequency, and Reasonable Expectation," American Jour. Phys., Vol. 17, p. 1, (1946).

12. Jaynes, E. T., "Note on Unique Decipherability," I. R. E. Trans. on Information Theory, Vol. IT-5, p. 98 (September, 1959).

13. Jaynes, E. T., "Information Theory and Statistical Mechanics," Phys. Rev. Vol. 106, p. 620; Vol. 108, p. 171 (1957).