Prior Probabilities

Edwin T. Jaynes

Department of Physics, Washington University, St. Louis, Missouri

In decision theory, mathematical analysis shows that once the sampling distributions, loss function, and sample are specified, the only remaining basis for a choice among different admissible decisions lies in the prior probabilities. Therefore, the logical foundations of decision theory cannot be put in fully satisfactory form until the old problem of arbitrariness (sometimes called "subjectiveness") in assigning prior probabilities is resolved.

The principle of maximum entropy represents one step in this direction. Its use is illustrated, and a correspondence property between maximum-entropy probabilities and frequencies is demonstrated. The consistency of this principle with the principles of conventional "direct probability" analysis is illustrated by showing that many known results may be derived by either method. However, an ambiguity remains in setting up a prior on a continuous parameter space because the results lack invariance under a change of parameter; thus a further principle is needed.

It is shown that in many problems, including some of the most important in practice, this ambiguity can be removed by applying methods of group theoretical reasoning which have long been used in theoretical physics. By finding the group of transformations on the parameter space which converts the problem into an equivalent one, a basic desideratum of consistency can be stated in the form of functional equations which impose conditions on, and in some cases fully determine, and "invariant measure" on the parameter space. The method is illustrated for the case of location and scale parameters, rate constants, and in Bernoulli trials with unknown probability of success.

In realistic problems, both the transformation group analysis and the principle of maximum entropy are needed to determine the prior. The distributions thus found are uniquely determined by the prior information, independently of the choice of parameters. In a certain class of problems, therefore, the prior distributions may now be claimed to be fully as "objective" as the sampling distributions.

I. Background of the problem

Since the time of Laplace, applications of probability theory have been hampered by difficulties in the treatment of prior information. In realistic problems of decision or inference, we often have prior information which is highly relevant to the question being asked; to fail to take it into account is to commit the most obvious inconsistency of reasoning and may lead to absurd or dangerously misleading results.

As an extreme example, we might know in advance that a certain parameter $\theta \leq 6$. If we fail to incorporate that fact into the equations, then a conventional statistical analysis might easily lead to the conclusion that the "best" estimate of θ is $\theta^* = 8$, and a shortest 90-percent confidence interval is $(7 \leq \theta \leq 9)$.

Few people will accept an estimate of a parameter which lies outside the parameter space, and so "orthodox" statistical principles such as efficient estimators or shortest confidence intervals can break down and leave no definite procedures for inference in the presence of this kind of prior information. Further examples of this phenomenon are given by Kendall and Stuart (1961).

With more "gentle" kinds of prior information, which do not absolutely exclude any interval for θ but only render certain intervals highly unlikely, the difficulty is less drastic but still present. Such cases are even more dangerous in practice because the shortcomings of orthodox principles, while just as real, are no longer obvious.

The Bayesian approach to statistics offers some hope of overcoming such difficulties since, of course, both the prior and the posterior distributions for θ will vanish outside the parameter space, and so the results cannot conflict with deductive reasoning. However, what determines the prior within the parameter space? After nearly two centuries of discussion and debate, we still do not seem to have the principles needed to translate prior information into a definite prior probability assignment.

For many years the orthodox school of thought, represented by most statisticians, has sought to avoid this problem by rejecting the use of prior probabilities altogether, except in the case where the prior information consists of frequency data. However, as the preceding example shows, this places a great restriction on the class of problems which can be treated. Usually the prior information does not consist of frequency data, but is nonetheless cogent. As Kendall and Stuart (1961) point out, this is a major weakness of the principle of confidence intervals.

With the rise of decision theory, this problem has assumed new importance. As we know, this development was started by Wald (1950) with the express purpose of finding a new foundation for statistics which would have the generality, but avoid the supposed mistakes, of the work of Bayes and Laplace. But after monumental labors, the mathematical situation uncovered by Wald finally led to a realization that the only consistent procedure of digesting information into the decision process is identical with application of Bayes' theorem, and that, once the loss function, sampling distribution, and sample are given, the only rational basis for choice among the admissible decisions lies in the prior probabilities.

Thus in modern decision theory, it appears that statistical practice has reached a level where the problem of prior probabilities can no longer be ignored or belittled. In current problems of engineering design, quality control, operations research, and irreversible statistical mechanics, we cannot translate the full problem into mathematical terms until we learn how to find the prior probability assignment which describes the prior information. In fact, as shown later, in some of the most important problems the prior information is the only information available, and so decisions must be based entirely on it. In the absence of any principle for setting up prior distributions, such problems cannot be treated mathematically at all.

The "personalistic" school of thought (Savage 1954,1962) recognizes this deficiency, but proceeds to overcompensate it by offering us many different priors for a given state of prior knowledge. Surely, the most elementary requirement of consistency demands that two persons with the same relevant prior information should assign the same prior probability. Personalistic doctrine makes no attempt to meet this requirement, but instead attacks it as representing a naive "necessary" view of probability, and even proclaims as one of its fundamental tenets (Savage 1954 p. 3) that we are free to violate it without being unreasonable. Consequently, the theory of personalistic probability has come under severe criticism from orthodox statisticians who have seen in it an attempt to destroy the "objectivity" of statistical inference by injecting the user's personal opinions into it.

Of course, no one denies that personal opinions are entitled to consideration and respect if they are based on factual evidence. For example, the judgment of a competent engineer as to the reliability of a machine, based on calculations of stresses, rate of wear, etc., is fully as cogent as anything we can learn from a random experiment; and methods of reliability testing which fail to take such information into account are not only logically inconsistent, but economically wasteful. Nevertheless, the author must agree with the conclusions of orthodox statisticians, that the notion of personalistic probability belongs to the field of psychology and has no place in applied statistics. Or, to state this more constructively, objectivity requires that a statistical analysis should make use, not of anybody's personal opinions, but rather the specific factual data on which those opinions are based.

An unfortunate impression has been created that rejection of personalistic probability automatically means the rejection of Bayesian methods in general. It will hopefully be shown here that this is not the case; the problem of achieving objectivity for prior probability assignments is not one of psychology or philosophy, but one of proper definitions and mathematical techniques, which is capable of rational analysis. Furthermore, results already obtained from this analysis are sufficient for many important problems of practice, and encourage the belief that with further theoretical development prior probabilities can be made fully as "objective" as direct probabilities.

It is sometimes held that this evident difference in the nature of direct and prior probabilities arises from the fact that the former have a clear frequency interpretation usually lacking in the latter. However, there is almost no situation of practice in which the direct probabilities are actually verified experimentally in the frequency sense. In such cases it is hard to see how the mere possibility of *thinking* about direct probabilities as frequencies in a nonexistent experiment can really be essential, or even relevant, to the problem.

Perhaps the real difference between the manifestly "public" nature of direct probabilities and the "private" nature of prior probabilities lies in the fact that in one case there is an established theory, accepted by all (*i.e.,* Bernoulli trials, etc.), which tells how to calculate them; while in the case of prior probabilities, no universally accepted theory exists as yet. If this view is correct, we would expect that with further development of probability theory, the distinction will tend to disappear. The two principles—maximum entropy and transformation groups—discussed in the following sections represent methods for calculating probabilities which apply indifferently to either.

II. The Basic Desideratum

To elaborate the point just made, a prior probability assignment not based on frequencies is necessarily "subjective" in the sense that it describes a state of knowledge, rather than anything which could be measured in an experiment. But if the methods are to have any relevance to science, the prior distributions must be completely "objective" in the sense that it is independent of the personality of the user. On this point, it is believed that even the most ardent Bayesian must agree with orthodox statisticians. The measure of success in producing an objective theory of decision or inference is just the extent to which we are able to eliminate all personalistic elements and create a completely "impersonalistic" theory.

Evidently, then, we need to find a middle ground between the orthodox and personalistic approaches, which will give us just one prior distribution for a given state of knowledge. Historically, orthodox rejection of Bayesian methods was not based at first on any ideological dogma about the "meaning of probability" and certainly not on any failure to recognize the importance of prior information; this has been noted by Kendell and Stuart (1961), Lehmann (1959) and many other orthodox writers. The really fundamental objection (stressed particularly in the remarks of Pearson and Savage 1962) was the lack of any principle by which the prior probabilities could be made objective in the aforementioned sense. Bayesian methods, for all their advantages, will not be entirely satisfactory until we face the problem squarely and show how this requirement may be met.

For later purposes it will be convenient to state this basic desideratum as follows: *in two problems where we have the same prior information, we should assign the same prior probabilities.*

This is stated in such a way that it seems psychologically impossible to quarrel with it; indeed, it may appear so trivial as to be without useful content. A major purpose of the present paper is to show that in many cases, in spite of first appearances, this desideratum may be formulated mathematically in a way which has nontrivial consequences.

Some kinds of prior information seems to vague to be translatable into mathematical terms. If we are told that, "Jones was very pleased at the suggestion that θ might be greater than 100," we have to concede that this does constitute prior information about θ; if we have great respect for Jones' sagacity, it might be relevant for inferences about θ. But how can this be incorporated into a mathematical theory of inference? There is a rather definite minimum requirement which the prior information must satisfy before it can be used by any presently known methods.

Definition 1: A piece of information I concerning a parameter θ will be called *testable* if, given any proposed prior probability assignment $f(\theta)d\theta$, there is a procedure which will determine unambiguously whether $f(\theta)$ does or does not agree with the information I.

As examples, consider the following statements.

I_1: "$\theta < 6$."

I_2: "The mean value of $\tanh^{-1}(1-\theta^2)$ in previous measurements was 1.37."

I_3: "In the eighteenth century, Laplace summarized his analysis of the mass of Saturn by writing, 'It is a bet of 11,000:1 that the error of this estimate is not 1/100 of its value.' He estimated this mass as 1/3512 of the sun's mass."

I_4: "There is at least a 90-percent probability that $\theta > 10$."

Statements I_1 nd I_2 clearly constitute testable information; they can be used immediately to restrict the form of a prior probability assignment. Statement I_3 becomes testable if we understand the exact meaning of Laplace's words, and very easily so if we know the additional historical fact that Laplace's calculations were based on the incomplete beta distribution. I_4 is also clearly testable, but it is perhaps less clear how it could lead to any unique prior probability assignment.

Perhaps in the future others will discover new principles by which nontestable prior information could be used in a mathematical theory of inference. For the present, however, we will restrict ourselves to a search for formal principles by which testable information can be converted into a unique prior probability assignment.

Fortunately, we are not without clues as to how this uniqueness problem might be solved. The principle of maximum entropy (*i.e.,* the prior probability assignment should be the one with the maximum entropy consistent with the prior knowledge) gives a definite rule for setting up priors. The rule is impersonal and has an evident intuitive appeal (Jaynes 1957, 1963, Good 1963, Kullback 1959, Wichmann 1963, and Dutta 1966) as the distribution which "assumes the least" about the unknown parameter. In applications it has a number of advantages, but also some shortcomings which prevent its being regarded as a complete solution to the problem.

We now survey these briefly and aim to supplement the principle in a way that retains the advantages, while correcting the shortcomings.

III. Maximum Entropy

We illustrate this method by a simple example which occurred in a physical problem (distribution of impurities in a crystal lattice), and for simplicity consider only a one-dimensional version. An impurity atom may occupy any of n different positions $\{x_1 \cdots x_n\}$, where $x_j = jL$, and L is a fixed length. From experiments on scattering of X rays, it has been determined that there is a moderate tendency to prefer sites at which $\cos(kx_j) > 0$, the specific datum being that in many previous instances the average value of $\cos(kx_j)$ was

$$\langle \cos(kx_j) \rangle = 0.3. \tag{1}$$

This is clearly testable information, and it is desired to find a probability assignment $p(j|I)$ for occupation of the jth site which incorporates the information I, given by (1), but assumes nothing further, from which statistical predictions about future instances can be made.

The mathematical problem is then to find the $p(j|I)$ which will maximize the entropy

$$H = -\sum_{j=1}^{n} p(j|I) \log p(j|I) \tag{2}$$

subject to the constraints $p(j|I) \geq 0$ and

$$\sum_{j=1}^{n} p(j|I) = 1 \tag{3}$$

$$\sum_{j=1}^{n} p(j|I) \cos(kx_j) = 0.3. \tag{4}$$

The solution is well known, and in this case takes the form

$$p(j|I) = \frac{1}{Z(\lambda)} \exp\{\lambda \cos(kx_j)\} \tag{5}$$

where $Z(\lambda)$ is the partition function

$$Z(\lambda) \equiv \sum_{j=1}^{n} \exp\{\lambda \cos(kx_j)\} \tag{6}$$

and the value of λ is to be determined from (4):

$$\langle \cos(kx) \rangle = \frac{\partial}{\partial \lambda} \log Z(\lambda) = 0.3. \tag{7}$$

In the case where $ka \ll 1$, $nka \gg 1$, we may approximate the discrete sums sufficiently well by integrals, leading to

$$Z(\lambda) \simeq n I_0(\lambda) \tag{8}$$

$$\langle \cos(mkx) \rangle \simeq \frac{I_m(\lambda)}{I_0(\lambda)} \tag{9}$$

where $I_m(\lambda)$ are the modified Bessel functions. From (1), and (9) in the case $m = 1$, we find $\lambda = 0.63$.

Having found the distribution for $p(j|I)$, we can now use it as the prior from which further information about the impurity location can be incorporated via Bayes' theorem. For example, suppose that if the impurity is at site j, the probability that a neutron incident on the crystal will be reflected is proportional to $\sin^2(kx_j)$. We acquire the new data: "n neutrons incident, r reflected." The posterior probability for the impurity to be at site j would then be

$$\begin{aligned} p(j|nr) &= A p(j|I) p(r|nj) \\ &= B \exp\{\lambda \cos(kx_j)\} \left[\sin^2(kx_j)\right]^r \left[\cos^2(kx_j)\right]^{n-r} \end{aligned} \tag{10}$$

where A, B are normalizing constants.

Alternatively, and representative of a large class of important problems which includes statistical mechanics, the prior distribution $p(j|I)$ may be used directly for certain kinds of decision or inference. For example, suppose that before the neutron reflection experiment, we wish to estimate the probability of reflection of r neutrons from n incident. Conditional only on the prior information (1), this probability is

$$\begin{aligned} p(r|n) &= \sum_{j=1}^{n} p(r|nj)p(j|I) \\ &= \binom{n}{r} \left\langle \left[\sin^2(kx_j)\right]^r \left[\cos^2(kx_j)\right]^{n-r} \right\rangle \end{aligned} \tag{11}$$

the expectation value being taken over the prior distribution (5). In the case $n = r = 1$, it reduces to the probability of reflection at a single trial; using (9) we find

$$\langle \sin^2(kx) \rangle = \frac{I_0 - I_2}{2I_0} = \lambda^{-1} \langle \cos(kx) \rangle = 0.48 \tag{12}$$

which is only slightly below the value 0.50 corresponding to a uniform prior distribution $p(j|I)$; thus in agreement with our intuition, the moderate constraint (1) is by no means sufficient to inhibit appreciably the occupation of sites for which $|\sin(kx)| \ll 1$. On the other hand, if the prior information had been $\langle \cos(kx) \rangle = 0.95$, repetition of the argument would yield $\langle \sin^2(kx) \rangle = 0.09$, indicating now a very appreciable inhibition.

The values of $\langle \sin^2(kx) \rangle$ thus calculated represent estimates of $\sin^2(kx)$ which are "optimal" in the sense that 1) they are "maximally noncommittal" with regard to all information except the specific datum given; and 2) they minimize the expected square of the error. Of course, in a problem as rudimentary as this, one does not expect that these estimates can be highly reliable; the information available is far too, meager to permit such a thing. But this fact, too is automatically incorporated into the maximum-entropy formalism; a measure of the reliability of the estimates is given by the expected "loss function," which in this case is just the variance of $\sin^2(kx)$ over the maximum-entropy distribution

$$\sigma^2 = \langle \sin^4(kx) \rangle - \langle \sin^2(kx) \rangle^2 = \frac{I_0^2 - 2I_2^2 + I_0 I_4}{8I_0^2} \tag{13}$$

from which we find, in the cases $\langle \cos(kx) \rangle = 0.3,\ 0.95$, the values $\sigma = 0.35$, $\sigma = 0.12$, respectively. Thus, if $\langle \cos(kx) \rangle = 0.3$, no accurate estimate of $sin^2 kx$ is possible; we can say only that it is reasonably likely to lie in the interval $(0.13, 0.83)$. With the prior datum $\langle \cos(kx) \rangle = 0.95$, we are in a somewhat better position, and can say that $\sin^2(kx)$ is reasonably likely to be less than 0.21.

Evidently the principle of maximum entropy can yield reliable predictions only of those quantities for which it leads to a sharply peaked distribution. If, for example, we find that a maximum-entropy distribution concentrates 99.99 percent of the probability on those values of x for which $6.72 < f(x) < 6.73$, we shall feel justified in predicting that $f(x)$ lies in that interval, and in attributing a very high (but not necessarily 99.99 percent) reliability to our perdition. Mathematically, both equilibrium and non-equilibrium statistical mechanics are equivalent to applying the principle of maximum entropy in just this way; and their success derives from the enormous number of possible microstates, which leads to very sharply peaked distributions (typically of relative width 10^{-12}) for the quantities of interest.

Let us now try to understand some conceptual problems arising from the principle of maximum entropy. A common objection to it is that the probabilities thus obtained have no frequency interpretation, and therefore cannot be relevant to physical applications; there is no reason to believe that distributions observed experimentally would agree with the ones found by maximum entropy. We wish to show that the situation is a great deal more subtle than that by demonstrating that 1) there is a sense in which maximum-entropy distributions do have a precise correspondence with frequencies; 2) in most realistic problems, however, this frequency connection is unnecessary for the usefulness of the principle; and 3) in fact, the principle is most useful in just those cases where the empirical distribution fails to agree with the one predicted by maximum entropy.

IV. The Correspondence Property

Application of the principle of maximum entropy does not require that the distribution sought be the result of any random experiment (in fact, its main purpose was to extend the range of applications of Bayesian methods to problems where the prior probabilities have no reasonable frequency interpretations, such problems being by far the most often encountered in practice). Nonetheless, nothing prevents us from applying it also in cases where the prior distribution is the result of some random experiment, and one would hope that there is some close correspondence between the maximum-entropy distributions and observable frequencies in such cases; indeed, any principle for assigning priors which lacks this correspondence property would surely contain logical inconsistencies.

We give a general proof for the discrete case. The quantity x can take on the values $\{x_1 \cdots x_n\}$ where n may be finite or countably infinite, and the x_i may be specified arbitrarily. The available information about x places a number of constraints on the probability distribution $p(x_i|I)$. We assume for convenience, although it is in no way necessary for our argument, that these take the form of mean values of several functions $\{f_1(x) \cdots f_m(x)\}$, where $m < n$. The probability distribution $p(x_i|I)$ which incorporates this information, but is free from all other assumptions, is then the one which maximizes

$$H = -\sum_{i=1}^{n} p(x_i|I) \log p(x_i|I) \tag{14}$$

subject to the constraints

$$\sum_{i=1}^{n} p(x_i|I) = 1 \tag{15}$$

$$\sum_{i=1}^{n} p(x_i|I) f_k(x_i) = F_k, \quad k = 1, 2, \cdots, m \tag{16}$$

where the F_k are the prescribed mean values. Again, the well-known solution is

$$p(x_i|I) = \frac{1}{Z(\lambda_1 \cdots \lambda_m)} \exp\{\lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)\} \tag{17}$$

with partition function

$$Z(\lambda_1 \cdots \lambda_m) = \sum_{i=1}^{n} \exp\{\lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)\} \tag{18}$$

in which the real constants λ_k are to be determined from the constraints (16), which reduce to the relations

$$F_k = \frac{\partial}{\partial \lambda_k} \log Z(\lambda_1 \cdots \lambda_m). \tag{19}$$

The distribution (17) is the one which is, in a certain sense, spread out as uniformly as possible without contradicting the given information, *i.e.,* it gives free rein to all possible variability of x allowed by the constraints. Thus it accomplishes, in at least one sense, the intuitive purpose of assigning a prior distribution; it agrees with what is known, but expresses a "maximum uncertainty" with respect to all other matters, and thus leaves a maximum possible freedom for our final decisions to be influenced by the subsequent sample data.

Suppose now that the value of x is determined by some random experiment; at each repetition of the experiment the final result is one of the values x_i. On the basis of the given information, what can we say about the frequencies with which the various x_i will occur? Let the experiment be repeated M times (we are particularly interested in the limit $M \to \infty$, because that is the situation referred to in the usual frequency theory of probability), and let every conceivable sequence of results be analyzed. Each trial could give, independently, say one of the results $\{x_i \cdots x_n\}$, and so there are a priori n^M conceivable detailed outcomes. However, many of these will be incompatible with the given information about mean values of the $f_k(x)$. We will, of course, assume that the result of the random experiment agrees with this information (if it did not, then the given information was false and we are doing the wrong problem). In the M repetitions of the experiment, the results x_1 will be obtained m_1 times, x_2 will be obtained m_2 times, etc. Of course,

$$\sum_{i=1}^{n} m_i = M \tag{20}$$

and if the specified mean values are in fact verified, we have the additional relations

$$\sum_{i=1}^{n} m_i f_k(x_i) = MF_k, \quad k = 1, \cdots, m. \tag{21}$$

If $m < n-1$, the constraints (20) and (21) are insufficient to determine the relative frequencies $f_i = m_i/M$. Nevertheless, we have strong grounds for predicting some choices of the f_i to others. For out of the original n^M conceivable results, how many would lead to a given set of sample numbers $\{m_1 \cdots m_n\}$? The answer is, of course, the multinomial coefficient

$$W = \frac{M!}{m_1! \cdots m_n!} = \frac{M!}{(Mf_1)! \cdots (Mf_m)!} \tag{22}$$

and so the set of frequencies $\{f_1 \cdots f_n\}$ which can be realized in the greatest number of ways is the one which maximizes (22) subject to the constraints (20) and (21). We may, equally well, maximize any monotonic increasing function of W, in particular $M^{-1} \log W$, but as $M \to \infty$ we have immediately from the Stirling approximation,

$$M^{-1} \log W \to -\sum_{i=1}^{n} f_i \log f_i = H_f. \tag{23}$$

It is now evident that, in (20)–(23) we have formulated exactly the same mathematical problem as in (14)–(16), and that this identity will persist whether or not the constraints take the form of mean values. Given any testable prior information, the *probability* distribution which maximizes the entropy is numerically identical with the *frequency* distribution which can be realized in the greatest number of ways.

The maximum in W is, furthermore, enormously sharp; to investigate this, let $\{f_i\}$ be the set of frequencies which maximize W and has entropy H_f and $\{f_i'\}$ be any other set of frequencies which agree with the constraints (20) and (21) and has entropy $H'_f < H_f$. The ratio [(number of ways in which $\{f_i\}$ could be realized) / (number of ways in which $\{f'_i\}$ could be realized)] grows asymptotically as

$$\frac{W}{W'} \sim \exp\{M(H_f - H'_f)\} \tag{24}$$

and passes all bounds as $M \to \infty$. Therefore, the distribution predicted by maximum entropy can be realized experimentally in overwhelmingly more ways than can any other. This is the precise connection between maximum-entropy distributions and frequencies promised earlier.

Now, does this property justify a prediction that the maximum-entropy distribution will, in fact, be observed in a real experiment? Clearly not, in the sense of deductive proof, for different people may have different amounts of information, which will lead them to set up different maximum-entropy distributions. Consider a specific case: Mr. A knows the mean value of $\langle f_1(x)\rangle$, $\langle f_2(x)\rangle$; but Mr. B knows in addition $\langle f_3(x)\rangle$. Each sets up a maximum-entropy distribution conditional on his information, and since Mr. B's entropy H_B is maximized subject to one further constraint, we will have

$$H_B \leq H_A. \tag{25}$$

We note two properties, easily verified from the forgoing equations. If Mr. B's additional information is redundant (in the sense that it is only what Mr. A would have predicted from his distribution), then $\lambda_3 = 0$, and the distribution is unchanged. In this case, and only in this case, we have equality in (25). Because of this property (which holds generally), it is never necessary when setting up a maximum-entropy problem to determine whether the different pieces of information used are independent; any redundant information will drop out of the equations automatically.

On the other hand, if the given pieces of information are logically contradictory (for example, if it turns out that $f_3(x) = f_1(x) + 2f_2(x)$, but the given mean values fail to satisfy $\langle f_3(x)\rangle = \langle f_1(x)\rangle + 2\langle f_2(x)\rangle$), then it will be found that (19) has no simultaneous solution with real λ_k. In this case, the method of maximum entropy breaks down, as it should, giving us no distribution at all.

In general, Mr. B's extra information will be neither redundant nor contradictory, and so he will find a maximum-entropy distribution different from that of Mr. A. The inequality will then hold in (25), indicating that Mr. B's extra information was "useful" in further narrowing down the rang of possibilities. Suppose now that we start performing the random experiment with Mr. A and Mr. B watching. Since Mr. A predicts a mean value $\langle f_3(x)\rangle$ different from the correct one known to Mr. B, it is clear that the experimental distribution cannot agree in all respects with Mr. A's prediction. We cannot be sure in advance that it will agree with Mr. B's prediction either, for there may be still further constraints $f_4(x), f_5(x), \cdots$, etc., operative in the experiment but unknown to Mr. B.

However, the property demonstrated above does justify the following weaker statement of frequency correspondence. If the information incorporated into the maximum-entropy analysis includes all the constraints actually operative in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally, because it can be realized in overwhelmingly the greatest number of ways.

Conversely, if the experiment fails to confirm the maximum-entropy prediction, and this disagreement persists on indefinite repetition of the experiment, then we will conclude that the physical mechanism of the experiment must contain additional constraints which were not taken into account in the maximum-entropy calculations. The observed deviations then provide a clue as to the nature of these new constraints. In this way, Mr. A can discover empirically that his information was incomplete.

Now the little scenario just described is an accurate model of just what did happen in one of the most important applications of statistical analysis, carried out by Gibbs. By the year 1900 it was known that in classical statistical mechanics, use of the canonical ensemble (which Gibbs derived as a maximum-entropy distribution over classical phase volume, based on a specified mean value of the energy) failed to predict thermodynamic properties (heat capacities, equations of state, equilibrium constants, etc.) correctly. Analysis of the data showed that the entropy of a real physical system was always less than the value predicted. At that time, therefore, Gibbs was in just the position of Mr. A in the scenario, and the conclusion was drawn that the microscopic laws of physics must involve an additional constraint not contained in the laws of classical mechanics.

In due course, the nature of this constraint was found; first by Plank in the case of radiation, then by Einstein and Debye for solids, and finally by Bohr for isolated atoms. The constraint consisted in the discreteness of the possible energy values, thenceforth called energy levels. By 1927, the mathematical theory by which these could be calculated was developed nearly to its present form.

Thus it is an historical fact that the first clues indicating the need for the quantum theory, and indicating some necessary features of the new theory, were uncovered by a seemingly "unsuccessful" application of the principle of maximum entropy. We may expect that such things will happen again in the future, and this is the basis of the remark that the principle of maximum entropy is most useful to us in just those cases where it fails to predict the correct experimental facts.

Since the history of this development is not well known (a fuller account is given elsewhere, Jaynes 1967), the following brief remarks seem appropriate here. Gibbs (1902) wrote his probability density in phase space in the form

$$w(q_1 \cdots q_n; p_1 \cdots p_n) = \exp\{\eta(q_1 \cdots q_n; p_1 \cdots p_n)\} \tag{26}$$

and called his function η the "index of probability of phase." He derived his canonical and grand canonical ensembles (Gibbs 1902 ch. 11) from constraints on average energy, and average energy and particle numbers, respectively, as (Gibbs 1902, p. 143) "the distribution in phase which without violating this condition gives the least value of the average index of probability of phase $\hat{\eta} \cdots$" This is, of course, just what we would describe today as maximizing the entropy subject to constraints.

Unfortunately, Gibbs did not give any clear explanation, and we can only conjecture whether he possessed one, as to why this particular function is to be minimized on the average, in preference to all others. Consequently, his procedure appeared arbitrary to many, and for sixty years there was controversy over the validity and justification of Gibbs' method. In spite of its enormous practical success when adapted to quantum statistics, few attempts were made to extend it beyond problems of thermal equilibrium.

It was not until the work of Shannon in our own time that the full significance and generality of Gibbs' method could be appreciated. Once we had Shannon's theorem establishing the uniqueness of entropy as an "information measure," it was clear that Gibbs' procedure was an example of a general method for inductive inference, whose applicability is in no way restricted to equilibrium thermodynamics or to physics.

V. Connection with Direct Probability Models

Another important conceptual point is brought out by comparing the frequency correspondence property of maximum-entropy distributions with those obtained from other theoretical models, for example, the standard model of Bernoulli trials. We wish to show that this difference is far less than is often supposed.

As noted previously, we are not entitled to assert, that the distribution predicted by maximum entropy must be observed in a real experiment; we can say only that this distribution is by far the most likely to be observed, provided that the information used includes all the constraints actually operative in the experiment. This requirement, while sufficient, is not always necessary; from the fact that the predicted distribution has been observed, we cannot conclude that no further constraints exist beyond those taken into account. We can conclude only that further constraints, if present, must be of such a nature that they do not affect the relative frequencies (although they might affect other observable things such as correlations).

Now what are we entitled to assert about frequency correspondence of probabilities calculated from the theory of Bernoulli trials? Clearly no probability calculation, whether based on maximum entropy or any other principle, can predict with certainty what the results of a real experiment must be; if the information available were sufficient to permit such a thing, we would have no need of probability theory at all.

In the theory of Bernoulli trials, we calculate the probability that we shall obtain r successes in n trials as

$$p(r|n) = \binom{n}{r} p^r (1-p)^{n-r} \tag{27}$$

in which p is regarded as a given number $0 < p < 1$. For finite n, there is no r in $0 \le r \le n$ which is absolutely excluded by this, and so the observed frequency of success $f \equiv r/n$ cannot be predicted with certainty. Nevertheless, we infer from (27) that, as n becomes very large, the frequency $f = p$ becomes overwhelmingly the most likely to be observed, provided that the assumptions which went into the derivation of (27) (numerical value of p, independence of different trials) correctly describe the conditions operative in the real experiment.

Conversely, if the observed frequency fails to agree with the predictions (and this tendency persists on indefinite repetitions of the experiment), we will conclude that the physical mechanism of the experiment is different from that assumed in the calculation, and the nature of the observed deviation gives a clue as to what is wrong in our assumptions.

On comparing these statements of probability-frequency correspondence, we see that there is virtually no difference in the logical situation between the principles of maximum entropy and of Bernoulli trials. In both cases, and in every other application of probability theory, the onus is on the user to make sure that all the information, which his common sense tells him is relevant to the problem, is actually incorporated into the equations. There is nothing in the mathematical theory which can determine whether this has been, in fact, accomplished; success can be known only a posterior from agreement with experiment. But in both cases, failure to confirm the predictions gives us an opportunity to learn more about the physical mechanism of the experiment.

For these reasons, we are entitled to claim that probabilities calculated by maximum entropy have just as much and just as little correspondence with frequencies as those calculated from any other principle of probability theory.

We can make this point still more strongly by exhibiting a mathematical connection between these two methods of calculation, showing that in many cases we can obtain identical results from use of either method. For this purpose, it is convenient to introduce some more of the vocabulary usually associated with information theory. Any random experiment may be regarded as a "message" transmitted to us by nature. The "alphabet" consists of the set of all possible outcomes of a single trial; on each repetition of the experiment, nature transmits to us one more letter of the message. In the case of Bernoulli trials, we are concerned with a message on a binary alphabet. Define the "random variables"

$$y_i \equiv \left\{ \begin{array}{ll} 1, & \text{if the } i\text{th trial yields success} \\ 0, & \text{if the } i\text{th trial yields failure} \end{array} \right\}. \tag{28}$$

On n repetitions of the experiment, we receive the message

$$M \equiv \{y_1, y_2, \cdots, y_n\} \tag{29}$$

and the total number of successes obtained is

$$r(M) \equiv \sum_{i=1}^{n} y_i. \tag{30}$$

From (27) we find that, for any n_i the expected number of successes is

$$\langle r \rangle = np. \tag{31}$$

Suppose now that we reverse our viewpoint, regard (31) as the primary given datum, and seek the probability of obtaining r successes in n trials by maximum entropy. A full probability analysis of the experiment requires that we consider, not just the probabilities on the 2-point sample space of a single trial, but rather the probabilities

$$P_M \equiv p\{y_0 \cdots y_n\} \tag{32}$$

on the 2^n-point sample space of all possible messages. The problem is then to find the distribution P_M which maximizes the entropy

$$H = -\sum_M P_M \log P_M \tag{33}$$

subject to the constraint (31). The result is

$$P_M = \frac{1}{Z(\lambda)} \exp\{\lambda r(M)\} \tag{34}$$

with the partition function

$$Z(\lambda) = \sum_M \exp\{\lambda r(M)\} = (\exp\{\lambda\} + 1)^n. \tag{35}$$

The value of λ is determined, as always, by (19):

$$\langle r \rangle = \frac{\partial}{\partial \lambda} \log Z = n\left(\exp\{-\lambda\} + 1\right)^{-1}$$

or

$$\lambda = \log \frac{\langle r \rangle}{n - \langle r \rangle} = \log \frac{p}{1-p}. \tag{36}$$

Using (35) and (36), the maximum-entropy distribution (34) reduces to

$$P_M = p^r (1-p)^{n-r}. \tag{37}$$

This is the probability of obtaining a specific message, with successes at specified trials. The probability of obtaining r successes regardless of the order then requires the additional binomial coefficient, and so we obtain precisely the result (27) of the Bernoulli model.

From a mathematical standpoint, therefore, it is immaterial whether we approach the theory of Bernoulli trials in the conventional way, or whether we regard it as an example of maximum-entropy inference on a "higher manifold" than the sample space of a single trial, in which the only information available is the mean value (31).

In a similar way, many other of the so-called "direct probability" calculations may be regarded equally well as the result of applying the principle of maximum entropy on a higher manifold. If we had considered a random experiment with m possible outcomes at a single trial, we would be concerned with messages on the alphabet of m symbols $\{A_1 \cdots A_m\}$, and repetition of the preceding argument leads immediately to the usual multinomial distribution.

We may, perhaps, feel that this result gives us a new insight into the nature of Bernoulli trials. The "independence of different trials" evident already from (34) arises here from the fact that the given information consisted only of statements about individual trials and said nothing about mutual properties of different trials. The principle of maximum entropy thus tells us that, if no information is available concerning correlations between different trials, then we should not assume any such correlations to exist. To do so would reduce the entropy of the distribution P_M and thus reduce the range of variability of different messages below that permitted by the data, *i.e.,* it would amount to introducing new arbitrary assumptions not warranted by the given information. The precise nature of this reduction is described by the asymptotic equipartition theorem (Feinstein 1958). The principle of maximum entropy is just the formal device which ensures that no such hidden arbitrary assumptions have been introduced, and so we are taking into account the full range of possibilities permitted by the information at hand.

If definite information concerning correlations is available, the maximum-entropy method readily digests this information. The usual theory of discrete stochastic processes can be derived by this same application of maximum entropy on a higher manifold, for particular kinds of information about correlations. To give only the simplest example, suppose that in our random experiment with m possible outcomes per trial, we are given information fixing the mean values not only of the "single-letter frequencies" $\langle f_i \rangle$, but also the "digram frequencies" $\langle f_{ij} \rangle$. The maximum-entropy distribution over messages will then take the form

$$P_M = \frac{1}{Z} \exp\left\{ \sum_i \lambda_i f_i(M) + \sum_{ij} \lambda_{ij} f_{ij}(M) \right\} \tag{38}$$

where $nf_i(M)$ is the number of times the letter A_i occurs in the message M, and $(n-1)f_{ij}(M)$ is the number of times the digram $A_i A_j$ occurs in M. The partition function Z is determined by the normalizing of (38). Calculation of the λ_i and the λ_{ij} from (19) is no longer trivial, however, we find the problem to be exactly solvable (Jaynes 1963a). For messages of finite length, there are small "end effects," but in the limit of long messages the maximum-entropy distribution (38) reduces to the distribution of a Markov chain with transition probabilities $p_{ij} = \langle f_{ij} \rangle / \langle f_i \rangle$, in agreement with the results of conventional methods.

In a similar way, if the given information includes expectations of trigram frequencies $\langle f_{ijk} \rangle$, we obtain the distribution of a higher type stochastic process, in which the probability of the outcome A_i at any trial depends on the results of the previous two trials, etc.

To point out the possibility of deriving so much of conventional "direct probability" analysis from maximum entropy on a higher manifold is, of course, in no way to suggest that conventional methods of analysis be abandoned in favor of maximum entropy (although this would bring a higher degree of unity into the field), because in these applications the conventional methods usually lead to shorter calculations. The pragmatic usefulness of maximum entropy lies rather in the fact that it is readily applied in many problems (in particular, setting up prior probability assignments) where conventional methods do no apply.

It is, however, important to realize the possibility of deriving much of conventional probability theory from the principle of maximum entropy, firstly, because it shows that this principle fits in neatly and consistently with the other principles of probability theory. Secondly, we still see from time to time some doubts expressed as to the uniqueness of the expression $(-p \log p)$; it has even been asserted that the results of maximizing this quantity have no more significance than those obtained by maximizing any other convex function. In pointing out the correspondence with frequencies and the fact that many other standard results of probability theory follow from the maximum-entropy principle, we have given a constructive answer to such objections. Any alternative expression to $(-p \log p)$ must surely reproduce all of these desirable properties before it could be taken seriously. It seems to the author impossible that any such alternative quantity could do so, and likely that a rigorous proof of this could now be given.

VI. Continuous Distributions

Thus far we have considered the principle of maximum entropy only for the discrete case and have seen that if the distribution sought can be regarded as produced by a random experiment, there is a correspondence property between probability and frequency, and the results are consistent with other principles of probability theory. However, nothing in the mathematics requires that any random experiment be in fact performed or conceivable; and so we interpret the principle in the broadest sense which gives it the widest range of applicability, *i.e.,* whether or not any random experiment is involved, the maximum-entropy distribution still represents the most "honest" description of our state of knowledge.

In such applications, the principle is easy to apply and leads to the kind of results we should want and expect. For example, in Jaynes (1963a) a sequence of problems of decision making under uncertainty (essentially, of inventory control) of a type which arises constantly in practice was analyzed. Here the state of nature was not the result of any random experiment; there was no sampling distribution and no sample. Thus it might be thought to be a "no data" decision problem, in the sense of Chernoff and Moses (1959). However, in successive stages of the sequence, there were available more and more pieces of prior information, and digesting them by maximum entropy led to a sequence of prior distributions in which the range of possibilities was successively narrowed down. They led to a sequence of decisions, each representing the rational one on the basis of the information available at that stage, which corresponds to intuitive common-sense judgments in the early stages where intuition was able to see the answer. It is difficult to see how this problem could have been treated at all without the use of the principle of maximum entropy, or some other device that turns out in the end to be equivalent to it.

In several years of routine application of this principle in problems of physics and engineering, we have yet to find a case involving a discrete prior where it fails to produce a useful and intuitively reasonable result. To the best of the author's knowledge, no other general method for setting up discrete priors has been proposed. It appears, then, that the principle of maximum entropy may prove to be the final solution to the problem of assigning discrete priors.

Use of this principle in setting up continuous prior distributions, however, requires considerably more analysis because at first glance the results appear to depend on the choice of parameters. We do not refer here to the well-known fact that the quantity

$$H' = -\int p(x) \log p(x) dx \tag{39}$$

lacks invariance under a change of variables $x \to y(x)$, for (39) is not the result of any derivation, and it turns out not to be the correct information measure for a continuous distribution. Shannon's theorem establishing (14) as an information measure goes through only for discrete distributions;

but to find the corresponding expression in the continuous case we can (in the absence of any more direct argument) pass to the limit from a discrete distribution. As shown previously (Jaynes 1963b), this leads instead to the quantity

$$H_c = -\int p(x) \log \frac{p(x)}{m(x)} dx \tag{40}$$

where $m(x)$ is an "invariant measure" function, proportional to the limiting density of discrete points. (In all applications so far studied, $m(x)$ is a well-behaved continuous function, and so we continue to use the notation of Riemann integrals; we call $m(x)$ a "measure" only to suggest the appropriate generalization, readily supplied if a practical problem should ever require it.) Since $p(x)$ and $m(x)$ transform in the same way under a change of variables, H_c is invariant. We examine the form of maximum-entropy inference based on this information measure, in which we may regard x as being either a one-dimensional or multidimensional parameter.

We seek a probability density $p(x)$ which is to be normalized:

$$\int p(x) dx = 1 \tag{41}$$

(we understand the range of integration to be the full parameter space); and we have information fixing the mean values of m different functions $f_k(x)$:

$$F_k = \int p(x) f_k(x) dx, \quad k = 1, 2, \ldots, m \tag{42}$$

where the F_f are the given numerical values. Subject to these constraints, we are to maximize (40). The solution is again elementary:

$$p(x) = Z^{-1} m(x) \exp\left\{\lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)\right\} \tag{43}$$

with the partition function

$$Z(\lambda_1, \cdots, \lambda_m) \equiv \int m(x) \exp\left\{\lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)\right\} dx \tag{44}$$

and the Lagrange multipliers λ_k are determined once again by (19). Our "best" estimate (by quadratic loss function) of any other quantity $q(x)$ is then

$$\langle q \rangle = \int q(x) p(x) dx. \tag{45}$$

It is evident from these equations that when we use (40) rather than (39) as our information measure not only our final conclusions (45), but also the partition function and Lagrange multipliers are all invariant under a change of parameter $x \to y(x)$. In applications, these quantities acquire definite physical meanings.

There remains,however, a practical difficulty. If the parameter space is not the result of any obvious limiting process, what determines the proper measure $m(x)$? The conclusions, evidently, will depend on which measure we adopt. This is the shortcoming from which the maximum-entropy principle, has suffered heretofore, and which must be cleared up before we can regard it as a full solution to the prior probability problem.

Let us note the intuitive meaning of this measure. Consider the one-dimensional case, and suppose it is known that $a < x < b$ but we have no other prior information. Then there are no Lagrange multipliers λ_k and (43) reduces to

$$p(x) = \left[\int_a^b m(x) dx \right]^{-1} m(x), \quad a < x < b. \tag{46}$$

Except for a constant factor, the measure $m(x)$ is also the prior distribution describing "complete ignorance" of x. The ambiguity is, therefore, just the ancient one which has always plagued Bayesian statistics; how do we find the prior representing "complete ignorance?" Once this problem is solved, the maximum-entropy principle will lead to a definite, parameter-independent method of setting up prior distributions based on any testable prior information. Since this problem has been the subject of so much discussion and controversy for 200 years, we wish to state what appears to us a constructive attitude toward it.

To reject the question, as some have done, on the grounds that the state of complete ignorance does not "exist" would be just as absurd as to reject Euclidean geometry on the grounds that a physical point does not exist. In the study of inductive inference, the notion of complete ignorance intrudes itself into the theory just as naturally and inevitably as the concept of zero in arithmetic.

If one rejects the consideration of complete ignorance on the grounds that the notion is vague and ill-defined, the reply is that the notation cannot be evaded in any full theory of inference. So if it is still ill-defined, then a major and immediate objective must be to find a precise definition which will agree with intuitive requirements and be of constructive use in a mathematical theory.

With this in mind, let us survey some previous thoughts on the problem. Bayes suggested, in one particular case, that we express complete ignorance by assigning a uniform prior probability density; and the domain of useful applications of this rule is certainly not zero, for Laplace was led to some of the most important discoveries in celestial mechanics by using it in analysis of astronomical data. However, Bayes' rule has the obvious difficulty that it is not invariant under a change of parameters, and there seems to be no criterion for telling us which parameterization to use. (We note in passing that the notions of an unbiased estimator, and efficient estimator, and a shortest confidence interval are all subject to just the same ambiguity with equally serious consequences, and so orthodox statistics cannot claim to have solved this problem any better than Bayes did.)

Jeffreys (1931 the 1957 edition, and 1939) suggested that we assign a prior $d\sigma/\sigma$ to a continuous parameter σ know to be positive, on the grounds that we are then saying the same thing whether we use the parameter σ or σ^m. Such a desideratum is surely a step in the right direction; however, it cannot be extended to more general parameter changes. We do not want (and obviously cannot have) invariance of the form of the prior under all parameter changes; what we want is invariance of content, but the rules of probability theory already determine how the prior must transform, under any parameter change, so as to achieve this.

The real problem, therefore, must be stated rather differently; we suggest that the proper question to ask is: "For which choice of parameters does a given form such as that of Bayes or Jeffreys apply?" Our parameter spaces seems to have a mollusk-like quality that prevents us from answering this, unless we can find a new principle that gives them a property of "rigidity."

Stated in this way, we recognize that problems of just this type have already appeared and have been solved in other branches of mathematics. In Riemannian geometry and general relativity theory, we allow arbitrary continuous coordinate transformations; yet the property of rigidity is maintained by the concept of the invariant line element, which enables us to make statements of definite geometrical and physical meaning independently of the choice of coordinates. In the theory of continuous groups, the group parameter space has just his mollusk-like quality until the introduction of invariant group measure by Wigner (1959), Harr (1933), and Pontryagin (1912). We seek to do something very similar to this for the parameter spaces of statistics.

The idea of utilizing groups of transformations in problems related to this was discussed by Poincaré (1912) and more recently by Fraser (1966), Hartigan (1964) and Stone (1965). In the following section we give three examples of a different group theoretical method of reasoning developed largely by Weyl and Wigner (1959), which has met with great success in physical problems and seems uniquely adapted to our problem.

VII. Transformation Groups—Examples

The method of reasoning is best illustrated by a simple example, which also happens to be one of the most important in practice. We sample from a continuous two-parameter distribution

$$p(dx|\mu\sigma) = h\left(\frac{x-\mu}{\sigma}\right)\frac{dx}{\sigma} \tag{47}$$

where $h(y)$ is a non-negative and normalized function, and consider the following problem.

Problem 1: Given a sample $\{x_1 \cdots x_n\}$, estimate μ and σ. The problem is indeterminate, both mathematically and conceptually, until we introduce a definite prior distribution

$$f(\mu,\sigma)d\mu d\sigma \tag{48}$$

but if we merely specify "complete initial ignorance," this does not seem to tell us which function $f(\mu,\sigma)$ to use.

Now what do we mean by the statement that we are completely ignorant of μ and σ, except for the knowledge that μ is a location parameter and σ a scale parameter? If we know the sampling distribution (47), we can hardly be ignorant of at least that much. To answer this we might reason as follows. If a change of scale can make the problem appear in any way different to us, then we were not completely ignorant; we must have had some kind of prior knowledge about the absolute scale of the problem. Likewise, if a shift of location can make the problem appear in any way different, then it must be that we had some kind of prior knowledge about location. In other words, complete ignorance of a location and scale parameter is a state of knowledge such that a change of scale and a shift of location does not change that state of knowledge. Suppose, therefore, that we carry out a change of variables $(x,\mu,\sigma) \to (x',\mu',\sigma')$ according to

$$\begin{aligned} \mu' &= \mu + b \\ \sigma' &= a\sigma \\ x' - u' &= a(x-\mu) \end{aligned} \tag{49}$$

where $(0 < a < \infty)$, $(-\infty < b < \infty)$. The distribution (47) expressed in the new variables is unchanged:

$$p(dx'|\mu'\sigma') = h\left(\frac{x'-\mu'}{\sigma'}\right)\frac{dx'}{\sigma'} \tag{50}$$

but the prior distribution is changed to $g(\mu',\sigma')d\mu'd\sigma'$ where from the Jacobian of the transformation (49)

$$g(\mu',\sigma') = a^{-1}f(\mu,\sigma). \tag{51}$$

Now let us consider a second problem.

Problem 2: Given a sample $\{x'_1 \cdots x'_n\}$, estimate μ' and σ'. If we are completely ignorant in the preceding sense, then we must consider Problems 1 and 2 as entirely equivalent for they have identical sampling distributions and our state of prior knowledge about μ' and σ' in Problem 2 is exactly the same as for μ and σ in Problem 1. But our basic desideratum of consistency demands that in two problems where we have the same prior information, we should assign the same prior probabilities. Therefore, f and g must be the same function:

$$f(\mu, \sigma) = g(\mu, \sigma) \tag{52}$$

whatever the values of (a, b). But the form of the prior is now uniquely determined, for combining (49), (51), and (52), we see that $f(\mu, \sigma)$ must satisfy the functional equation

$$f(\mu, \sigma) = af(\mu + b, a\sigma) \tag{53}$$

whose general solution is

$$f(\mu, \sigma) = \frac{(\text{const})}{\sigma} \tag{54}$$

which is the Jeffreys rule.

As another example, not very different mathematically but differently verbalized, consider a Possion process. The probability that exactly n events will occur in a time interval t is

$$p(n|\lambda t) = \exp\left\{-M\frac{(\lambda t)^n}{n!}\right\} \tag{55}$$

and by observing the number of events we wish to estimate the rate constant λ. We are initially completely ignorant of λ except for the knowledge that it is a rate constant of physical dimensions $(\text{seconds})^{-1}$, *i.e.,* we are completely ignorant of the absolute time scale of the process.

Suppose, then, that two observers, Mr. X and Mr. X', whose watches run at different rates so their measurements of a given interval are related by $t = qt'$, conduct this experiment. Since they are observing the same physical experiment, their rate constants must be related by $\lambda' t' = \lambda t$, or $\lambda' = q\lambda$. They assign prior distributions

$$p(d\lambda|X) = f(\lambda)d\lambda \tag{56}$$

$$p(d\lambda'|X') = g(\lambda')d\lambda' \tag{57}$$

and if these are mutually consistent (*i.e.,* they have the same content), it must be that $f(\lambda)d\lambda = g(\lambda')d\lambda'$; or $f(\lambda) = qg(\lambda')$. But Mr. X and Mr. X' are both completely ignorant, and they are in the same state of knowledge, and so f and g must be the same function: $f(\lambda) = g(\lambda)$. Combining those relations gives the functional equation $f(\lambda) = qf(q\lambda)$ or

$$p(d\lambda|X) \sim \lambda^{-1}d\lambda. \tag{58}$$

To use any other prior than this will have the consequence that a change in the time scale will lead to a change in the form of the prior, which would imply a different state of prior knowledge; but if we are completely ignorant of the time scale, then all time scales should appear equivalent.

As a third and less trivial example, where intuition did not anticipate the result, consider Bernoulli trials with an unknown probability of success. Here the probability of success is itself the parameter θ to be estimated. Given θ, the probability that we shall observe r successes in n trials is

$$p(r|n\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r} \tag{59}$$

and again the question is: What prior distribution $f(\theta)d\theta$ describes "complete initial ignorance" of θ?

In discussing this problem, Laplace followed the example of Bayes and answered the question with the famous sentence: "When the probability of a simple event is unknown, we may suppose all values between 0 and 1 as equally likely." In other words, Bayes and Laplace used the uniform prior $f_B(\theta) = 1$. However, Jeffreys (1939) and Carnap (1952) have noted that the resulting rule of succession does not seem to correspond well with the inductive reasoning which we all carry out intuitively. Jeffreys suggested that $f(\theta)$ ought to give greater weight to the end-points $\theta = 0, 1$ if the theory is to account for the kind of inferences made by a scientist.

For example, in a chemical laboratory we find a jar containing an unknown and unlabeled compound. We are at fist completely ignorant as to whether a small sample of this compound will dissolve in water or not. But having observed that one small sample does dissolve, we infer immediately that all samples of this compound are water soluble, and although this conclusion does not carry quite the force of deductive proof, we feel strongly that the inference was justified. Yet the Bayes-Laplace rule leads to a negligible small probability of this being true, and yields only a probability of 2/3 that the next sample tested will dissolve.

Now let us examine this problem from the standpoint of transformation groups. There is a conceptual difficulty here, since $f(\theta)d\theta$ is a "probability of a probability." However, it can be removed by carrying the notation of a split personality to extremes; instead of supposing that $f(\theta)$ describes the state of knowledge of any one person, imagine that we have a large population of individuals who hold varying beliefs about the probability of success, and that $f(\theta)$ describes the distribution of their beliefs. It is possible that, although each individual holds a definite opinion, the population as a whole is completely ignorant of θ? What distribution $f(\theta)$ describes a population in a state of total confusion on the issue?

Since we are concerned with a consistent extension of probability theory, we must suppose that each individual reasons according to the mathematical rules (Bayes' theorem, etc.) of probability theory. The reason they hold different beliefs is, therefore, that they have been given different and conflicting information; one man has read the editorials of the St. Louis Post-Dispatch, another the Los Angeles Times, one has read the Daily Worker, another the National Review, etc., and nothing in probability theory tells one to doubt the truth of what he has been told in the statement of the problem.

Now suppose that, before the experiment is performed, one more definite piece of evidence E is given simultaneously to all of them. Each individual will change his state of belief according to Bayes' theorem; Mr. X, who had previously held the probability of success to be

$$\theta = p(S|X) \tag{60}$$

will change it to

$$\theta' = p(S|EX) = \frac{p(S|X)p(E|SX)}{p(E|SX)p(S|X) + p(E|FX)p(F|X)} \tag{61}$$

where $p(F|X) = 1 - p(S|X)$ is his prior belief in probability of failure. This new evidence thus generates a mapping of the parameter space $0 \le \theta \le 1$ onto itself, given from (61) by

$$\theta' = \frac{a\theta}{1 - \theta + a\theta} \tag{62}$$

where

$$a = \frac{p(E|SX)}{p(E|FX)}. \tag{63}$$

If the population as a whole can learn nothing from this new evidence, then it would seem reasonable to say that the population has been reduced, by conflicting propaganda, to a state of total confusion on the issue. We therefore define the state of "total confusion" or "complete ignorance" by the condition that after the transformation (62), the number of individuals who hold beliefs in any given range $\theta_1 < \theta < \theta_2$ is the same as before.

The mathematical problem is again straightforward. The original distribution of beliefs $f(\theta)$ is shifted by the transformation (62) to a new distribution $g(\theta')$ with

$$f(\theta)d\theta = g(\theta')d\theta' \tag{64}$$

and, if the population as a whole learned nothing, then f and g must be the same function:

$$f(\theta) = g(\theta). \tag{65}$$

Combining (62), (64), and (65), we find that $f(\theta)$ must satisfy the functional equation

$$af\left(\frac{a\theta}{1-\theta-a\theta}\right) = (1-\theta+a\theta)^2 f(\theta). \tag{66}$$

This may be solved directly by eliminating a between (62) and (66) or, in the more usual manner, by differentiating with respect to a and setting $a = 1$. This leads to the differential equation

$$\theta(1-\theta)f'(\theta) = (2\theta - 1)f(\theta) \tag{67}$$

whose solution is

$$f(\theta) = \frac{(\text{const})}{\theta(1-\theta)} \tag{68}$$

which has the qualitative property anticipated by Jeffreys. Now that the imaginary population of individuals has served its purpose of revealing the transformation group (62) of the problem, let them coalesce again into a single mind (that of a statistician who wishes to estimate θ), and let us examine the consequences of using (68) as our prior distribution.

If we had observed r success in n trials, then from (59) and (68) the posterior distribution of θ is (provided that $r \geq 1$, $n - r \geq 1$)

$$p(d\theta|rn) = \frac{(n-1)!}{(r-1)!(n-r-1)!}\theta^{r-1}(1-\theta)^{n-r-1}d\theta. \tag{69}$$

This distribution has expectation value and variance

$$\langle\theta\rangle = \frac{r}{n} = f \tag{70}$$

$$\sigma^2 = \frac{f(1-f)}{n+1}. \tag{71}$$

Thus the "best" estimate of the *probability* of success, by the criterion of quadratic loss function, is just equal to the observed *frequency* of success f; and this is also equal to the probability of success at the next trial, in agreement with the intuition of everybody who has studied Bernoulli trials. On the other hand, the Bayes-Laplace uniform prior would lead instead to the mean value $\langle\theta\rangle_B = (r+1)/(n+2)$ of the rule of succession, which has always seemed a bit peculiar.

For interval estimation, numerical analysis shows that the conclusions drawn from (69) are for all practical purposes the same as those based on confidence intervals [*i.e.,* the shortest 90-percent confidence interval for θ is nearly equal to the shortest 90-percent posterior probability interval determined from (69)]. If $r \gg 1$ and $(n - r) \gg 1$, the normal approximation to (71) will be valid, and the $100P$ percent posterior probability interval is simply $(f \pm q\sigma)$, where q is the $(1 + P)/2$ percentile of the normal distribution; for the 90-, 95-, and 99-percent levels, $q = 1.645$, 1.960, and 2.576, respectively. Under conditions where this normal approximation is valid, the difference between this result and the exact confidence interval are generally less than the difference between various published confidence interval tables, which have been calculated from different approximation schemes.

If $r = (n - r) = 1$, (69) reduces to $p(d\theta|r, n) = d\theta$, the uniform distribution which Bayes and Laplace took as their prior. Therefore, we can now interpret the Bayes-Laplace prior as describing not a state of complete ignorance, but the state of knowledge in which we have observed one success and one failure. It thus appears that the Bayes-Laplace choice will be the appropriate prior if the prior information assures us that it is physically possible for the experiment to yield either a success or a failure, while the distribution of complete ignorance (68) describes a "pre-prior" state of knowledge in which we are not even sure of that.

If $r = 0$, or $r = n$, the derivation of (69) breaks down and the posterior distribution remains unnormalizable, proportional to $\theta^{-1}(1 - \theta)^{n-1}$ or $\theta^{n-1}(1 - \theta)^{-1}$, respectively. The weight is concentrated overwhelmingly on the value $\theta = 0$ or $\theta = 1$. The prior (68) thus accounts for the kind of inductive inference noted in the case of chemicals, which we all make intuitively. However, ounce we have seen at least one success and one failure, then we know that the experiment is a true binary one, in the sense of physical possibility, and from that point on all posterior distributions (69) remain normalized, permitting definite inferences about θ.

The transformation group method therefore yields a prior which appears to meet the common objections raised against the Laplace rule of succession; but we also see that whether (68) or the Bayes-Laplace prior is appropriate depends on the exact prior information available.

To summarize the above results: if we merely specify complete initial ignorance, and cannot hope to obtain any definite prior distribution, because such a statement is too vague to define any mathematically well-posed problem. We are defining what we mean by complete ignorance far more precisely if we can specify as set of operations which we recognize as transforming the problem into an equivalent one, and the desideratum of consistency then places nontrivial restrictions on the form of the prior.

VII. Transformation Groups—Discussion

Further analysis shows that, if the number of independent parameters in the transformation group is equal to the number of parameters in the statistical problem, the "fundamental domain" of the group, Wigner (1959), reduces to a point and the form of the prior is uniquely determined; thus specification of such a transformation group is an exhaustive description of a state of knowledge.

If the number of parameters in the transformation group is less than the number of statistical parameters, the fundamental domain is of higher dimensionality, and the prior will be only partially determined. For example, if in the group (49) we had specified only the change of scale operation and not the shift of location, repetition of the argument would lead to the prior $f(\mu, \sigma) = \sigma^{-1}k(\mu)$, where $k(\mu)$ is an arbitrary function.

It is also readily verified that the transformation group method is consistent with the desideratum of invariance under parameter changes mentioned earlier, *i.e.,* that while the form of the prior cannot be invariant under all parameter changes, its content should be. If the transformation group (49) had been specified in terms of some other choice of parameters (α, β), the form of the transformation equations and the functional equations would, of course, be different, but the prior to which they would lead in the (α, β) space would be just the one that we obtained by solving the problem in the (μ, σ) space and transforming the result to the (α, β) space by the usual Jacobian rule.

The method of reasoning illustrated here is somewhat reminiscent of Laplace's "principle of indifference." However, we are concerned here with indifference between problems, rather than indifference between events. The distinction is essential, for indifference between events is a matter of intuitive judgment on which our intuition often fails even when there is some obvious geometrical symmetry (as Bertrand's paradox shows). However, if a problem is formulated in a sufficiently careful way, indifference between problems is a matter that is determined by the statement of a problem, independently of our intuition; none of the proceeding transformation groups corresponded to any particularly obvious geometrical symmetry.

More generally, if we approach a problem with the charitable presumption that it has a definite solution, then every circumstance left unspecified in the statement of the problem defines an invariance property (*i.e.,* a transformation to an equivalent problem) which that solution must have. Recognition of this leads to a resolution of the Bertrand paradox; here we draw straight lines "at random" intersecting a circle and ask for the distribution of chord lengths. But the statement of the problem does not specify the exact position of the circle; therefore, if there is any definite solution, it must not depend on this circumstance. The condition that the solution be invariant under infinitesimal displacements of the circle relative to the random straight lines uniquely determines the solution.

In such problems, furthermore, the transformation group method is found to have a frequency correspondence property rather like that of the maximum-entropy principle. If (as in the Bertrand problem) the distribution sought can be regarded as the result of a random experiment, then the distribution predicted by invariance under the transformation group is by far the most likely to be observed experimentally, because it requires by far the least "skill," consistently to produce any other would require a "microscopic" degree of control over the exact conditions of the experiment. Proof of the statements in the last two paragraphs will be deferred to later.

The transformation group derivation enables us to see the Jeffreys prior probability rule in a new light. It has, perhaps, always been obvious that the real justification of the Jeffreys rule cannot lie merely in the fact that the parameter is positive. As a simple example, suppose that μ is known to be a location parameter; then both intuition and the preceding analysis agree that a uniform prior density is the proper way to express complete ignorance of μ. The relation $\mu = \theta - \theta^{-1}$ defines a 1:1 mapping of the region $(-\infty < \mu < \infty)$ onto the region $(0 < \theta < \infty)$; but the Jeffreys rule cannot apply to the parameter θ, consistency demanding that its prior density be taken proportional to $d\mu = (1 + \theta^{-2})d\theta$. It appears that the fundamental justification of the Jeffreys rule is not merely that a parameter is positive, but that it is a *scale parameter.*

The fact that the distributions representing complete ignorance found by transformation groups cannot be normalized may be interpreted in two ways. One can say that it arises simply from the fact that our formulation of the notation of complete ignorance was an idealization that does not strictly apply in any realistic problem. A shift of location from a point in St. Louis to a point in the Andromeda nebula, or a change of scale from the size of an atom to the size of our galaxy, does not transform any problem of earthly concern into a completely equivalent one. In practice we will always have some kind of prior knowledge about location and scale, and in consequence the group parameters (a, b) cannot vary over a truly infinite range. Therefore, the transformations (49) do not, strictly speaking, form a group. However, over the range which does express our prior ignorance, the above kind of arguments still apply. Within this range, the functional equations and the resulting form of the priors must still hold.

However, our discussion of maximum entropy shows a more constructive way of looking at this. Finding the distribution representing complete ignorance is only the first step in finding the prior for any realistic problem. The pre-prior distribution resulting from a transformation group does not strictly represent any realistic state of knowledge, but it does define the invariant measure for our parameter space, without which the problem of finding a realistic prior by maximum entropy is mathematically indeterminate.

IX. Conclusion

The analysis given here provides no reason to think that specifying a transformation group is the only way in which complete ignorance may be precisely defined, or that the principle of maximum entropy is the only way to converting testable information into a prior distribution. Furthermore, the procedures described here are not necessarily applicable in all problems, and so it remains an open question whether other approaches may be as good or better. However, before we would be in a position to make any comparative judgments, it would be necessary that some definite alternative procedure be suggested.

At present, lacking this, one can only point out some properties of the methods here suggested. The class of problems in which they can be applied is that in which 1) the prior information is testable; and 2) in the case of a continuous parameter space, the statement of the problem suggests some definite transformation group which establishes the invariant measure. We note that satisfying these conditions is, to a large extent, simply a mater of formulating the problem more completely than is usually done.

If these conditions are met, then we have the means for incorporating prior information into our problem, which is independent of our choice of parameters and is completely impersonal, allowing no arbitrary choice on the part of the user. Few orthodox procedures and, to the best of the author's knowledge, no other Bayesian procedures, enjoy this complete objectivity. Thus while the above criticisms are undoubtedly valid, it seems apparent that this analysis does constitute an advance in the precision with which we are able to formulate statistical problems, as well as an extension of the class of problems in which statistical methods can be used. The fact that this has proved possible gives hope that further work along these lines—in particular, directed toward learning how to formulate problems so that condition 2) is satisfied—may yet lead to the final solution of this ancient but vital puzzle, and thus achieve full objectivity for Bayesian methods.

REFERENCES

Bayes, Rev. Thomas, (1763) "An Essay Toward Solving a Problem in the Doctrine of Chances," Phil. Trans. Roy. Soc. pp. 370-418. Photographic reproduction in E. C. Molina (1963). Reprint, with biographical note by G. A. Barnard in Biometrika **45**, 293-313 (1958) and in Pearson & Kendall (1970). The date is slightly ambiguous; this work was read before the Royal Society in 1763, but not actually published until 1764. Further biographical information on Thomas Bayes (1702-1761) is given by Holland (1962). Stigler (1983) and Zabell (1989) present evidence that Bayes may have found his result and communicated it to friends as early as 1748, in response to a challenge from David Hume (1711-1776).

Carnap, R., (1952) *The Continuum of Inductive Methods,* University of Chicago Press, Chicago.

Chernoff H., and Mosess, L. E., (1959) Elementary Decision Theory, Wiley & Sons, Inc., N. Y.

Dutta, M,. (1966) "On maximum entropy estimation," *Sankhya,* ser. **A,** vol. 28, pt. 4, pp. 319-328.

Fraser, D. A. S., (1966) "On Fiducial Inference," *Ann. Math. Statist., 32,* pp. 661-676.

Gibbs, J. W., (1902) *Elementary Principles in Statistical Mechanics,* Yale University Press, New Haven, Conn.

Good, I. J., (1963) "Maximum entropy for hypothesis formulation," *Ann. Math. Statist.,* **34,** pp. 911-930.

Harr, A. (1933) "Der Massbegriff in der Theorie der Kontinuierlichen Gruppen," *Ann. Math.,* **34,** pp. 147-169.

Hartigan, J., (1964) "Invariant Prior Distributions," *Ann. Math. Statists,,* **35,** pp. 836-845.

Jaynes, E. T., (1957) "Information Theory and Statistical Mechanics I," Phys. Rev. **106**, pp. 620-630; pt. II, *ibid.* **108**, pp. 171-190.

Jaynes, E. T., (1963a) "New Engineering Applications of Information Theory," in *Engineering Uses of Random Function Theory and Probability*, J. L. Bogdanoff and F. Kozin, editors, H. Wiley & Sons, Inc., N. Y., pp. 163-203.

Jaynes, E. T., (1963b) "Information Theory and Statistical Mechanics," in *Statistical Physics*, K. W. Ford, editor, W. A. Benjamin, Inc., pp. 181-218.

Jaynes, E. T., (1967) "Foundations of Probability Theory and Statistical Mechanics," in *Delaware Seminar in Foundations of Physics*, M. Bunge, editor, Springer-Verlag, Berlin. Reprinted in Jaynes (1983).

Jeffreys, H., (1939) *Theory of Probability*, Clarendon Press, Oxford; Later editions, 1948, 1961, 1967, 1988. Appreciated in our Preface.

Jeffreys, H., (1957) *Scientific Inference*, Cambridge University Press, Cambridge.

Kendall, M. G., and Stuart, A., (1961) *The Advanced Theory of Statistics,* Vol 2, C. Griffin and Co., Ltd., London.

Kullback, S., (1959) *Information Theory and Statistics,* Wiley & Sons, Inc., N. Y.

Lehmann, E. L., (1959) *Testing Statistical Hypotheses*, 2nd. edition, 1986, Wiley & Sons, Inc., N. Y.

Pearson, E. S., (1962) Discussion in Savage.

Pontryagin, L,. (1946) *Topological Groups,* Princeton University Press, Princeton, N. J.

Poincaré, H., (1921) *Calcul des Probabilitiès,* Gauthier-Villars, Paris.

Savage, L. J., (1954) *Foundations of Statistics*, Wiley & Sons, Inc., N. Y. Second Revised edition, 1972, by Dover Publications, Inc., New York.

Savage, L. J., (1962) *The Foundations of Statistical Inference: A Discussion*, G. A. Barnard & D. R. Cox, editors, Methuen, London.

Stone, M., (1965) "Right Harr measure for convergence in probability to quasi-posterior distributions," *Ann. Math. Statist.,* **36,** pp. 449-453.

Wald, A., (1950) *Statistical Decision Functions,* Wiley & Sons, Inc., N. Y.

Wichmann, E. H., (1966) "Density matrices arising from incomplete measurements," *J. Math. Phys,.* **4,** pp. 884-897.

Wigner, E. P., (1959) *Group Theory,* Academic Press, New York.