12

Ignorance priors and transformation groups

> Ignorance is preferable to error and he is less remote from the truth who believes nothing than he who believes what is wrong.
>
> *Thomas Jefferson (1781)*

The problem of translating prior information uniquely into a prior probability assignment represents the as yet unfinished half of probability theory, though the principle of maximum entropy in the preceding chapter provides one important tool. It is unfinished because it has been rejected for many decades by those who were unable to conceive of a probability distribution as representing information; but, just because of that long neglect, many current scientific, engineering, economic, and environmental problems are today calling out for new solutions to this problem, without which important new applications cannot proceed.

12.1 What are we trying to do?

It is curious that, even when different workers are in substantially complete agreement on what calculations should be done, they may have radically different views as to what we are actually doing and why we are doing it. For example, there is a large Bayesian community, whose members call themselves 'subjective Bayesians', who have settled into a position intermediate between 'orthodox' statistics and the theory expounded here. Their members have had, for the most part, standard orthodox training; but then they saw the absurdities in it and defected from the orthodox philosophy, while retaining the habits of orthodox terminology and notation.

These habits of expression put subjective Bayesians under a severe handicap. While perceiving that probabilities cannot represent only frequencies, they still regard sampling probabilities as representing frequencies of 'random variables'. But for them prior and posterior probabilities represent only private opinions, which are to be updated, in accordance with de Finetti's principle of coherence. Fortunately, this leads to the Bayesian algorithm, so we do the same calculations.

Subjective Bayesians face an awkward ambiguity at the beginning of a problem, when one assigns prior probabilities. If these represent merely prior opinions, then they are basically arbitrary and undefined; it seems that only private introspection could assign them, and

different people will make different assignments. Yet most subjective Bayesians continue to use a language which implies that there exists some unknown 'true' prior probability distribution in a real problem. In our view, problems of inference are ill-posed until we recognize three essential things.

(A) The prior probabilities represent our prior *information*, and are to be determined, not by introspection, but by *logical analysis* of that information.
(B) Since the final conclusions depend necessarily on both the prior information and the data, it follows that, in formulating a problem, one must specify the prior information to be used just as fully as one specifies the data.
(C) Our goal is that inferences are to be completely 'objective' in the sense that two persons with the same prior information must assign the same prior probabilities.

If one fails to specify the prior information, a problem of inference is just as ill-posed as if one had failed to specify the data. Indeed, since the time of Laplace, applications of probability theory have been hampered by difficulties in the treatment of prior information. In realistic problems of inference, it is typical that we have cogent prior information, highly relevant to the question being asked; to fail to take it into account is to commit the most obvious inconsistency of reasoning, and it may lead to absurd or dangerously misleading results.

Having specified the prior information, we then have the problem of translating that information into a specific prior probability assignment. It is this formal translation process that represents fully half of probability theory, as it is needed for real applications; yet it is entirely absent from orthodox statistics, and only dimly perceived in subjective Bayesian theory.

Just as zero is the natural starting point in adding a column of numbers, the natural starting point in translating a number of pieces of prior information is the state of complete ignorance. In the previous chapter we have seen that for discrete probabilities the principle of maximum entropy tells us, in agreement with our obvious intuition, that complete ignorance, but for specification of a finite set of possibilities, is represented by a uniform prior probability assignment. For continuous probabilities the problem is much more difficult, because intuition fails us and we must resort to formal desiderata and principles. In this chapter we examine the use of the mathematical tool of transformation groups for this purpose.

Some object to the very attempt to represent complete ignorance, on the grounds that a state of complete ignorance does not 'exist'. We would reply that a perfect triangle does not exist either; nevertheless, a surveyor who was ignorant of the properties of perfect triangles would not be competent to do his job. Complete ignorance is, for us, an ideal limiting case of real prior information, in exactly the same sense that a perfect triangle is an ideal limiting case of the real triangles made by surveyors. If we have not learned how to deal with complete ignorance, we are hardly in a position to solve a real problem.

The relatively simple problems examined up till now could be dealt with by reasonable common sense, which could see, nearly always, what the prior ought to be. When we advance to more complicated problems, a formal theory of how to find ignorance priors becomes more and more necessary. The principle of maximum entropy suffices in many cases, but other

principles such as transformation groups, marginalization theory, and coding theory, should also be available in our toolbox. In this chapter we develop the method of transformation groups. Before beginning that development, we first, as a way of introduction, discuss the principle of maximum entropy for continuous distributions, and show how this naturally leads to the idea of assigning distributions to represent complete ignorance.

12.2 Ignorance priors

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 having been 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 about 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.

12.3 Continuous distributions

Use of the principle of maximum entropy 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

$$H' = -\int \mathrm{d}x\; p(x|I)\log[p(x|I)] \tag{12.1}$$

lacks invariance under a change of variables $x \to y(x)$, for (12.1) 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 (11.23) as an information measure goes through only for discrete distributions; to find the corresponding expression in the continuous case we can pass to the limit from a discrete distribution. The following argument can be made as rigorous as we please, but at considerable sacrifice of clarity.

In the discrete entropy expression

$$H_I^{\mathrm{d}} = -\sum_{i=1}^{n} p_i \log[p_i], \tag{12.2}$$

we suppose that the discrete points x_i, $i = 1, 2, \ldots, n$, become more and more numerous, in such a way that, in the limit $n \to \infty$,

$$\lim_{n\to\infty} \frac{1}{n}\,(\text{number of points in } a < x < b) = \int_a^b \mathrm{d}x\, m(x). \tag{12.3}$$

If this passage to the limit is sufficiently well-behaved, it will also be true that adjacent differences $(x_{i+1} - x_i)$ in the neighborhood of any particular value of x will tend to zero so that

$$\lim_{n\to\infty} [n(x_{i+1} - x_i)] = [m(x_i)]^{-1}\,. \tag{12.4}$$

The discrete probability distribution p_i will go over into a continuous probability $p(x|I)$, according to the limiting form of

$$p_i = p(x_i|I)(x_{i+1} - x_i) \tag{12.5}$$

or, from (12.4),

$$p_i \to p(x_i|I)\,[nm(x_i)]^{-1}\,. \tag{12.6}$$

Consequently, the discrete entropy (12.2) goes over into the integral

$$H_I^{\mathrm{d}} \to \int \mathrm{d}x\; p(x|I)\log\left[\frac{p(x|I)}{nm(x)}\right]. \tag{12.7}$$

In the limit, this contains an infinite term $\log(n)$; if we subtract this, the difference will, in the cases of interest, approach a definite limit, which we take as the continuous information measure:

$$H_I^{\mathrm{c}} \equiv \lim_{n\to\infty}\left[H_I^{\mathrm{d}} - \log(n)\right] = -\int \mathrm{d}x\; p(x|I)\log\left[\frac{p(x|I)}{m(x)}\right]. \tag{12.8}$$

The 'invariant measure' function, $m(x)$, is 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 notion 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 $p(x|I)$ and $m(x)$ transform in the same way under a change of variables, H_I^c is invariant.

We seek a probability density $p(x|I)$ which is normalized:

$$\int \mathrm{d}x\; p(x|I) = 1 \tag{12.9}$$

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

$$F_k = \int \mathrm{d}x\; p(x|I) f_k(x), \quad k = 1, 2, \ldots, m, \tag{12.10}$$

where the F_f are the given numerical values. Subject to these constraints, we are to maximize (12.8). The solution is again elementary:

$$p(x|I) = Z^{-1} m(x) \exp\left\{\lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)\right\}, \tag{12.11}$$

with the partition function

$$Z(\lambda_1, \ldots, \lambda_m) \equiv \int \mathrm{d}x\, m(x) \exp\left\{\lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)\right\}, \tag{12.12}$$

and the Lagrange multipliers λ_k are determined by

$$F_k = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k} \qquad k = 1, \ldots, m. \tag{12.13}$$

Our 'best' estimate (by quadratic loss function) of any other quantity $q(x)$ is then

$$\langle q \rangle = \int \mathrm{d}x\, q(x) p(x|I). \tag{12.14}$$

It is evident from these equations that when we use (12.8) rather than (12.1) as our information measure not only our final conclusions (12.14), 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 until now, 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 (12.11) reduces to

$$p(x|I) = \left[\int_a^b \mathrm{d}x\, m(x)\right]^{-1} m(x), \quad a < x < b. \tag{12.15}$$

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 notion 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; 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; 1939, 1957 edn) suggested that we assign a prior $\mathrm{d}\sigma/\sigma$ to a continuous parameter σ known 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 seem 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 this mollusk-like quality until the introduction of invariant group measure by Harr (1933), Pontryagin (1946), and Wigner (1959). 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 Hartigan (1964), Stone (1965) and Fraser (1966). In the following sections we give four examples of a different group theoretical method of reasoning developed largely by Wigner (1959) and Weyl (1961), which has met with great success in physical problems and seems uniquely adapted to our problem.

12.4 Transformation groups

The method of reasoning is best illustrated by some simple examples, the first of which also happens to be one of the most important in practice.

12.4.1 Location and scale parameters

We sample from a continuous two-parameter distribution

$$p(x|\nu\sigma) = \phi(x, \nu, \sigma)\,\mathrm{d}x \tag{12.16}$$

and consider problem A, as follows.

Problem A

Given a sample $\{x_1, \ldots, x_n\}$, estimate ν and σ. The problem is indeterminate, both mathematically and conceptually, until we introduce a definite prior distribution

$$p(\nu\sigma|I)\,\mathrm{d}\nu\,\mathrm{d}\sigma = f(\nu, \sigma)\,\mathrm{d}\nu\,\mathrm{d}\sigma, \tag{12.17}$$

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

Suppose we carry out a change of variables to the new quantities $\{x', \nu', \sigma'\}$ according to

$$\begin{aligned} \nu' &= \nu + b \\ \sigma' &= a\sigma \\ x' - \nu' &= a(x - \nu), \end{aligned} \tag{12.18}$$

where $0 < a < \infty$, $-\infty < b < \infty$. The distribution (12.16) expressed in the new variables is

$$p(x'|\nu'\sigma') = \psi(x', \nu', \sigma') = \phi(x, \nu, \sigma)\,\mathrm{d}x, \tag{12.19}$$

or, from (12.18),

$$\psi(x', \nu', \sigma') = a^{-1}\phi(x, \nu, \sigma). \tag{12.20}$$

Likewise, the prior distribution is changed to $g(\nu', \sigma')$, where, from the Jacobian of the transformation (12.18),

$$g(\nu', \sigma') = a^{-1} f(\nu, \sigma). \tag{12.21}$$

The above relations will hold whatever the distributions $\phi(x, \nu, \sigma)$, $f(\nu, \sigma)$.

Now suppose the distribution (12.16) is invariant under the group of transformations (12.18), so that ψ and ϕ are the same *function:*

$$\psi(x, \nu, \sigma) = \phi(x, \nu, \sigma), \tag{12.22}$$

whatever the values of a, b. The condition for this invariance is that $\phi(x, \nu, \sigma)$ must satisfy the functional equation

$$\phi(x, \nu, \sigma) = a\phi(ax - a\nu + \nu + b, \nu + b, a\sigma). \tag{12.23}$$

Differentiating with respect to a, b and solving the resulting differential equation, we find that the general solution of (12.23) is

$$\phi(x, \nu, \sigma) = \frac{1}{\sigma} h\left(\frac{x - \nu}{\sigma}\right), \tag{12.24}$$

where $h(q)$ is an arbitrary function. Thus, the usual definition of a location parameter ν and a scale parameter σ is equivalent to specifying that the distribution shall be invariant under the group of transformations (12.18).

What do we mean by the statement that we are 'completely ignorant' of ν and σ except for the knowledge that ν is a location parameter and σ is a scale parameter? 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 information about the absolute scale of the problem. Likewise, if a shift of location can make the problem appear in any way different, then we must have had some prior information about location. In other words, 'complete ignorance' of a location and a scale parameter is a state of knowledge such that *a change of scale and shift of location does not change that state of knowledge.* We shall presently have to state this more carefully, but first let us see its consequences. Consider, therefore, problem B.

Problem B

Given a sample $\{x'_1, \ldots, x'_n\}$, estimate ν' and σ'. If we are 'completely ignorant' in the above sense, then we must consider A and B as entirely equivalent problems; they have identical sampling distributions, and our state of prior knowledge about ν' and σ' in problem B is exactly the same as for ν and σ in problem A.

Our basic desideratum now acquires a nontrivial content; for we have formulated two problems in which we have the same prior information. Consistency demands, therefore,

that we assign the same prior probability distribution in them. Thus, f and g must be the same function:

$$f(\nu, \sigma) = g(\nu, \sigma) \tag{12.25}$$

whatever the values of (a, b). But the form of the prior distribution is now uniquely determined; for, combining (12.18), (12.21), and (12.25), we see that $f(\nu, \sigma)$ must satisfy the functional equation

$$f(\nu, \sigma) = af(\nu + b, a\sigma), \tag{12.26}$$

whose general solution is

$$f(\nu, \sigma) = \frac{\text{const.}}{\sigma} \tag{12.27}$$

which is the Jeffreys rule!

We must not jump to the conclusion that the prior (12.27) has been determined by the form (12.24) of the population. Indeed, it would be very disconcerting if the form of the prior were determined merely by the form of the population from which we are sampling; any principle which led to such a result would be suspect. Examination of the above reasoning shows, however, that the result (12.27) was uniquely determined by the *transformation group* (12.18), and not by the form of the distribution (12.24).

To illustrate this, note that there is more than one transformation group under which (12.24) is invariant. In the transformations (12.18) we carry out a change of scale by a factor a and a translation b. Denoting this operation by the symbol (a, b), we can carry out the transformation (a_1, b_1), then (a_2, b_2), and, from (12.18), obtain the composition law of group elements:

$$(a_2, b_2)(a_1, b_1) = (a_2 a_1, b_2 + b_1). \tag{12.28}$$

Thus the group (12.18) is Abelian, the direct product of two one-parameter groups. It has a faithful representation in terms of the matrices

$$\begin{pmatrix} a & 0 \\ 0 & \exp\{b\} \end{pmatrix}. \tag{12.29}$$

Now consider the group of transformations in which we first carry out a change of scale a on the quantities, and follow this by a translation b. This group is given by

$$\begin{aligned} \nu' &= a\nu + b \\ \sigma' &= a\sigma \\ x' &= ax + b. \end{aligned} \tag{12.30}$$

These transformations have the composition law

$$(a_2, b_2)(a_1, b_2) = (a_2 a_1, a_2 b_1 + b_2), \tag{12.31}$$

and so the group (12.30) is non-Abelian; it has a faithful representation in terms of the matrices

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}, \tag{12.32}$$

which cannot be reduced to diagonal form. Therefore, (12.18) and (12.30) are entirely different groups.

If we specify the transformation group (12.30) instead of (12.18), (12.21) and (12.23) are modified to

$$g(v', \sigma') = a^{-2} f(v, \sigma), \tag{12.33}$$

and

$$\phi(x, v, \sigma) = a\phi(ax + b, av + b, a\sigma). \tag{12.34}$$

But we find that the general solution of (12.34) is also (12.24); and so both groups define location and scale parameters equally well. However, their consequences for the prior are different; for the functional equation (12.26) is modified to

$$f(v, \sigma) = a^2 f(av + b, a\sigma), \tag{12.35}$$

whose general solution is

$$f(v, \sigma) = \frac{\text{const.}}{\sigma^2}. \tag{12.36}$$

Thus, the state of knowledge which is invariant under the group (12.18) is *not* the same as that which is invariant under (12.30); and we see a new subtlety in the concept of 'complete ignorance'. In order to define it unambiguously, it is not enough to say merely, 'A change of scale and shift of location does not change that state of knowledge'. We must specify the precise manner in which these operations are to be carried out; i.e. *we must specify a definite group of transformations.*

We thus face the question: Which group, (12.18) or (12.30), really describes the prior information? The difficulty with (12.30) lies in the equations $x' = ax + b$, $v' = ax + b$; thus, the change of scale operation is to be carried out about two points denoted by $x = 0$, $v = 0$. But, if we are 'completely ignorant' about location, then the condition $x = 0$ has no particular meaning; what determines this fixed point about which the change of scale is to be carried out?

In every problem which I have been able to imagine, it is the group (12.18), and therefore the Jeffreys prior probability rule, which seems appropriate. Here the change of scale involves only the difference $\{x - v\}$; thus it is carried out about a point which is itself arbitrary, and so no 'fixed point' is defined by the group (12.18). However, it will be interesting to see whether others can produce examples in which the point $x = 0$ always has a special meaning, justifying the stronger prior (12.36).

To summarize: if we merely specify 'complete initial ignorance', we 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 this state of knowledge far more precisely if we can specify a set of operations which we recognize as transforming the problem into an equivalent one. Having found such a set of operations, the basic desideratum of consistency then places nontrivial restrictions on the form of the prior.

12.4.2 A Poisson rate

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

$$p(n|\lambda t) = \exp\left\{-\frac{(\lambda t)^n}{n!}\right\}, \tag{12.37}$$

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 such that 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

$$p(\mathrm{d}\lambda|X) = f(\lambda)\,\mathrm{d}\lambda, \tag{12.38}$$

$$p(\mathrm{d}\lambda'|X') = g(\lambda')\,\mathrm{d}\lambda', \tag{12.39}$$

and if these are mutually consistent (i.e. they have the same content), it must be that $f(\lambda)\mathrm{d}\lambda = g(\lambda')\mathrm{d}\lambda'$; 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

$$p(\mathrm{d}\lambda|X) \sim \lambda^{-1}\mathrm{d}\lambda. \tag{12.40}$$

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.

12.4.3 Unknown probability for success

As a third and less trivial example, where intuition did not anticipate the result, consider Bernoulli trials with an unknown probability for success. Here the probability for success

is itself the parameter θ to be estimated. Given θ, the probability that we shall observe r successes in n trials is

$$p(r|n\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}, \tag{12.41}$$

and again the question is: What prior distribution $f(\theta)\,\mathrm{d}\theta$ 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 for a simple event is unknown, we may suppose all values between zero and one 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 first 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 negligibly small probability for 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 $f(\theta)\,\mathrm{d}\theta$ is a 'probability for a probability'. However, it can be removed by carrying the notion 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 for success, and that $f(\theta)$ describes the distribution of their beliefs. Is it 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 for success to be

$$\theta = p(S|X), \tag{12.42}$$

will change it to

$$\theta' = p(S|EX) = \frac{p(S|X)p(E|SX)}{p(E|SX)p(S|X) + p(E|FX)p(F|X)}, \tag{12.43}$$

where $p(F|X) = 1 - p(S|X)$ is his prior belief in probability for failure. This new evidence thus generates a mapping of the parameter space $0 \leq \theta \leq 1$ onto itself, given from (12.43) by

$$\theta' = \frac{a\theta}{1 - \theta + a\theta}, \tag{12.44}$$

where

$$a = \frac{p(E|SX)}{p(E|FX)}. \tag{12.45}$$

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 (12.44), 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 for beliefs $f(\theta)$ is shifted by the transformation (12.44) to a new distribution $g(\theta')$ with

$$f(\theta)\,\mathrm{d}\theta = g(\theta')\,\mathrm{d}\theta', \tag{12.46}$$

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

$$f(\theta) = g(\theta). \tag{12.47}$$

Combining (12.44), (12.46), and (12.47), we find that $f(\theta)$ must satisfy the functional equation

$$af\left(\frac{a\theta}{1 - \theta - a\theta}\right) = (1 - \theta + a\theta)^2 f(\theta). \tag{12.48}$$

This may be solved directly by eliminating a between (12.44) and (12.48) or, in the more usual manner, by differentiating with respect to a and setting $a = 1$. This leads to the differential equation

$$\theta(1 - \theta)f'(\theta) = (2\theta - 1)f(\theta), \tag{12.49}$$

whose solution is

$$f(\theta) = \frac{\text{const.}}{\theta(1 - \theta)}, \tag{12.50}$$

which has the qualitative property anticipated by Jeffreys. Now that the imaginary population of individuals has served its purpose of revealing the transformation group (12.44) 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 (12.50) as our prior distribution.

If we had observed r successes in n trials, then from (12.41) and (12.50) the posterior distribution for θ is (provided that $r \geq 1, n - r \geq 1$)

$$p(\mathrm{d}\theta|rn) = \frac{(n-1)!}{(r-1)!(n-r-1)!}\theta^{r-1}(1-\theta)^{n-r-1}\mathrm{d}\theta. \tag{12.51}$$

This distribution has expectation value and variance

$$\langle\theta\rangle = \frac{r}{n} = f, \tag{12.52}$$

$$\sigma^2 = \frac{f(1-f)}{n+1}. \tag{12.53}$$

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 for 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 (12.51) are, for all practical purposes, the same as those based on confidence intervals (i.e. the shortest 90% confidence interval for θ is nearly equal to the shortest 90% posterior probability interval determined from (12.51)). If $r \gg 1$ and $(n-r) \gg 1$, the normal approximation to (12.51) will be valid, and the $100P\%$ 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% 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 is generally less than the difference between various published confidence interval tables, which have been calculated from different approximation schemes.

If $r = (n-r) = 1$, (12.51) reduces to $p(\mathrm{d}\theta|r,n) = \mathrm{d}\theta$, 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 for complete ignorance (12.50) 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 (12.51) 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 (12.50) thus accounts for the kind of inductive inference noted in the case of chemicals, which we all make intuitively. However, once 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 (12.51) 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 (12.50) or the Bayes–Laplace prior is appropriate depends on the exact prior information available.

12.4.4 Bertrand's problem

Finally, we give an example where transformation groups may be used to find more informative priors. Bertrand's problem (Bertrand, 1889) was stated originally in terms of drawing a straight line 'at random' intersecting a circle. It will be helpful to think of this in a more concrete way; presumably, we do no violence to the problem (i.e. it is still just as 'random') if we suppose that we are tossing straws onto the circle, without specifying how they are tossed. We therefore formulate the problem as follows.

A long straw is tossed at random onto a circle; given that it falls so that it intersects the circle, what is the probability that the chord thus defined is longer than a side of the inscribed equilateral triangle? Since Bertrand proposed it in 1889, this problem has been cited to generations of students to demonstrate that Laplace's 'principle of indifference' contains logical inconsistencies. For there appear to be many ways of defining 'equally possible' situations, and they lead to different results. Three of these are: assign uniform probability density to (A) the linear distance between centers of chord and circle, (B) angles of intersections of the chord on the circumference, (C) the center of the chord over the interior area of the circle. These assignments lead to the results $p_A = 1/2$, $p_B = 1/3$, and $p_C = 1/4$, respectively.

Which solution is correct? Of the ten authors cited (Bertrand 1889; Borel 1909; Poincaré 1912; Uspensky 1937; Northrop 1944; von Mises 1957; Gnedenko 1962; Kendell and Moran 1963; Mosteller 1965), only Borel is willing to express a definite preference, although he does not support it by any proof. Von Mises takes the opposite extreme, declaring that such problems (including the similar Buffon needle problem) do not belong to the field of probability theory at all. The others, including Bertrand, take the intermediate position of saying simply that the problem has no definite solution because it is ill-posed, the phrase 'at random' being undefined.

In works on probability theory, this state of affairs has been interpreted, almost universally, as showing that the principle of indifference must be totally rejected. Usually, there is the further conclusion that the only valid basis for assigning probabilities is frequency in some random experiment. It would appear, then, that the only way of answering Bertrand's question is to perform the experiment.

But do we really believe that it is beyond our power to predict by 'pure thought' the result of such a simple experiment? The point at issue is far more important than merely resolving a geometric puzzle; for, as discussed further in the conclusion of this chapter, applications of probability theory to physical experiments usually lead to problems of just this type; i.e. they appear at first to be undetermined, allowing many different solutions with nothing to choose among them. For example, given the average particle density and total energy of a gas, predict its viscosity. The answer, evidently, depends on the exact spatial and velocity distributions of the molecules (in fact, it depends critically on position–velocity correlations), and nothing in the given data seems to tell us which distribution to assume. Yet physicists *have* made definite choices, guided by the principle of indifference, and they *have* led us to correct and nontrivial predictions of viscosity and many other physical phenomena.

Thus, while in some problems the principle of indifference has led us to paradoxes, in others it has produced some of the most important and successful applications of probability theory. To reject the principle without having anything better to put in its place would lead to consequences so unacceptable that for many years even those who profess the most faithful adherence to the strict frequency definition of probability have managed to overlook these logical difficulties in order to preserve some very useful solutions.

Evidently, we ought to examine the apparent paradoxes such as Bertrand's more closely; there is an important point to be learned about the application of probability theory to real physical situations.

It is evident that if the circle becomes sufficiently large, and the tosser sufficiently skilled, various results could be obtained at will. However, in the limit where the skill of the tosser must be described by a 'region of uncertainty' large compared with the circle, the distribution for chord lengths must surely go into one unique function obtainable by 'pure thought'. A viewpoint toward probability theory which cannot show us how to calculate this function from first principles, or even denies the possibility of doing this, would imply severe – and, to a physicist, intolerable – restrictions on the range of useful applications of probability theory.

An invariance argument was applied to problems of this type by Poincaré (1912), and cited more recently by Kendall and Moran (1963). In this treatment we consider straight lines drawn 'at random' in the xy plane. Each line is located by specifying two parameters (u, v) such that the equation of the line is $ux + vy = 1$, and one can ask: Which probability density $p(u, v)\,\mathrm{d}u\,\mathrm{d}v$ has the property that it is invariant in *form* under the group of Euclidean transformations (rotations and translations) of the plane? This is a readily solvable problem (Kendall and Moran 1963), with the answer $p(u, v) = (u^2 + v^2)^{-3/2}$.

Yet evidently this has not seemed convincing; for later authors have ignored Poincaré's invariance argument, and have adhered to Bertrand's original judgment that the problem has no definite solution. This is understandable, for the statement of the problem does not specify that the distribution for straight lines is to have this invariance property, and we do not see any compelling reason to expect that a rain of straws produced in a real experiment would have it. To assume this would seem to be an intuitive judgment resting on no stronger

grounds than the ones which led to the three different solutions above. All of this amounts to trying to guess what properties a 'random' rain of straws should have, by specifying the intuitively 'equally possible' events; and the fact remains that different intuitive judgments lead to different results.

The viewpoint just expressed, which is by far the most common in the literature, clearly represents one valid way of interpreting the problem. If we can find another viewpoint according to which such problems *do* have definite solutions, *and define the conditions under which these solutions are experimentally verifiable,* then, while it would perhaps be overstating the case to say that this new viewpoint is more 'correct' in principle than the conventional one, it will surely be more useful in practice.

We now suggest such a viewpoint, and we understand from the start that we are not concerned at this stage with *frequencies* of various events. We ask rather: Which probability distribution describes our *state of knowledge* when the only information available is that given in the above statement of the problem? Such a distribution must conform to the desideratum of consistency formulated in Chapter 1: in two problems where we have the same state of knowledge we must assign the same probabilities. The essential point is this: if we start with the assumption that Bertrand's problem has a definite solution *in spite of the many things left unspecified,* then the statement of the problem automatically implies certain invariance properties, which in no way depend on our intuitive judgments. After the solution is found, it may be used as a prior for Bayesian inference whether or not it has any correspondence with frequencies; any frequency connections that may emerge will be regarded as an additional bonus, which justify its use also for direct physical prediction.

Bertrand's problem has an obvious element of rotational symmetry, recognized in all the proposed solutions; however, this symmetry is irrelevant to the distribution for chord lengths. There are two other 'symmetries' which are highly relevant: neither Bertrand's original statement nor our restatement in terms of straws specified the exact size of the circle, or its exact location. If, therefore, the problem is to have any definite solution at all, it must be 'indifferent' to these circumstances; i.e. it must be unchanged by a small change in the size or position of the circle. This seemingly trivial statement, as we will see, fully determines the solution.

It would be possible to consider all these invariance requirements simultaneously by defining a four-parameter transformation group, whereupon the complete solution would appear suddenly, as if by magic. However, it will be more instructive to analyze the effects of these invariances separately, and see how each places its own restrictions on the form of the solution.

Rotational invariance

Let the circle have radius R. The position of the chord is determined by giving the polar coordinates (r, θ) of its center. We seek to answer a more detailed question than Bertrand's: What probability density $f(r, \theta)\mathrm{d}A = f(r, \theta)\, r\, \mathrm{d}r\, \mathrm{d}\theta$ should we assign over the interior

area of the circle? The dependence on θ is actually irrelevant to Bertrand's question, since the distribution for chord lengths depends only on the radial distribution

$$g(r) = \int_0^{2\pi} \mathrm{d}\theta\, f(r, \theta). \tag{12.54}$$

However, intuition suggests that $f(r, \theta)$ should be independent of θ, and the formal transformation group argument deals with the rotational symmetry as follows.

The starting point is the observation that the statement of the problem does not specify whether the observer is facing north or east; therefore, if there is a definite solution, it must not depend on the direction of the observer's line of sight. Suppose, therefore, that two different observers, Mr X and Mr Y, are watching this experiment. They view the experiment from different directions, their lines of sight making an angle α. Each uses a coordinate system oriented along his line of sight. Mr X assigns the probability density $f(r, \theta)$ in his coordinate system S; and Mr Y assigns $g(r, \theta)$ in his system S_α. Evidently, if they are describing the same situation, then it must be true that

$$f(r, \theta) = g(r, \theta - \alpha), \tag{12.55}$$

which expresses a simple change of variables, transforming a fixed distribution f to a new coordinate system; this relation will hold whether or not the problem has rotational symmetry.

But now we recognize that, because of the rotational symmetry, the problem appears exactly the same to Mr X in his coordinate system as it does to Mr Y in his. Since they are in the same state of knowledge, our desideratum of consistency demands that they assign the same probability distribution; and so f and g must be the same function:

$$f(r, \theta) = g(r, \theta). \tag{12.56}$$

These relations must hold for all α in $0 \leq \alpha \leq 2\pi$; and so the only possibility is $f(r, \theta) = f(r)$.

This formal argument may appear cumbersome when compared with our obvious flash of intuition; and of course it is, when applied to such a trivial problem. However, as Wigner (1931) and Weyl (1946) have shown in other physical problems, it is this cumbersome argument that generalizes at once to nontrivial cases where our intuition fails us. It always consists of two steps: we first find a transformation equation like (12.55) which shows how two problems are related to each other, irrespective of symmetry; then a symmetry relation like (12.56) which states that we have formulated two equivalent *problems*. Combining them leads in most cases to a functional equation which imposes some restriction on the form of the distribution.

Scale invariance

The problem is reduced, by rotational symmetry, to determining a function $f(r)$, normalized according to

$$\int_0^{2\pi} \mathrm{d}\theta \int_0^R r\mathrm{d}r\, f(r) = 1. \tag{12.57}$$

Again, we consider two different problems; concentric with a circle of radius R, there is a circle of radius aR, $0 < a \leq 1$. Within the smaller circle there is a probability $h(r)\, r\, \mathrm{d}r\, \mathrm{d}\theta$ which answers the question: given that a straw intersects the smaller circle, what is the probability that the center of its chord lies in the area $\mathrm{d}A = r\, \mathrm{d}r\, \mathrm{d}\theta$?

Any straw that intersects the small circle will also define a chord on the larger one; and so, within the small circle $f(r)$ must be proportional to $h(r)$. This proportionality is, of course, given by the standard formula for a conditional probability, which in this case takes the form

$$f(r) = 2\pi h(r) \int_0^{aR} r\mathrm{d}r\, f(r) \qquad 0 < a \leq 1, \quad 0 \leq r \leq aR. \tag{12.58}$$

This transformation equation will hold whether or not the problem has scale invariance.

But we now invoke scale invariance; to two different observers with different size eyeballs, the problems of the large and small circles would appear exactly the same. If there is any unique solution independent of the size of the circle, there must be another relationship between $f(r)$ and $h(r)$, which expresses the fact that one problem is merely a scaled-down version of the other. Two elements of area $r\, \mathrm{d}r\, \mathrm{d}\theta$ and $(ar)\, \mathrm{d}(ar)\, \mathrm{d}\theta$ are related to the large and small circles, respectively, in the same way; and so they must be assigned the same probabilities by the distributions $f(r)$ and $h(r)$, respectively:

$$h(ar)\,(ar)\,\mathrm{d}(ar)\,\mathrm{d}\theta = f(r)\, r\, \mathrm{d}r\, \mathrm{d}\theta, \tag{12.59}$$

or

$$a^2 h(ar) = f(r), \tag{12.60}$$

which is the symmetry equation. Combining (12.58) and (12.60), we see that invariance under change of scale requires that the probability density satisfy the functional equation

$$a^2 f(ar) = 2\pi f(r) \int_0^{aR} u\mathrm{d}u\, f(u) \qquad 0 < a \leq 1, \quad 0 \leq r \leq R. \tag{12.61}$$

Differentiating with respect to a, setting $a = 1$, and solving the resulting differential equation, we find that the most general solution of (12.61) satisfying the normalization condition (12.57) is

$$f(r) = \frac{q r^{q-2}}{2\pi R^q}, \tag{12.62}$$

where q is a constant in the range $0 < q < \infty$, not further determined by scale invariance.

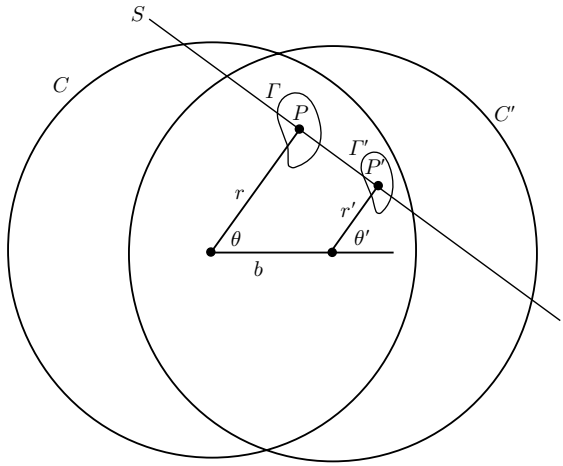

Fig. 12.1. A Straw S intersects two slightly displaced circles C and C'.

We note that the proposed solution B in the introduction has now been eliminated, for it corresponds to the choice $f(r) \sim 1/\sqrt{(R^2 - r^2)}$, which is not of the form (12.62). This means that if the intersections of chords on the circumference were distributed in angle uniformly and independently on one circle, this would not be true for a smaller circle inscribed in it; i.e. the probability assignment of B could be true for, at most, only one size of circle. However, solutions A and C are still compatible with scale invariance, corresponding to the choices $q = 1$ and $q = 2$, respectively.

Translational invariance

We now investigate the consequences of the fact that a given straw S can intersect two circles C, C' of the same radius R, but with a relative displacement b. Referring to Figure 12.1, the midpoint of the chord with respect to circle C is the point P, with coordinates (r, θ); while the same straw defines a midpoint of the chord with respect to C' at the point P' whose coordinates are (r', θ'). From Figure 12.1 the coordinate transformation $(r, \theta) \rightarrow (r', \theta')$ is given by

$$r' = |r - b\cos\theta|, \tag{12.63}$$

$$\theta' = \begin{cases} \theta & r > b\cos\theta \\ \theta + \pi & r < b\cos\theta. \end{cases} \tag{12.64}$$

As P varies over the region Γ, P' varies over Γ', and vice versa; thus the straws define a 1:1 mapping of Γ onto Γ'.

Now we note the translational symmetry; since the statement of the problem gave no information about the location of the circle, the problems of C and C' appear exactly the same to two slightly displaced observers O and O'. Our desideratum of consistency then demands that they assign probability densities in C and C', respectively, which have the same form (12.62) with the same value of q.

It is further necessary that these two observers assign equal probabilities to the regions Γ and Γ', respectively, since (a) they are probabilities of the same event, and (b) the probability that a straw which intersects one circle will also intersect the other, thus setting up this correspondence, is also the same in the two problems. Let us see whether these two requirements are compatible.

The probability that a chord intersecting C will have its midpoint in Γ is

$$\int_\Gamma r\mathrm{d}r\mathrm{d}\theta\, f(r) = \frac{q}{2\pi R^q}\int_\Gamma \mathrm{d}r\mathrm{d}\theta\, r^{q-1}. \tag{12.65}$$

The probability that a chord intersecting C' will have its midpoint in Γ' is

$$\frac{q}{2\pi R^q}\int_{\Gamma'} \mathrm{d}r'\mathrm{d}\theta'\,(r')^{q-1} = \frac{q}{2\pi R^q}\int_\Gamma \mathrm{d}r\mathrm{d}\theta\;|r - b\cos\theta|^{q-1}\,, \tag{12.66}$$

where we have transformed the integral back to the variables (r, θ) by use of (12.63) and (12.64), noting that the Jacobian is unity. Evidently, (12.65) and (12.66) will be equal for arbitrary Γ if and only if $q = 1$; and so our distribution $f(r)$ is now uniquely determined.

The proposed solution C in the introduction is thus eliminated for lack of translational invariance; a rain of straws which had the property assumed with respect to one circle could not have the same property with respect to a slightly displaced one.

We have found that the invariance requirements determine the probability density

$$f(r,\theta) = \frac{1}{2\pi Rr}, \quad 0 \le r \le R, \quad 0 \le \theta \le 2\pi, \tag{12.67}$$

corresponding to solution A in the introduction. It is interesting that this has a singularity at the center, the need for which can be understood as follows. The condition that the midpoint (r, θ) falls within a small region Δ imposes restrictions on the possible directions of the chord. But as Δ moves inward, as soon as it includes the center of the circle all angles are suddenly allowed. Thus there is an infinitely rapid change in the 'manifold of possibilities'.

Further analysis (almost obvious from contemplation of Figure 12.1) shows that the requirement of translational invariance is so stringent that it already determines the result (12.67) uniquely; thus the proposed solution B is incompatible with either scale or translational invariance, and in order to find (12.67) it was not really necessary to consider scale invariance. However, the solution (12.67) would in any event have to be tested for scale invariance, and if it failed to pass that test we would conclude that the problem as stated has *no* solution; i.e. although at first glance it appears underdetermined, it would have to be regarded, from the standpoint of transformation groups, as overdetermined. As luck would have it, these requirements *are* compatible; and so the problem has one unique solution.

The distribution for chord lengths follows at once from (12.67). A chord whose midpoint is at (r, θ) has a length $L = 2\sqrt{(R^2 - r^2)}$. In terms of the reduced chord lengths, $x \equiv L/2R$,

we obtain the universal distribution law

$$p(x)\,\mathrm{d}x = \frac{x\,\mathrm{d}x}{\sqrt{(1-x^2)}}, \quad 0 \le x \le 1, \tag{12.68}$$

in agreement with Borel's conjecture (1909).

Frequency correspondence

From the manner of its derivation, the distribution (12.68) would appear to have only a subjective meaning; while it describes the only possible state of knowledge corresponding to a unique solution in view of the many things left unspecified in the statement of Bertrand's problem, we have as yet given no reason to suppose that it has any relation to frequencies observed in the actual experiment. In general, of course, no such claim can be made; the mere fact that my state of knowledge gives me no reason to prefer one event over another is not enough to make the events occur equally often! Indeed, it is clear that no 'pure thought' argument, whether based on transformation groups or any other principle, can predict with certainty what must happen in a real experiment. And we can easily imagine a very precise machine which tosses straws in such a way as to produce any distribution for chord lengths we please on a given circle.

Nevertheless, we are entitled to claim a definite frequency correspondence for the result (12.68). For there is one 'objective fact' which *has* been proved by the above derivation: any rain of straws which does *not* produce a frequency distribution agreeing with (12.68) will necessarily produce different distributions on different circles.

This is all we need in order to predict with confidence that the distribution (12.68) *will* be observed in any experiment where the 'region of uncertainty' is large compared with the circle. For, if we lack the skill to toss straws so that, with certainty, they intersect a given circle, then surely we lack *a fortiori* the skill consistently to produce different distributions on different circles *within* this region of uncertainty!

It is for this reason that distributions predicted by the method of transformation groups turn out to have a frequency correspondence after all. Strictly speaking, this result holds only in the limiting case of 'zero skill', but, as a moment's thought will show, the skill required to produce any appreciable deviation from (12.68) is so great that in practice it would be difficult to achieve even with a machine.

These conclusions seem to be in direct contradiction to those of von Mises (1957), who denied that such problems belong to the field of probability theory at all. It appears to us that if we were to adopt von Mises' philosophy of probability theory strictly and consistently, the range of legitimate physical applications of probability theory would be reduced almost to the vanishing point. Since we have made a definite, unequivocal prediction, this issue has now been removed from the realm of philosophy into that of verifiable fact. The predictive power of the transformation group method can be put to the test quite easily in this and other problems by performing the experiments.

The Bertrand experiment has, in fact, been performed by the writer and Dr Charles E. Tyler, tossing broom straws from a standing position onto a 5 inch diameter circle

drawn on the floor. Grouping the range of chord lengths into ten categories, 128 successful tosses confirmed Eq. (12.68) with an embarrassingly low value of chi-squared. However, experimental results will no doubt be more convincing if reported by others.

12.5 Comments

Bertrand's problem has a greater importance than appears at first glance, because it is a simple crystallization of a deeper paradox which has permeated much of probability theory from its beginnings. In 'real' physical applications, when we try to formulate the problem of interest in probability terms, we find almost always that a statement emerges which, like Bertrand's, appears too vague to determine any definite solution, because apparently essential things are left unspecified.

We elaborate the example noted in the introduction of the preceding section. Given a gas of N molecules in a volume V, with known intermolecular forces, total energy E, predict its molecular velocity distribution, the pressure, distribution for pressure fluctuations, viscosity, thermal conductivity, and diffusion constant. Here again the viewpoint expressed by most writers on probability theory would lead one to conclude that the problem has no definite solution because it is ill-posed; the things specified are grossly inadequate to determine any unique probability distribution over microstates. If we reject the principle of indifference, and insist that the only valid basis for assigning probabilities is frequency in some random experiment, it would again appear that the only way of determining these quantities is to perform the experiment.

It is, however, a matter of record that over a century ago, without benefit of any frequency data on positions and velocities of molecules, James Clark Maxwell was able to predict all these quantities correctly by a 'pure thought' probability analysis, which amounted to recognizing the 'equally possible' cases. In the case of viscosity, the predicted dependence on density appeared at first to contradict common sense, casting doubt on Maxwell's analysis. But when the experiments were performed they confirmed Maxwell's prediction, leading to the first great triumph of kinetic theory. These are solid, positive accomplishments; and they cannot be made to appear otherwise merely by deploring Maxwell's use of the principle of indifference.

Likewise, we calculate the probability for obtaining various hands at poker; and we are so confident of the results that we are willing to risk money on bets which the calculations indicate are favorable to us. But underlying these calculations is the intuitive judgment that all distributions of cards are equally likely; and with a different judgment our calculations would give different results. Once again we are predicting definite, verifiable facts by 'pure thought' arguments based ultimately on recognizing the 'equally possible' cases; and yet present statistical doctrine, both orthodox and personalistic, denies that this is a valid basis for assigning probabilities!

The dilemma is thus apparent. On the one hand, one cannot deny the force of arguments which, by pointing to such things as Bertrand's paradox, demonstrate the ambiguities and

dangers in the principle of indifference. On the other hand, it is equally undeniable that use of this principle has, over and over again, led to correct, nontrivial, and useful predictions. Thus it appears that, although we cannot wholly accept the principle of indifference, we cannot wholly reject it either; to do so would be to cast out some of the most important and successful applications of probability theory.

The transformation group method grew out of the writer's conviction that the principle of indifference has been unjustly maligned in the past; what it has needed was not blanket condemnation, but recognition of the proper way to apply it. We agree with most other writers on probability theory that it is dangerous to apply this principle at the level of indifference between *events,* because our intuition is a very unreliable guide in such matters, as Bertrand's paradox illustrates.

The principle of indifference may, in our view, be applied legitimately at the more abstract level of indifference between *problems;* because that is a matter that is definitely determined by the statement of a problem, independently of our intuition. Every circumstance left unspecified in the statement of a problem defines an invariance property which the solution must have if there is to be any definite solution at all. The transformation group, which expresses these invariances mathematically, imposes definite restrictions on the form of the solution, and in many cases fully determines it.

Of course, not all invariances are useful. For example, the statement of Bertrand's problem does not specify the time of day at which the straws are tossed, the color of the circle, the luminosity of Betelgeuse, or the number of oysters in Chesapeake Bay; from which we infer, correctly, that if the problem as stated is to have a unique solution, it must not depend on these circumstances. But this would not help us unless we had previously thought that these things might be germane.

Study of a number of cases makes it appear that the aforementioned dilemma can now be resolved as follows. We suggest that the cases in which the principle of indifference has been applied successfully in the past are just the ones in which the solution can be 'reverbalized' so that the actual calculations used are seen as an application of indifference between problems, rather than events.

The transformation group derivation of the Jeffreys prior enables us to see that prior 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 $\mathrm{d}\mu = (1 + \theta^{-2})\,\mathrm{d}\theta$. 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 notion 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 (12.50) 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.

Our discussion of maximum entropy has shown a more constructive way of looking at this, however. 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.