Uncertainty in prediction and in inference *

J.Hilgevoord and J. Uffink
Department of History and Foundations of Science
University of Utrecht,
P.O.Box 80.000, 3508 TA Utrecht, the Netherlands

Abstract

The concepts of uncertainty in prediction and inference are introduced and illustrated using the diffraction of light as an example. The close relationship between the concepts of uncertainty in inference and resolving power is noted. A general quantitative measure of uncertainty in inference can be obtained by means of the so-called statistical distance between probability distributions. When applied to quantum mechanics, this distance leads to a measure of the distinguishability of quantum states, which essentially is the absolute value of the matrix element between the states. The importance of this result to the quantum mechanical uncertainty principle is noted. The second part of the paper provides a derivation of the statistical distance on basis of the so-called method of support.

One summer night, under Sicilian skies, we discussed Fisher's concept of likelihood with John Bell. He didn't like it! To him, the words likelihood and probability meant just the same thing. We argued that Fisher had used the term likelihood to indicate something fundamentally different from probability. But John was not impressed; he proposed that we use the word Fisherhood, if we really had to.

In this article we will discuss the concept Fisher had in mind when he used the word likelihood, and try to show its relevance to physics and to the uncertainty principle in particular. And although there is no immediate connection with the problems relating to Bell's theorem, we want to dedicate this article to John Bell for his sixtieth birthday, because his work, more than anything else, has brought

*appeared in *Foundations of Physics* **21** (1991), 323-341

about the renaissance of the interest in the conceptual problems of quantum mechanics that we witness today.

1. Introduction and summary

Consider the following problem in probability theory: an urn is filled with black and white stones in some definite proportion and is well-shaken. Now a single stone is drawn. How well can one predict the colour of this stone? The standard answer is that this depends on the proportion of black and white stones. If nearly all the stones are of the same colour one feels almost certain about the outcome of this experiment. If, on the other hand, the proportions are equal one feels completely uncertain about the outcome.

In problems of this kind we are dealing with uncertainty about the result of an experiment, based on a given description of the experimental arrangement. There is, however, also a second kind of question. Suppose that some information about the experimental arrangement is missing. For example, the proportion of black and white stones in an urn is unknown. Yet a number of outcomes of the experiment have been observed. For example, a number of stones have been drawn from the urn (with replacement and reshaking) and we have found in 10 draws 7 black and 3 white stones. What can be inferred about the contents of the urn, and with what (un)certainty?

The above illustrates two kinds of problems that typically arise in probability theory and statistics, the problem of prediction and the problem of inference. The uncertainties which are connected with these problems are conceptually quite distinct. The first is an uncertainty about an outcome when a probability distribution is given; the second is an uncertainty about a probability distribution given an outcome. We shall refer to these uncertainties as uncertainties of the first and second kind.

The distinction we encounter here is by no means confined to problems relating to urns and the like, as can be seen from the following example. Consider a plane light wave which is diffracted by a slit in a screen. We assume that the direction of incidence is given. The solid curve in fig. 1 shows the diffraction pattern which is observed on a photographic plate behind the slit. (If l is the

distance between the screen and the plate, a the width of the slit, and λ the wavelength of the light, we assume $l \gg a \gg \lambda$.) If the intensity of the light is very low, the diffraction pattern is seen to build up gradually as a result of the impact of many individual photons, each producing a small black dot. The curve in fig. 1 can be taken to represent the probability for photons to arrive at different points on the photographic plate. Given this probability distribution one may ask: with what uncertainty can one predict the place where the next photon will arrive? It is usual to express this uncertainty by means of some measure of the width or spread of the probability distribution. Forgoing for the moment a precise definition of this measure, its numerical value, in the present example, is evidently of the order of magnitude of the width of the central diffraction peak, i.e. $\approx \frac{l\lambda}{a}$.

Now, suppose that the direction of the incoming plane wave is unknown, and that a photon has been detected on the photographic plate, what can then be said about the direction of the incoming beam? This question arises in the theory of image formation if the slit is part of an optical instrument. Light beams from different directions form diffraction patterns which are shifted with respect to each other on the photographic plate. Hence, our problem is to assign the detected photon to one of a set of shifted diffraction patterns, i.e. we must make an inference about probability distributions from a given outcome. In the present case, the most reasonable candidate is the pattern which has its maximum at the point where the photon has been detected. But this assignment is not completely certain; slightly shifted patterns cannot be ruled out, i.e. light beams from slightly different directions could also have produced this particular impact. Forgoing again a precise definition of the uncertainty involved in this problem, it is evident that the numerical value of this uncertainty is again of the order of magnitude of the width of the central diffraction peak.

There is a close relation between the second kind of uncertainty, in this example, and the resolving power of an optical instrument. In a microscope or a telescope, the aperture of the instrument acts as a slit, and the image of a point source is a diffraction pattern of essentially the same shape as the one in fig. 1, the width of the central peak being determined by the size of the aperture. Dis-

tinguishing different point sources therefore amounts to distinguishing between shifted diffraction patterns, and the limited resolving power of the instrument results from an uncertainty in assigning detected photons to a diffraction pattern. According to Rayleigh's criterion two diffraction patterns, as in fig. 1, are said to be just distinguishable if the shift between them is so large that the maximum of the one coincides with the first minimum of the other. This gives a resolution distance of the order of magnitude of the width of the central peak.

The example of diffraction of light at a slit shows that the two kinds of uncertainty introduced at the beginning of this article have clear physical meanings. The example is a less clear illustration of the fundamental difference between these uncertainties in that their magnitudes happen to be about equal. However, consider a grating consisting of N identical slits of width a, separated by a center-to-center distance b. The diffraction pattern formed by such a grating is roughly indicated in fig. 2. The envelope of the pattern (broken line) is, apart from a normalization factor, the diffraction pattern of a single slit of width a; the width of its central peak is $\approx \frac{l\lambda}{a}$. The pattern itself consists of narrow peaks, the principal maxima, separated by $N-2$ subsidiary maxima. The width of these peaks is $\approx \frac{l\lambda}{Nb}$. If one now asks the same two questions as before, the corresponding uncertainties turn out to be very different. The uncertainty in the prediction of the point where the next photon will land depends on the bulk of the probability distribution. If 90% of the total probability is concentrated on a small subset of the possible outcomes, the uncertainty in the prediction of the outcome is small; if 90% of the total probability is distributed more or less evenly over a large subset of possible outcomes, the uncertainty in predicting the outcome is large. In the present example, the bulk of the probability lies within the central peak of the envelope of the diffraction pattern. If we choose to measure the uncertainty by the length of a connected interval on which this bulk is located, the uncertainty is $\approx \frac{l\lambda}{a}$. If we allow the interval to be disconnected, its total length is equal to the number of peaks within the central peak of the envelope (which is $\approx \frac{b}{a}$) times the width of these peaks. The resulting uncertainty now is $\approx \frac{l\lambda}{Na}$. Measures of bulk-width of either type can be taken, but whatever measure one chooses, the uncertainty in prediction depends on the bulk of the probability

distribution. If, on the other hand, we are to infer the direction of the incoming beam from the observation of a photon, we must ask how well mutually shifted diffraction patterns can be distinguished. If we apply the Rayleigh criterion, the corresponding uncertainty is determined by the width of the principal maxima, and its value, therefore, is $\approx \frac{l\lambda}{Nb}$. Thus, the two uncertainties, in this example, differ numerically. The uncertainty of the second kind is much smaller than that of the first kind. This explains why a grating is regarded to be such an accurate instrument: not because it allows one to make better predictions about the location of the impact of a photon, but because it allows one to make much sharper inferences about parameters of the incoming beam than would be possible with a single slit. More importantly, however, the two kinds of uncertainty are found to depend on different features of the diffraction pattern and, indeed, on different parameters of the experimental set-up.

Considering the above examples, it will not come as a surprise that uncertainties of the first and second kind are relevant to the quantum mechanical uncertainty principle. In his famous microscope experiment, Heisenberg used the resolving power to quantify the uncertainty in the position of the object. We have shown elsewhere [(1)] how his argument can be construed as relating an uncertainty of the second kind to one of the first kind. Likewise, the uncertainties occurring in many other applications of the uncertainty principle can be interpreted in this way. In the standard literature on the uncertainty principle, starting with the work of Heisenberg himself, no sharp distinction is made between the two kinds of uncertainty; both are taken to represent the spread of the bulk of a given probability distribution. This may be a consequence of the fact that most authors, when discussing the uncertainty principle, have in mind simple probability distributions, like the one in fig. 1, with only one main peak.

When the distinction, both conceptually and numerically, between uncertainties in prediction and inference has been recognized, the next important point is to look for general quantitative measures of these uncertainties, especially of those of the second kind. Quantitative measures for the bulk-width of a given distribution are well-known; most of the measures of uncertainty that have been used to express the uncertainty principle are of this kind. Such measures are

discussed systematically in ref. 2. In the rest of this article we shall deal with measures of uncertainty of the second kind only.

Measures of uncertainty of the second kind seem to be pretty well unexplored in physics, the exception being the resolving power. Indeed, the notion of such an uncertainty does not seem to appear in quantum mechanics texts at all, although, as we have seen, the concept is thoroughly physical. On the other hand, the corresponding problem in statistics, i.e. the problem of distinguishing between probability distributions on the basis of given evidence, has been intensively studied by statisticians. The purpose of this article is to communicate some of this work and to show how it applies to quantum mechanics. The result, which is both simple and remarkable, will be summarized below.

Let $p_i^{(1)}$ and $p_i^{(2)}$, $(i = 1, \ldots, m)$, be two probability distributions, for simplicity taken to be discrete. These distributions are possible descriptions of an experiment. How well can they be distinguished, given an outcome of the experiment, and with what (un)certainty can this be done? The answer to this question depends on the outcome. If the two probability distributions assign very different probabilities to this particular outcome, they are well distinguished and the uncertainty involved in choosing between them is small. If, on the other hand, they assign almost equal probabilities to the outcome, they are not well distinguished and the uncertainty is large. One may, however, also look for criteria for the distinguishability of two probability distributions which are independent of a particular outcome. We shall use the term 'uncertainty of the second kind' to cover both cases. The 'second-kindness' refers primarily to the fact that criteria of this sort involve a comparison between probability distributions instead of characterizing a single distribution.

In particular, W.K. Wootters [3] has defined, on the basis of purely statistical considerations, the concept of a statistical distance between two probability distributions. The larger (smaller) the statistical distance between two probability distributions, the easier (harder) it is to distinguish between them. This distance

is found to be

$$d(p^{(1)}, p^{(2)}) = \arccos \sum_i \sqrt{p_i^{(1)} p_i^{(2)}}, \tag{1}$$

which may be regarded as the angle between two unit vectors with components $\sqrt{p_i^{(1)}}$ and $\sqrt{p_i^{(2)}}$. This distance vanishes when the two probability distributions coincide, and it attains its maximum value $\pi/2$ when the two vectors are orthogonal, i.e. when each outcome which has a positive probability according to one distribution has zero probability according to the other. Note the occurrence of the square roots of the probabilities, suggesting quantum mechanics, but appearing in a purely classical context. Following Wootters, let us apply this result to quantum mechanics. In quantum mechanics the problem takes the following form. Let $\psi^{(1)}$ and $\psi^{(2)}$ be two quantum states we want to distinguish on basis of the results of measurements. If A is a non-degenerate observable with a complete orthonormal set of eigenstates ϕ_i, the probability distributions over the eigenvalues of A in the states $\psi^{(1)}$ and $\psi^{(2)}$ are $|\langle\phi_i|\psi^{(1)}\rangle|^2$ and $|\langle\phi_i|\psi^{(2)}\rangle|^2$ respectively. The corresponding statistical distance (1) is

$$\arccos \sum_i |\langle\phi_i|\psi^{(1)}\rangle\langle\phi_i|\psi^{(2)}\rangle|$$

which is the statistical distance between $\psi^{(1)}$ and $\psi^{(2)}$ relative to the measurement of A. However, other measurements are in principle possible; so we may ask for the most discriminating measurement, i.e. the one producing the largest statistical distance between $\psi^{(1)}$ and $\psi^{(2)}$. It is easily seen that the most discriminating observables are the ones that have either $\psi^{(1)}$ or $\psi^{(2)}$ among their eigenstates. The resulting largest distance is

$$d(\psi^{(1)}, \psi^{(2)}) = \arccos |\langle\psi^{(1)}|\psi^{(2)}\rangle|. \tag{2}$$

This is called the *absolute* statistical distance between the quantum states $\psi^{(1)}$ and $\psi^{(2)}$. This simple result is completely in harmony with what a quantum physicist would intuitively expect. According to the usual interpretation $|\langle\psi^{(1)}|\psi^{(2)}\rangle|^2$ is the probability of finding the system in the state $\psi^{(2)}$ when it was prepared in the state $\psi^{(1)}$. We now find the same quantity appearing in a purely statistical criterion for the distinguishability of the two quantum states.

Crucial to this result is the occurrence of the square roots of $p^{(1)}$ and $p^{(2)}$ in the classical expression (1).

The distance (2) provides a natural measure of uncertainty of the second kind. Applying it, for example, to our shifted diffraction patterns, we can define this uncertainty as the smallest shift producing a fixed statistical distance between the original and the shifted pattern. Obviously, we have here a generalization of Rayleigh's criterion. Applying it in a similar way to an unstable state (the shift now being a shift in time!) we get a generalization of the concept of the half-life of a state. It is a remarkable fact that an uncertainty relation exists between uncertainties of the second kind, defined on basis of (2), and uncertainties of the first kind [1,4]. This relation provides a satisfactory mathematical basis for Heisenberg's microscope argument. It also leads to a completely general uncertainty relation for the lifetime and line width of a decaying state [1]. The next section will be devoted to a more elaborate analysis of the statistical distance, using a somewhat different approach from the one taken by Wootters.

2. Statistical distance

2.1 The method of support

The problem of statistical inference can be formulated as follows. Consider an experiment with a specified set X of possible outcomes. We shall assume in our notation that X is a continuum, although this is not essential to the argument, and denote its elements as $x \in X$. A class of probability distributions over X, representing candidate descriptions of the preparation of the experiment, is also assumed to be given. We assume that these distributions are labeled (uniquely) by a parameter $\theta \in \Theta$, where Θ is some arbitrary parameter space, which may be discrete, continuous, multi-dimensional or whatever.

Thus, each parameter value θ specifies a probability distribution $p_\theta(x)$. We assume that one of these distributions provides the actual or true description of the experiment, but that we don't know which one. The question is then: suppose we have observed an outcome x, or a set of outcomes $x = (x_1, \ldots, x_n)$ in n independent repetitions of the experiment, what can we say about the

unknown parameter θ?

Intuitively, one would feel that some sort of judgement should be possible on the basis of the observed data, especially when n is large. But it is also clear that in general this judgement cannot take the form of a definite conclusion or deduction which rules out certain values of θ as being inconsistent with the data.

How then should we proceed in formulating such a judgement? This problem is fraught with difficulties and controversies, and no generally accepted answer is available. Instead, the several procedures which have been proposed have led to the development of different schools which are still engaged in heated debate.

On one side there are those who follow the Bayesian method. Here, one starts from the basic principle that in a case where the actual probability distribution is unknown it is always legitimate to assign 'prior' probabilities to the various probability distributions. On the other side are the orthodox methods which start by constructing certain functions from the data (test functions, estimators, etc.) of which certain 'nice properties' are demanded to hold in the long run, i.e. upon repeated application of the procedure.

This is not the place to go into the controversies surrounding these methods. (Cf. V. Barnett [5].) Both schools have found eminent supporters as well as eloquent critics. It is sufficient for our purpose to note that any statistical method must somehow appeal to some non-trivial basic principles, or achieve nothing.

There is also a third method of statistical inference which is called 'likelihood inference' or the 'method of support' and has been expounded by I. Hacking [6] and A.W.F. Edwards [7] . We shall adopt this method in the sequel. The basic principles of this method are:

a. All the information provided by the data x about the value of θ is contained in the function

$$L_x(\theta) \equiv p_\theta(x)$$

b. The ratio $L_x(\theta_0)/L_x(\theta_1)$ can be interpreted as a degree of relative support, in the sense that the data provide stronger support for θ_0 than for θ_1 if, and in so far as, this ratio exceeds unity.

The function $L_x(\theta)$ was introduced and named the 'likelihood of θ' by R.A. Fisher in 1922. Fisher wanted to discriminate between the study of $p_\theta(x)$ as a distribution of x for given θ and as a function of θ for given x. The latter standpoint is of course the most important for statistical inference. But $p_\theta(x)$ is not a probability distribution over θ, so that a different name seemed appropriate. However, Fisher also wanted to convey a certain meaning to the concept of likelihood, as opposed to probability. Yet the fact that these terms are synonymous in everyday speech has not helped to clarify the distinction. Terms like 'support' (Jeffreys [8], Hacking, Edwards), 'decisiveness of evidence' (Jeffreys [9]) 'weight of evidence' (Good [10]) or perhaps 'information for discrimination in favor of θ_0 against θ_1' (Kullback [11]) seem to be more apt to express the purported meaning.

Indeed, if one finds data x such that $p_{\theta_0}(x)$ is close to unity, whereas $p_{\theta_1}(x)$ is close to zero, one cannot conclude that these data make θ_0 'more likely' or 'more probable' than θ_1, without making an assumption about the prior probability of these parameters. But it does not seem unreasonable to state that these data by themselves provide more support or stronger evidence for θ_0 than for θ_1, quite irrespective of the question what their prior probabilities were, or indeed, whether such probabilities were defined at all.

The method of support is of course no less controversial than any other method of statistical inference. But it has some intuitive appeal and the fact that writers holding very different viewpoints on the nature of probability have embraced this method may lend some respectability to it.

It should be noted that 'support', as it is used above, is only a relative concept. It defines an ordering relation on the parameter space. Any strictly monotonous function of the likelihood ratio would render the same service and would therefore also be acceptable as a measure of support. It is customary and convenient to take the logarithm of the likelihood function as a mathematical measure of support. Thus,

$$S_x(\theta) = \log L_x(\theta)$$

is taken as a definition of the support function. Although this choice is somewhat

arbitrary it has the important advantage that support becomes additive under independent repetitions of the experiment. Note also that the support function may change under coordinate transformations on X. Indeed, if $y(x)$ is some bijective transformation, the support function transforms into

$$\hat{S}_y(\theta) = \log \hat{p}_\theta(y) = \log \left(p_\theta(x) |\frac{dx}{dy}| \right) = S_x(\theta) + \log |\frac{dx}{dy}|$$

But differences in support for various parameter values, which are the only quantities of interest for our purpose, are in fact invariant under such coordinate transformations.

The method described above thus works in practice by attaching relative degrees of support to all candidate parameter values θ. In particular, the value θ_{max} for which $S_x(\theta)$ is maximal is the best supported or maximum likelihood estimate. But we can read off more from the support function than just its maximum. This is of particular importance when one asks how certain we are of this estimate.

Let us illustrate this by a simple example. Imagine stones are drawn from an urn containing black and white stones in a proportion $\theta/(1-\theta)$. $(0 \leq \theta \leq 1.)$ Suppose we have made 10 draws (with replacement) and found 7 black and 3 white stones. The best supported value of θ is then 0.7. Now suppose we have made 100 draws and found 70 black, 30 white. The best supported value is still 0.7, but now one would feel much more certain about this guess since the amount of data is so much larger. It is easy to see this difference reflected in the shape of the support or likelihood functions. In the case of 100 draws the support function drops of much more rapidly around its maximum than in the case of 10 draws.

It seems natural then to consider some measure of the width of the support function as an indication of the uncertainty in our guess. This width simply represents the size of a region of θ values which are nearly as well supported as the best supported value. A particularly simple expression which may serve as an inverse measure of this width is the curvature of the support function at its

maximum value, i.e.

$$I_x = -\left.\frac{d^2}{d\theta^2} S_x(\theta)\right|_{\theta=\theta_{max}}$$

Edwards has called this the 'observed information in the data'. The motivation for this choice is simply that if the support function is smooth, it may be approximated in the neighbourhood of its maximum by the parabola:

$$S_x(\theta) \approx S_x(\theta_{max}) - \frac{1}{2} I_x \, (\theta - \theta_{max})^2 \tag{3}$$

The quantity $\frac{1}{\sqrt{I_x}}$ measures the width of this parabola.

The above expression gives reasonable values in the example considered above, as well as in many other examples, including the shifted diffraction patterns of the previous section, but it is clear that it can only serve as a rather provisional and tentative element in our discussion. In general, support functions may take arbitrary forms. They need not have a unique maximum. They need not be continuous or differentiable, or the approximation (3) may be bad. Moreover, this choice makes 'information' dependent on the parametrization of the distributions. For example, Pitman [12] has noted that if we put $\phi = (\theta - \theta_{max})^3$ any 'information' one might have about θ will transform into zero 'information' about ϕ. This behavior is quite undesirable, because an inference about the unknown value of θ is equivalent to an inference about the unknown value of ϕ. Furthermore, the above choice only works as long as θ is a continuous one-dimensional parameter. However, we shall see that all these obstacles can be overcome.

2.2 Expected support

The support function depends, of course, on the actual data x obtained from the experiment. In a theoretical study we can only compare the various support functions associated with each possible set of data that might be produced by the experiment. However, since the number of possible sets of data is often enormous, such a study would not be feasible in general. Therefore, we shall approach the subject from a slightly different angle. Suppose that the experiment is described

by the distribution $p_{\theta_0}(x)$. What form of the support function would we then expect? The expected support function can be written as

$$\langle S(\theta)\rangle_{\theta_0} = \int p_{\theta_0}(x) S_x(\theta)\, dx = \int p_{\theta_0}(x) \log p_\theta(x)\, dx \tag{4}$$

Note that the expected support no longer depends on observed data but on the (hypothetically) true value θ_0. Note also that due to the lucky choice of the logarithm in the measure of support, the expected support is proportional to the actual support one would obtain in n independent repetitions of the experiment if the various possible outcomes occurred in relative frequencies equal to the probabilities $p_{\theta_0}(x)$.

Expressions of the form (4) have been extensively studied in information theory, and are well-known for their remarkable properties [11]. To mention only one of them: $\langle S(\theta)\rangle_{\theta_0}$ attains its maximum value for $\theta = \theta_0$, and this maximum is unique. Thus, the true value has maximum expected support. This indicates that although support functions can behave wildly, expected support functions are better-behaved.

Let us now return to the notion of uncertainty of the second kind. According to our previous choice we may take

$$I(\theta_0) = -\frac{d^2}{d\theta^2} \langle S(\theta)\rangle_{\theta_0} \Big|_{\theta=\theta_0} \tag{5}$$

as an inverse measure of the uncertainty to be expected in our inference. Fisher called this ' the amount of information to be anticipated from our observation' or also the 'intrinsic accuracy of the distribution'. Today it is commonly called 'Fisher information'. The meaning of this quantity is of course similar to that of I_x in the sense that $\frac{1}{\sqrt{I(\theta_0)}}$ gives a rough measure of the width of the expected support function around its maximum.

Note that $I(\theta_0)$ is additive under repetitions of the experiment, so that the amount of information to be anticipated in an n-fold repetition of the experiment is just n times the information of a single performance. Equivalently, the size of the uncertainty region decreases by a factor $\frac{1}{\sqrt{n}}$. Also, $I(\theta_0)$ is invariant under coordinate transformations on X, but not under parameter transformations on

Θ. Under fairly general regularity conditions $I(\theta_0)$ can also be written as

$$I(\theta_0) = \int \frac{1}{p_\theta}\left(\frac{dp_\theta}{d\theta}\right)^2 dx\bigg|_{\theta=\theta_0} = 4\int \left(\frac{d\sqrt{p_\theta}}{d\theta}\right)^2 dx\bigg|_{\theta=\theta_0}$$

showing more explicitly that the quantity is always non-negative. The Fisher information is particularly well-known for its appearance in the famous Cramér-Rao inequality [13,14], by which $\frac{1}{\sqrt{I(\theta_0)}}$ puts a lower bound to the standard deviation of all unbiassed estimators for θ. This explains in what sense it represents the 'intrinsic accuracy' with which θ can be estimated.

The Fisher information is a concept which may fruitfully be compared to that of the resolving power of an optical instrument. Indeed, when the problem is to distinguish between a set of shifted probability distributions, i.e. when $p_\theta(x) = p(x-\theta)$, the Fisher information becomes a constant, independent of the hypothetically true value θ_0. Likewise, as long as the assumption can be made that the diffraction patterns differ only by a shift, the resolving power is also a constant, characterizing the optical instrument. If one calculates the Fisher information for the patterns discussed in the previous section one obtains $I = \frac{4}{3}(\frac{a}{\ell\lambda})^2$ in the case of a single slit, and $I \approx \frac{4}{3}(\frac{Nb}{\ell\lambda})^2$ for the grating, so that $\frac{1}{\sqrt{I}}$ indeed gives a reasonable size for the resolving distance for those distributions. But the Fisher information is obviously much more generally defined than a resolving power based on the Rayleigh, or some similar, criterion, and it can also be applied to distributions of arbitrary shape (Gaussians, Cauchy distributions, etc.).

However, the most important point of similarity between the Fisher information and the resolving power is that both expressions are intended to represent uncertainty in an inference problem. This is reflected by the fact that they do not look at the spread of a single probability distribution over the set of possible outcomes (as would be the case for measures of uncertainty of the first kind), but rather at the sensitivity of the probability distributions to changes of θ.

Still, however useful and important the Fisher information may be in statistical theory, it cannot be the end of the story. As noted before, a conventional change of the parametrization may change the value of I. Also, the expression

is entirely useless for distributions failing to meet the regularity conditions, such as

$$p_\theta(x) = \begin{cases} 1 & \frac{1}{2} - \theta \leq x \leq \frac{1}{2} + \theta \\ 0 & \text{elsewhere} \end{cases} \tag{6}$$

2.3 Statistical distance

It has been observed by several authors (Rao [14] , Jeffreys [15]) that the Fisher information provides a geometry on the parameter space Θ. To introduce this point of view, it may be useful to consider first the case where $p_\theta(x) = p(x - \theta)$, for some regular distribution p. In this case $\frac{1}{\sqrt{I}}$ represents a region of uncertainty, the size of which is independent of the true value θ_0. Notice also that when θ has a physical dimension, $\frac{1}{\sqrt{I}}$ has the same dimension. It then becomes attractive to regard $\frac{1}{\sqrt{I}}$ as a natural unit of uncertainty, which may be used to 'normalize' the parametrization. That is, we replace θ by the dimensionless parameter $s = \frac{1}{2}\sqrt{I}\,\theta$. The meaning of such a normalized parameter is not difficult to see: If for two parameter values s_1 and s_2 we have $|s_1 - s_2| \ll 1$ then, even if one of these parameters should happen to be the true one, we may not expect to obtain decisive support for this value against the other from the outcome of the experiment. The distributions characterized by these parameter values are then hard to distinguish experimentally. Thus, normalizing the parameter provides a way of expressing how far apart two distributions are in a statistical sense.

If the Fisher information is not independent of θ_0, this idea becomes somewhat more involved. One may however still perform this normalization locally to obtain an infinitesimal distance element between θ_0 and $\theta_0 + d\theta$:

$$ds = \frac{1}{2}\sqrt{I(\theta_0)}\,d\theta$$

This can be integrated to obtain a total distance between the distributions p_{θ_0} and p_{θ_1}:

$$s(p_{\theta_0}, p_{\theta_1}) = \frac{1}{2}\int_{\theta_0}^{\theta_1} \sqrt{I(\theta)}\,d\theta \tag{7}$$

This normalized parameter defines a distance between the two probability distributions. This distance is invariant under parameter transformations on Θ;

it depends only on the set of distributions $\{p_\theta\}$ with which we started. This step thus removes one more obstacle we encountered in the use of the Fisher information. But there are still several left. In the first place, it should be noted that while a small distance between two distributions implies that they are hard to distinguish, it is not necessarily true that they are easy to distinguish if the distance (7) is large. The distribution displayed in fig. 2 provides an example of this phenomenon. Here, $\frac{1}{\sqrt{I}}$ is of the order of magnitude of the width of the principal maxima, so that any two distributions which are shifted by this or a larger amount will differ by a statistical distance greater than unity. However, if we consider a shift of the order of the distance between the principal maxima, the two distributions will no longer be so easy to distinguish because of the recurring overlap of the different maxima. A large value of the statistical distance (7) between two probability distributions generally implies only that there are many distinguishable pairs between them, it does not imply that these distributions are themselves easily distinguishable. Moreover, this distance function is still useless for distributions like (6), which have an infinite distance between any two different values of θ. But these obstacles can also be removed as we will now show.

The geometrical meaning of the Fisher information becomes perhaps even more evident if we pass to a case where θ is a multi-dimensional parameter, $\vec{\theta} = (\theta_1, \ldots, \theta_k)$. In this case the natural analogue of (5) is

$$\begin{aligned} I_{ij}(\vec{\theta}_0) &= -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \langle S(\vec{\theta}) \rangle_{\vec{\theta}_0} \Big|_{\vec{\theta}=\vec{\theta}_0} \\ &= \int \frac{1}{p_{\vec{\theta}}} \frac{\partial p_{\vec{\theta}}}{\partial\theta_i} \frac{\partial p_{\vec{\theta}}}{\partial\theta_j}\, dx \Big|_{\vec{\theta}=\vec{\theta}_0} \\ &= 4 \int \frac{\partial \sqrt{p_{\vec{\theta}}}}{\partial\theta_i} \frac{\partial \sqrt{p_{\vec{\theta}}}}{\partial\theta_j}\, dx \Big|_{\vec{\theta}=\vec{\theta}_0} \qquad i,j = 1 \ldots k \end{aligned}$$

Fisher's 'amount of information' is now not a quantity but a matrix. This positive definite and symmetrical matrix endows the parameter space with a metric which is invariant under parameter transformations. Indeed, if we define a general

infinitesimal element of distance in parameter space by

$$ds^2 = \frac{1}{4}\sum_{ij} I_{ij}(\vec{\theta})d\theta_i d\theta_j,$$

this distance is manifestly invariant under any regular parameter transformation.

From this geometrical point of view it is natural to regard the previous case of the one-dimensional parameter as a one-dimensional subset, i.e. as a curve in a metrical space of higher dimension. The distance (7) then measures the statistical length along this curve. But in this higher dimensional space one can make a distinction that would not be possible in a one-dimensional space, namely between length and distance. The distance between two points is defined as the length of the shortest curve that connects them in parameter space.

This idea becomes particularly fruitful if we take the parameter space Θ to map the set of *all* probability distributions over X. In that case Θ generally becomes a space of infinite dimension. For simplicity, let us first consider the case where X contains a finite number of possible outcomes, $X = \{x_1, \ldots, x_m\}$. In this case $m-1$ independent parameters suffice to map the totality of all probability distributions over X. The parameter space Θ is then a Riemannian space of dimension $m-1$, endowed with the metrical tensor:

$$I_{ij} = 4\sum_{\alpha=1}^{m} \frac{\partial\sqrt{p_{\vec{\theta}}(x_\alpha)}}{\partial\theta_i}\frac{\partial\sqrt{p_{\vec{\theta}}(x_\alpha)}}{\partial\theta_j} \qquad i,j = 1\ldots, m-1.$$

By choosing an appropriate system of coordinates, such a Riemannian space can always be represented as a curved subspace immersed in a Euclidean space of higher dimension. For the present case such a representation is obtained if we choose as coordinates $y_\alpha(\vec{\theta}) = \sqrt{p_{\vec{\theta}}(x_\alpha)}$. The distance element then takes the simple form:

$$ds^2 = \frac{1}{4}\sum_{i,j=1}^{m-1} I_{ij}d\theta_i d\theta_j = \sum_{\alpha=1}^{m}(dy_\alpha)^2$$

which is the usual distance element in m-dimensional Euclidean space. The parameter space Θ can thus be represented as an $m-1$-dimensional hypersurface in this space. This hypersurface is determined by the condition that the probability distributions be normalized, i.e. $\sum_\alpha (y_\alpha)^2 = 1$, and it is thus seen to be the

surface of a unit hypersphere. The condition $y_\alpha(\vec{\theta}) \geq 0$, for all α, further restricts this surface to a convex segment of the hypersphere. Thus, quite remarkably, the metric defined by the Fisher information gives Θ the geometry of the surface of a sphere.

Now consider a regular one-dimensional curve $\{p_\theta\}$ on the surface of this sphere connecting the points representing the probability distributions p_{θ_0} and p_{θ_1}. The shortest such curve is, of course, a piece of a great circle and the length of this curve is just the arc length between these points. The statistical distance between two probability distributions, therefore, is given by:

$$d(p_{\theta_0}, p_{\theta_1}) = \arccos \sum \sqrt{p_{\theta_0} p_{\theta_1}}.$$

In the case where X is a continuum, the above argument becomes more technical. The conclusion, however, is quite analogous, namely that the statistical distance is now given by

$$d(p_{\theta_0}, p_{\theta_1}) = \arccos \int \sqrt{p_{\theta_0} p_{\theta_1}}\, dx. \tag{8}$$

This expression, finally, represents a generally useful measure of the distinguishability of two distributions. The distance (8), of course, is always finite. It is invariant under parameter tranformations on Θ as well as coordinate transformations on X. It is also independent of the original curve $\{p_\theta\}$ from which p_{θ_0} and p_{θ_1} where taken. It depends merely on the overlap integral of the two probability distributions.

Returning once more to the diffraction pattern of fig. 2 we see that the distance (8) as a function of the shift first increases rapidly to attain roughly its maximum value when the principal peaks no longer overlap, but decreases again when different principal maxima start to overlap, etc. This behavior is just what one would expect from a distance that is supposed to measure distinguishability. It also shows that the curve in parameter space that corresponds to shifting the pattern is not a geodesic curve.

For the family of distributions (6) we see that the distance between two adjacent points on the curve is finite and of the order

$$d(p_\theta, p_{\theta+d\theta}) = \sqrt{2d\theta},$$

whereas the distance between these points taken along the curve corresponding to shifting the distributions is infinite. The fact that the distance element is here proportional to $\sqrt{d\theta}$, instead of $d\theta$ itself, reflects the irregularity of this curve. That the distinguishability should increase so rapidly for small shifts is not

References

[1] J. Hilgevoord and J. Uffink, in *Sixty-two Years of Uncertainty: Historical, Philosophical and Physics Inquiries into the Foundations of Quantum Mechanics*, A.I. Miller ed. (to appear).

[2] J. Uffink, *Measures of Uncertainty and the Uncertainty Principle*, (University of Utrecht thesis, 1990).

[3] W.K. Wootters, *Phys. Rev.* **D23**, 357 (1981).

[4] J. Uffink and J. Hilgevoord, *Found. Phys.* **15**, 925 (1985).

[5] V. Barnett, *Comparative Statistical Inference*, (John Wiley and Sons, New York, 2nd ed., 1982).

[6] I. Hacking, *Logic of Statistical Inference*, (Cambridge University Press, Cambridge, 1965).

[7] A.W.F. Edwards, *Likelihood*, (Cambridge University Press, Cambridge, 1972).

[8] H. Jeffreys, *Proc. Camb. Phil. Soc.* **32**, 618, (1936).

[9] H. Jeffreys, *Theory of Probability*, (2nd ed. Clarendon Press, Oxford, 1948).

[10] I.J. Good, *Probability and the Weighing of Evidence* (Charles Griffin, London, 1950).

[11] S. Kullback, *Information Theory and Statistics*, (John Wiley and Sons, New York, 1959).

[12] E.J.G. Pitman *Some Basic Theory for Statistical Inference*, (Chapman and Hall, London, 1979).

[13] M.G. Kendall and A. Stuart, *the Advanced Theory of Statistics*, Vol. 2, 3th edn. (Charles Griffin, London, 1973), p.9.

[14] C.R. Rao, *Bull. Calcutta Math. Soc.*, **37**, (1945).

[15] See Ref. 9, p. 157.

[16] See ref. 9. p. 143.

[17] A. Bhattacharyya *Bull. Calcutta Math. Soc.* **35**, 99, (1943).

[18] S. Kakutani *Ann. Math.*, **49**, 214, (1948).

Figure Captions

Fig. 1 Diffraction patterns of a slit.

Fig. 2 Diffraction pattern of a grating.