Journal of Statistical Physics, Vol. 6, No. 2/3, 1972

The Interpretation of Inductive Probabilities

Arthur Hobson[1]

Received February 25, 1972

An apparent inconsistency in the inductive logic interpretation of probabilities is examined and resolved.

KEY WORDS: Probability; inductive logic; information theory; Jaynes.

In an interesting recent article,[1] Friedman and Shimony (hereafter referred to as FS) presented an example which seems to imply an inconsistency in the "inductive logic interpretation" of probability theory.[2] According to the inductive logic interpretation, probability theory is the formalism for inductive reasoning. Any such inconsistency would have serious implications for the "information theory approach" to statistical mechanics proposed by Jaynes,[3,4] since this approach appears to require the inductive logic interpretation of probability theory. FS suggest three possible ways of avoiding the inconsistency, but do not ascertain whether one of them will actually resolve the difficulty.

The purpose of this note is to present the apparent inconsistency in a more straightforward and more general light, and to discuss its resolution. It will first be shown that the example of FS is a special case of a very general phenomenon: the proper resolution of the difficulty will then be given; finally, it will be shown that the resolutions suggested by FS are invalid.

Consider a repeatable experiment with possible outcomes (on a single trial) labeled i ($i = 1, 2,..., r$). We make the following definitions:

Supported by a grant from the National Science Foundation.

[1] Department of Physics, University of Arkansas, Fayetteville, Arkansas.

B: Background data which describe the experiment and give the possible outcomes, and which may give other data relative to single trials. However, we assume that the data B are symmetric with respect to different trials (i.e., do not distinguish between trials), and that the data B make no reference to sequences of trials (i.e., the data B do not link the outcome of one trial to the outcome of any other trial).

$f(i)$: A random variable (r.v.), i.e., a numerical function of the possible outcomes in a single trial.[5]

$\bar{f}$: The average of $f(i)$ over an infinite sequence of trials, defined by

$$\bar{f} = \lim_{n\to\infty} (1/n) \sum_{k=1}^{n} f(i_k)$$

where i_k represents the outcome of the kth trial. Note that $\bar{f}$ is itself a r.v. defined on the joint experiment consisting of an infinite sequence of trials.

D_F: The proposition that, in an infinite sequence of trials, $\bar{f}$ takes on the numerical value F.

Although it is not relevant to the following argument, it is worth noting that the probabilities $P(i \mid BD_F)$ that i will occur on a single trial given B and D_F are exponential in the r.v. $f(i)$,[3] since D_F implies the constraint

$$\sum_{i=1}^{r} f(i)\, P(i \mid BD_F) = F$$

Thus, D_F is essentially the same as the data d_ϵ of FS.

The apparent contradiction noted by FS arises from the following theorem:

Theorem. On the basis of B, D_F has probability one when $F = \langle f \rangle$ and probability zero when $F \neq \langle f \rangle$, where $\langle f \rangle$ is defined by

$$\langle f \rangle = \sum_{i=1}^{r} f(i)\, P(i \mid B)$$

In other words, the probability density of D_F, given B, is

$$\rho(D_F \mid B) = \delta(F - \langle f \rangle)$$

Proof. Since by assumption the data B treat all trials identically and independently, Jaynes's maximum-entropy prescription leads to a joint

n-trial distribution which is symmetric and uncorrelated (see, for instance, p. 78 of Ref. 4):

$$P(i_1 i_2 \cdots i_n \mid B) = \prod_{k=1}^{n} P(i_k \mid B)$$

It is now an elementary consequence of probability theory (called the "law of large numbers")[5] that $\bar{f}$ is concentrated around the single-trial mean $\langle f \rangle$ in the manner indicated in the theorem.

The apparent contradiction is now as follows: According to the theorem, $\bar{f} = \langle f \rangle$ with probability one (on the basis of B), so it appears that we can confidently *predict* $\bar{f}$ on the basis of B. But this seems absurd: A mere description of the experiment cannot tell us the long-run average of every r.v. $f(i)$.

An example will help sharpen the argument and the apparent inconsistency. Suppose the experiment is the tossing of a single die. Then B is the information that the die is a cube, with i spots on the ith side ($i = 1,..., 6$). On the basis of B, we obviously have (using Jaynes's approach,[3,4] or simply using the standard "principle of insufficient reason"[1-4])

$$P(i \mid B) = 1/6 \qquad i = 1,..., 6$$

The single-trial mean of the r.v. $f(i) = i$ (the number of spots showing) is then $\langle i \rangle = 3.5$. The above theorem tells us that, in an infinite series of trials, the average value $\bar{i}$ is 3.5 with probability one (on the basis of B). But this seems absurd. For instance, the die might be weighted in such a way that only even numbers can occur, in which case an actual measurement of $\bar{i}$ might yield a value close to $\bar{i} = 4$.[2] Nevertheless, it is still true that, on the basis of the data B (which do *not* include any information about weighting), $\bar{i} = 3.5$ with probability one. This seems paradoxical.

When the difficulty is presented in the above manner (rather than via the somewhat complicated example of FS), it becomes obvious that the apparent paradox is due to the usual misunderstanding regarding the interpretation of inductive probabilities: The source of the difficulty is that inductive predictions (even when they are "certain," i.e., true with probability one) are only the *best* predictions possible on the basis of the given data. The predictions are not *deduced* from the data, they are only *induced* from the data. Thus, even if the data are true, the predictions (including even predictions which are "certain") may turn out to be experimentally wrong. In the die example,

[2] Of course, we cannot actually carry out an infinite number of trials to measure $\bar{i}$ exactly; the best we can do is make a large but finite number of trials in order to approximate $\bar{i}$ experimentally.

there is nothing in the data B which deductively implies $\bar{i} = 3.5$. Nevertheless, we can *induce* (on the basis of B) that $\bar{i} = 3.5$ with probability one. But this inductive conclusion might be wrong, even though B is true: Information relevent to the prediction of $\bar{i}$ (for instance, information about weighting of the die) might not be included in B, in which case any estimate, even an estimate which is "certain," based on B alone is likely to be wrong. To state the situation more succinctly, $P(X \mid Y) = 1$ does *not* say that Y implies X [although the converse is true: if Y implies X, then $P(X \mid Y) = 1$].

Thus, the resolution of the difficulty is simply that inductive predictions, even when they are "certain," may turn out to be wrong if the data on which they were based are incomplete in some important respect. Nevertheless, they are still the best predictions available on the basis of the data. Experimental evidence E that such predictions are wrong then means that the original data B are either incorrect or incomplete in some important respect. The new evidence E should then be used (along with B) in making further predictions; this is precisely what Jaynes's approach is designed to do.[3,4]

Real-life examples of this situation are abundant. For instance, in the Stern–Gerlach experiment (where neutral silver atoms are passed through an inhomogeneous magnetic field and then allowed to impinge on a screen), if the background data B include no information about the quantization of the spin of the valence electron, then we obtain the inductive prediction that, with near certainty, the pattern of impact points on the screen will show an even apread from top to bottom. Note that this is an *inductive*, not *deductive*, prediction: The atoms enter the apparatus in "random," i.e., unknown, orientations, so we cannot use mechanics to deduce the precise pattern from precisely known initial conditions. Experimentally, the predicted pattern is not observed: Only two small impact points are observed, one at the top and the other at the bottom of the previously predicted pattern. Thus one must reevaluate the data. This is, in fact, precisely how the quantization of angular momentum was discovered.

Concerning the three resolutions proposed by FS: The above analysis shows that the proposition D_F (denoted d_ϵ by FS) can be well-defined, that probabilities $P(D_F \mid B)$ are well-defined, and that there is no need to reject or restrict any of the principles of inductive probabilities. Thus, all three proposed resolutions are invalid.

ACKNOWLEDGMENT

I would like to thank Professor Harry S. Robertson of the University of Miami for bringing the article by Friedman and Shimony to my attention.

REFERENCES

1. K. Friedman and A. Shimony, *J. Stat. Phys.* **3**:381 (1971).
2. Pierre Simon de Laplace, *A Philosophical Essay on Probabilities* (Dover, New York, 1951); J. M. Keynes, *A Treatise on Probability* (Macmillan, London and New York, 1921); Harold Jeffreys, *Theory of Probability*, 3rd Ed. (Oxford Univ. Press, London, 1961); R. T. Cox, *The Algebra of Probable Inference* (The Johns Hopkins Press, Baltimore, 1961).
3. E. T. Jaynes, *Phys. Rev.* **106**:620 (1957); *Phys. Rev.* **108**:171 (1957); in *Statistical Physics* (Brandeis Lectures in Theoretical Physics, Vol. 3), K. W. Ford, ed. (Benjamin, New York, 1963); in *Delaware Seminar in the Foundations of Physics*, M. Bunge, ed. (Springer, New York, 1967).
4. A. Hobson, *Concepts in Statistical Mechanics* (Gordon and Breach, New York, 1971).
5. A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1965).