# INFERENCE, METHOD AND DECISION

# SYNTHESE LIBRARY

MONOGRAPHS ON EPISTEMOLOGY,

LOGIC, METHODOLOGY, PHILOSOPHY OF SCIENCE,

SOCIOLOGY OF SCIENCE AND OF KNOWLEDGE,

AND ON THE MATHEMATICAL METHODS OF

SOCIAL AND BEHAVIOURAL SCIENCES



VOLUME 115

ROGER D. ROSENKRANTZ
*Virginia Polytechnic Institute and State University, Blacksburg, U.S.A.*

# INFERENCE, METHOD AND DECISION

*Towards a Bayesian Philosophy of Science*

D. REIDEL PUBLISHING COMPANY
DORDRECHT-HOLLAND / BOSTON-U.S.A.

Library of Congress Cataloging in Publication Data

Rosenkrantz, Roger D. 1938-
Inference, method, and decision.

(Synthese library; v. 115)
Includes bibliographies and indexes.
1. Science–Philosophy. 2. Bayesian statistical decision theory. I. Title.
Q175.R554 501 77–9399
ISBN-13: 978-90-277-0818-2 e-ISBN-13: 978-94-010-1237-9
DOI: 10.1007/978-94-010-1237-9

Published by D. Reidel Publishing Company,
P.O. Box 17, Dordrecht, Holland

Sold and distributed in the U.S.A., Canada, and Mexico
by D. Reidel Publishing Company, Inc.
Lincoln Building, 160 Old Derby Street, Hingham,
Mass. 02043, U.S.A.

*To Margaret*

# PREFACE

This book grew out of previously published papers of mine composed over a period of years; they have been reworked (sometimes beyond recognition) so as to form a reasonably coherent whole.

Part One treats of informative inference. I argue (Chapter 2) that the traditional principle of induction in its clearest formulation (that laws are confirmed by their positive cases) is clearly false. Other formulations in terms of the 'uniformity of nature' or the 'resemblance of the future to the past' seem to me hopelessly unclear. From a Bayesian point of view, 'learning from experience' goes by conditionalization (Bayes' rule). The traditional stumbling block for Bayesians has been to find objective probability inputs to conditionalize upon. Subjective Bayesians allow any probability inputs that do not violate the usual axioms of probability. Many subjectivists grant that this liberality seems prodigal but own themselves unable to think of additional constraints that might plausibly be imposed. To be sure, if we could agree on the correct probabilistic representation of 'ignorance' (or absence of pertinent data), then all probabilities obtained by applying Bayes' rule to an 'informationless' prior would be objective. But familiar contradictions, like the Bertrand paradox, are thought to vitiate all attempts to objectify 'ignorance'. Building on the earlier work of Sir Harold Jeffreys, E. T. Jaynes, and the more recent work of G. E. P. Box and G. E.Tiao, I have elected to bite this bullet. In Chapter 3, I develop and defend an objectivist Bayesian approach.

Given a number of constraints, say, mean values or symmetries, Jaynes' *maximum entropy rule* singles out that prior distribution which, among all those consistent with the given constraints, maximizes entropy (a measure of uncertainty). Far from 'generating knowledge out of ignorance', this rule enjoins us not to pretend to knowledge we do not possess. I regard it as a postulate of rationality, on a par with the 'sure-thing principle' or the expected utility rule.

Jaynes shows that, of all distributions consonant with the given constraints, the maximum entropy distribution is realized in by far the highest proportion of possible experimental outcomes. Consequently, if an actual outcome fails to fit the maximum entropy distribution at all well, one can be practically certain either that additional constraints are operative or that the given constraints are not wholly correct. (In that event, the initial distribution is not revised by conditionalization, but retracted.) To many of my readers, this will have a whiff of rationalism about it, but I think it a defensible rationalism. The given constraints are presumably empirical, but given their correctness and exhaustiveness, the correct distribution should be discoverable by pure thought in the same way that abstract argument leads one to the normal distribution for a trait that depends on a large number of independent factors. I contend there are no conceptual distinctions to be drawn between data distributions and prior distributions; both are arrived at by the same kind of reasoning and both are equally 'subjective' or 'objective'. Whether a given distribution counts as one or the other is largely a function of the use to be made of it. For this reason, maximum entropy distributions are often amenable to empirical check, and where they fail, other or different constraints are invariably discovered. Further justification of Jaynes' rule is found in its essential agreement with another rule of Sir Harold Jeffreys', and a more general rule recently devised by Box and Tiao. The confluence of these methods (which are based on quite disparate intuitions) is evidence of their soundness in the same way that the agreement of different definitions of 'recursive' evidences Church's thesis.

Invariance requirements are especially powerful constraints, often sufficient to uniquely determine a distribution. Jaynes has generated a prior for the Bertrand problem by an invariance argument and empirically verified it. In actual fact, absence of data rarely, if ever, connotes total ignorance, and consequently, it is a mistake to suppose that any old transformation of parameters is admissible. We must rather uncover the admissible re-parametrizations by asking for the equivalent reformulations of a problem, and these will often be implicit in the statement of the problem itself. Thus, we might be asked for the probability that $k$ is the first significant digit of an entry in a table of data (say, the areas of the world's largest islands), $k = 1, \ldots, 9$. Since nothing is said about the scale units employed, scale invariance is implied. That is, if there were such a 'law of first digits', it should be independent of scale.

Given the intuitive adequacy of the rules for generating 'no data' priors, the objectivist position begins to seem quite tenable. Even conditionalization has a derivative status, from this perspective. Any correct rule for revising probabilities in the light of the additional information that the 'conditioning event' has occurred should have the property of leading from a correct representation of the one state of partial knowledge to a correct representation of the other. That Bayes' rule does have this property is compelling evidence that it is the (uniquely) correct solution of the 'kinematical' problem.

From the standpoint of objectivist Bayesianism, one compares theories by their (objective) probabilities. I argue, in Part Three, that probabilities are sufficient to guide research, and against the advisability of imposing formal rules of acceptance and rejection for theories. It is also a corollary of a thoroughgoing Bayesian philosophy of science that other traits of theories which have been thought important enter objectively only to the extent that they are reflected in higher probability. Simplicity (measured in a way that generalizes the familiar 'paucity-of-parameters' criterion) is shown to be such a trait in Chapter 5, and the implications of this fundamental observation are traced out in the remainder of Part Two (and are felt in Part Three, as well). Since the detailed discussion of these matters is rather technical, a brief and more intuitive account may be useful to the reader.

I show, specifically, that in a broad spectrum of cases, the 'simpler' of two theories in equally good agreement with the data is better supported, and so also, confirmed. 'Support' and 'confirmation' are used throughout in a precise sense related to probability. A theory is *confirmed* by an observation when its probability is increased thereby. As for 'support', suppose that we have a partition of hypotheses, $H_i$ (exactly one must be true). Then the relative support accorded them by an observation $x$ is registered by the conditional probabilities, $P(x/H_i)$. Considered as a function of an hypothesis $H$, $P(x/H)$ is called the *likelihood function* (associated with $x$). The 'likeliest' or best hypothesis maximizes the likelihood function; it accords what was observed the highest probability. Bayes' rule for revising probabilities states that the new (or 'posterior') probability of $H$ is proportional to the prior probability, $P(H)$, and its likelihood, $P(x/H)$, and inversely proportional to $P(x)$, in symbols: $P(H/x) = P(H)P(x/H)/P(x)$.

Likelihood, then, measures the impact of the data on probability: the posterior probability is a blend of the initially given information, as represented by $P(H)$, and the sample information, as embodied in the likelihood function. Of two competing hypotheses, $H$ and $K$, the 'likelier' will be confirmed (its probability will go up, while that of the other hypothesis goes down). Indeed, using Bayes' rule, $P(H/x) : P(K/x) = [P(H) : P(K)]\,[P(x/H) : P(x/K)]$, since $P(x)$ cancels. This expresses the *posterior odds* as a product of the *prior odds*, $P(H) : P(K)$, and the *likelihood ratio*, $P(x/H) : P(x/K)$.

Scientific theories often contain adjustable parameters which must be estimated from the data each time the theory is applied. How might the support of two such theories be compared? One approach, often used, is to compare the support of their 'likeliest' or 'best-fitting' special cases (obtained by fitting the free parameters). The two theories, however, may differ in the number of their free parameters, and typically, a theory with more free parameters will fit more possible observations (it will be less 'simple', in my sense). This feature is surely relevant to any comparative evaluation of the theories. But it is unclear, at a purely judgmental level, how much to discount for a theory's comparative lack of simplicity when comparing two theories in the indicated way.

A Bayesian solution of the problem would take the following form. Let the two hypotheses, $H$ and $K$, have special cases $H_1, H_2, H_3$ and $K_1, K_2, K_3$ with prior probabilities and likelihoods as shown below:

$$P(x/H_1) = 1/5, P(x/H_2) = 1/10, P(x/H_3) = 1/20 \,,$$

$$P(H_1) = 6/20, P(H_2) = 3/20, P(H_3) = 1/20 \,;$$

$$P(x/K_1) = 1/4, P(x/K_2) = 1/6, P(x/K_3) = 1/12 \,,$$

$$P(K_1) = 3/20, P(K_2) = 4/20, P(K_3) = 3/20 \,.$$

Here I am assuming, of course, that the hypotheses $H_1 - H_3, K_1 - K_3$ are mutually exclusive and jointly exhaustive. Since a theory holds if some special case of it holds, the probability of a theory is the sum of the probabilities of its special cases (or, more accurately, of its 'ultimate' special cases, obtained by fixing the values of all its free parameters). To obtain posterior probabilities, we must therefore sum the posterior probabilities of the theory's special cases. Since $P(H_i/x) \propto P(x/H_i)P(H_i)$, we have:

$$P(H/x) \propto P(x/H_1)P(H_1) + P(x/H_2)P(H_2) + P(x/H_3)P(H_3)$$

$$+ (1/5)\,(6/20) + (1/10)\;\;(3/20) + (1/20)\;\;(1/20) = 0.0775.$$

Likewise, $P(K/x) \propto 0.0833$. Since $P(H/x$ and $P(K/x)$ must continue to sum to 1, $P(H/x) = 0.4819$ and $P(K/x) = 0.5181$. Hence, $K$ is better supported (and so confirmed), its probability increasing ever so slightly, while that of $H$ suffers a corresponding decrement.

We have obtained the support by averaging over the likelihoods of the special cases of the theory, each such likelihood being weighted by the probability of the corresponding special case. When we add a parameter, complicating a theory, we add to the number of special cases over which we must average the likelihood function, and because comparatively few of its special cases will fit the data at all well, the average likelihood is degraded.

Both accuracy and simplicity are therefore reflected in a theory's average likelihood or support, but in a mathematically determinate way that obviates the need for a judgmental assessment of the relative importance of these two factors.

Given the determinate rate of exchange between accuracy and simplicity, we can say by how much a theory's accuracy must be improved in complicating it for its support to increase. (That the support increases means that the accuracy gained is sufficient to offset the simplicity lost.) For example, a circular orbit may fail to fit the plotted positions of a planet very well, and the question arises whether an elliptical orbit (which contains the eccentricity as an additional parameter) would be better supported. (To compare the average likelihoods in this case requires a probabilistic treatment of the observational errors.) A circle is, of course, a special case of an ellipse, and so the two hypotheses are not logically exclusive. Strictly speaking, Bayes' rule can only be applied to a comparison of exclusive alternatives, but the difficulty can be circumvented in practice by taking logical differences: e.g., we compare the hypothesis of a circle to that of a *proper* ellipse (viz. an ellipse of positive eccentricity). The method of average likelihood is applied to some interesting examples from genetics in Chapter 5 and to curve-fitting and related problems in Chapter 11, where its performance is compared with that of several non-Bayesian methods.

One additional point is worth noting. It is evident that the average of the likelihood function can never exceed its maximum value. This observation is mathematically trivial, yet fraught with methodological significance. For it translates into the statement that a theory (with parameters) can never be better supported than its best-fitting (or 'likeliest') special case. We will see (Chapter 7) that the two main arguments Copernicus marshalls on behalf of the heliocentric hypothesis are but applications of this principle. And the Copernican example illustrates the applicability of Bayesian methods to deterministic, as well as to probabilistic, theories, for the former typically include a probabilistic treatment of error.

One friend who read this book in manuscript remarked that it seemed to fall between two stools, for many of the scientific examples are akin to those one might encounter in a statistics text, while others are more like those found in works on the history and philosophy of science. This aspect of the book is not unintentional. My deliberate aim here is to bring these formerly disparate areas of study together in one unified treatment – whence my subtitle. The symbiosis should be mutually enriching, and of interest to philosophers and historians of science, scientists, and statisticians alike.

The Bayesian position which I develop and the applications of it to issues in the philosophy of science owes much to the writings of I. J. Good and E. T. Jaynes. If the book makes their important work accessible to a wider philosophical public, I will not have labored in vain. To Good, I am also indebted for much stimulating conversation and many useful comments.

My early interest in the subject developed while I was a graduate student at Stanford, and the members of my dissertation committee, Professors Patrick Suppes, Jaakko Hintikka and Joseph Sneed, provided an especially exciting and supportive environment in which to develop my ideas. They have continued to provide much stimulation and support in the intervening years. My thinking was further shaped by participation in a discussion group organized by the late Allan Birnbaum. Allan encouraged my own tendency to view statistics and philosophy of science as a connected whole, and, more than that, to test one's views and methods in the crucible of actual scientific research. His patient but sharp criticisms of Bayesian methods incited me to investigate their applicability in a serious way and the fruits of that search are scattered through Chapters 5, 7, 9, 10 and 11. The other members of the group, Ronald Giere and Isaac Levi, have also done much to help me avert minor skirmishes and focus critical attention on the more significant issues that separate my position from theirs.

The list of those who have influenced me could go on and on, but I would be remiss indeed if I failed to mention Stephen Spielman, Barry Loewer, W. K. Goosens, Kenneth Friedman, and my former student, Robert Laddaga. Their characteristically insightful remarks have helped to advance my thinking at crucial points, forced its clearer expression, and saved me from serious errors. I owe them all a very large debt.

Earlier versions of Chapters 5, 7, and 9 were delivered at the 1973 and 1975 University of Western Ontario Workshops organized by W. A. Harper, C. K. Hooker and James Leach. A talk that evolved into Chapter 1 was delivered at Ernest Adams' Berkeley seminar, while another which grew into Chapter 11 was delivered at Patrick Suppes' probability seminar at Stanford (both in 1975). I am grateful to all of these men for the opportunity to try out my ideas, and to the participants in those seminars and workshops for their many useful comments and criticisms.

I would also like to thank Mrs. Dabney Whipple for her patient and accurate typing of the manuscript and her help in preparing it for publication. Last, but far from least, my wife, Margaret, to whom the book is dedicated, programmed the empirical Bayes/orthodox comparisons of Chapter 11 and provided the moral and material support without which the book could not have been written.

# TABLE OF CONTENTS