Springer Series in Statistics

Advisors:
D. Brillinger, S. Fienberg, J. Gani,
J. Hartigan, K. Krickeberg

Springer Series in Statistics

L. A. Goodman and W. H. Kruskal, Measures of Association for Cross Classifications. x, 146 pages, 1979.

J. O. Berger, Statistical Decision Theory: Foundations, Concepts, and Methods. xiv, 425 pages, 1980.

R. G. Miller, Jr., Simultaneous Statistical Inference, 2nd edition. xvi, 299 pages, 1981.

P. Brémaud, Point Processes and Queues: Martingale Dynamics. xviii, 354 pages, 1981.

E. Seneta, Non-Negative Matrices and Markov Chains. xv, 279 pages, 1981.

F. J. Anscombe, Computing in Statistical Science through APL. xvi, 426 pages, 1981.

J. W. Pratt and J. D. Gibbons, Concepts of Nonparametric Theory. xvi, 462 pages, 1981.

V. Vapnik, Estimation of Dependences based on Empirical Data. xvi, 399 pages, 1982.

H. Heyer, Theory of Statistical Experiments. x, 289 pages, 1982.

L. Sachs, Applied Statistics: A Handbook of Techniques. xxviii, 706 pages, 1982.

M. R. Leadbetter, G. Lindgren and H. Rootzen, Extremes and Related Properties of Random Sequences and Processes. xii, 336 pages, 1983.

H. Kres, Statistical Tables for Multivariate Analysis. xxii, 504 pages, 1983.

J. A. Hartigan, Bayes Theory. xii, 145 pages, 1983.

J. A. Hartigan

Bayes Theory

Springer-Verlag
New York Berlin Heidelberg Tokyo

J. A. Hartigan

Department of Statistics
Yale University
Box 2179 Yale Station
New Haven, CT 06520
U.S.A.

AMS Classification: 62 A15

Library of Congress Cataloging in Publication Data

Hartigan, J. A.
Bayes theory.
(Springer series in statistics)
Includes bibliographies and index.
1. Mathematical statistics. I. Title. II. Series.
QA276.H392 1983 519.5 83-10591

With 4 figures.

Typeset by Thomson Press (India) Limited, New Delhi, India.

9 8 7 6 5 4 3 2 1

ISBN-13:978-1-4613-8244-7 e-ISBN-13:978-1-4613-8242-3
DOI: 10.1007/978-1-4613-8242-3

To Jenny

Preface

This book is based on lectures given at Yale in 1971–1981 to students prepared with a course in measure-theoretic probability.

It contains one technical innovation—probability distributions in which the total probability is infinite. Such improper distributions arise embarrassingly frequently in Bayes theory, especially in establishing correspondences between Bayesian and Fisherian techniques. Infinite probabilities create interesting complications in defining conditional probability and limit concepts.

The main results are theoretical, probabilistic conclusions derived from probabilistic assumptions. A useful theory requires rules for constructing and interpreting probabilities. Probabilities are computed from similarities, using a formalization of the idea that the future will probably be like the past. Probabilities are objectively derived from similarities, but similarities are subjective judgments of individuals.

Of course the theorems remain true in any interpretation of probability that satisfies the formal axioms.

My colleague David Potlard helped a lot, especially with Chapter 13. Dan Barry read proof.

Contents

CHAPTER 1

Theories of Probability

1.0. Introduction

A theory of probability will be taken to be an axiom system that probabilities must satisfy, together with rules for constructing and interpreting probabilities. A person using the theory will construct some probabilities according to the rules, compute other probabilities according to the axioms, and then interpret these probabilities according to the rules; if the interpretation is unreasonable perhaps the original construction will be adjusted.

To begin with, consider the simple finite axioms in which there are a number of elementary events just one of which must occur, events are unions of elementary events, and the probability of an event is the sum of the non-negative probabilities of the elementary events contained in it.

There are three types of theory—logical, empirical and subjective. In logical theories, the probability of an event is the rational degree of belief in the event relative to some given evidence. In empirical theories, a probability is a factual statement about the world. In subjective theories, a probability is an individual degree of belief; these theories differ from logical theories in that different individuals are expected to have different probabilities for an event, even when their knowledge is the same.

1.1. Logical Theories: Laplace

The first logical theory is that of Laplace (1814), who defined the probability of an event to be the number of favorable cases divided by the total number of cases possible. Here cases are elementary events; it is necessary to identify equiprobable elementary events in order to apply Laplace's theory. In many gambling problems, such as tossing a die or drawing from a shuffled deck of cards, we are willing to accept such equiprobability judgments because of the apparent physical indistinguishability of the elementary events—the particular face of the die to fall, or the particular card to be drawn. In other problems, such as the probability of it raining tomorrow, the equiprobable alternatives are not easily seen. Laplace, following Bernoulli (1713) used the principle of insufficient reason which specifies that probabilities of two events will be equal if we have no reason to believe them different. An early user of this principle was Thomas Bayes (1763), who apologetically postulated that a binomial parameter p was uniformly distributed if nothing were known about it.

The principle of insufficient reason is now rejected because it sets rather too many probabilities equal. Having an unknown p uniformly distributed is different from having an unknown $\sqrt{p}$ uniformly distributed, yet we are equally ignorant of both. Even in the gambling case, we might set all combination of throws of n dice to have equal probability so that the next throw has probability 1/6 of giving an ace no matter what the results of previous throws. Yet the dice will always be a little biased and we want the next throw to have higher probability of giving an ace if aces appeared with frequency greater than 1/6 in previous throws.

Here, it is a consequence of the principle of insufficient reason that the long run frequency of aces will be 1/6, and this prediction may well be violated by the observed frequency. Of course any finite sequence will not offer a strict contradiction, but as a practical matter, if a thousand tosses yielded 1/3 aces, no gambler would be willing to continue paying off aces at 5 to 1. The principle of insufficient reason thus violates the skeptical principle that you can't be sure about the future.

1.2. Logical Theories: Keynes and Jeffreys

Keynes (1921) believed that probability was the rational belief in a proposition justified by knowledge of another proposition. It is not possible to give a numerical value to every such belief, but it is possible to compare some pairs of beliefs. He modified the principle of insufficient reason to a principle of indifference—two alternatives are equally probable if there is no relevant evidence relating to one alternative, unless there is corresponding evidence relating to the other. This still leaves a lot of room for judgment; for example, Keynes asserts that an urn containing n black and white balls in unknown proportion will produce each sequence of white and black balls with equal probability, so that for large n the proportion of white balls is very probably near 1/2. He discusses probabilities arising from analogy, but does not present methods for practical calculation of such probabilities. Keynes's theory does not succeed because it does not provide reasonable rules for computing probabilities, or even for making comparisons between probabilities.

Jeffreys (1939) has the same view of probability as Keynes, but is more constructive in presenting many types of prior distributions appropriate for different statistical problems. He presents an "invariant" prior distribution for a continuous parameter indexing a family of probability distributions, thus escaping one of the objections to the principle of insufficient reason. The invariant distribution is however inconsistent in another sense, in that it may generate conditional distributions that are not consistent with the global distribution. Jeffreys rejects it in certain standard cases.

Many of the standard prior probabilities used today are due to Jeffreys, and he has given some general rules for constructing probabilities. He concedes (1939, p. 37) that there may not be an agreed upon probability in some cases, but argues (p. 406) that two people following the same rules should arrive at the same probabilities. However, the many rules stated frequently give contradictory results.

The difficulty with Jeffreys's approach is that it is not possible to construct unique probabilities according to the stated rules; it is not possible to infer what Jeffreys means by probability by examining his constructive rules; it is not possible to interpret the results of a Jeffreys calculation.

1.3. Empirical Theories: Von Mises

Let $x_1, x_2, \ldots, x_n, \ldots$ denote an infinite sequence of points in a set. Let $f(A)$ be the limiting proportion of points lying in a set A, if that limit exists. Then f satisfies the axioms of finite probability. In frequency theories, probabilities correspond to frequencies in some (perhaps hypothetical) sequence of experiments. For example "the probability of an ace is 1/6" means that if the same die were tossed repeatedly under similar conditions the limiting frequency would be 1/6.

Von Mises (1928/1964) declares that the objects under study are not single events but sequences of events. Empirically observed sequences are of course always finite. Some empirically observed sequences show approximate convergence of relative frequencies as the sample size increases, and approximate random order. Von Mises idealizes these properties in an infinite sequence or *collective* in which each elementary event has limiting frequency that does not change when it is computed on any subsequence in a certain family. The requirement of invariance is supposed to represent the impossibility (empirically observed) of constructing a winning betting system.

Non trivial collectives do not exist satisfying invariance over all subsequences but it is a consequence of the strong law of large numbers that collectives exist that are invariant over any specified countable set of subsequences. Church (1940) suggests selecting subsequences using recursive functions, functions of integer variables for which an algorithm exists that will compute the value of the function for any values of the arguments in finite time on a finite computing machine. There are countably many recursive functions so the collective exists, although of course, it cannot be constructed. Further interesting mathematical developments are due to Kolmogorov (1965) who defines a finite sequence to be random if an algorithm required to compute it is sufficiently complex, in a certain sense; and to Martin-Löf (1966) who establishes the existence of finite and infinite random sequences that satisfy all statistical tests.

How is the von Mises theory to be applied? Presumably to those finite sequences whose empirical properties of convergent relative frequency and approximate randomness suggested the infinite sequence idealization. No rules are given by von Mises for recognizing such sequences and indeed he criticizes the "erroneous practice of drawing statistical conclusions from short sequences of observations" (p. ix). However the Kolmogorov or Martin-Lof procedures could certainly be used to recognize such sequences.

How does frequency probability help us learn? Take a long finite "random" sequence of 0's and 1's. The frequency of 0's in the first half of the sequence will be close to the frequency of 0's in the second half of the sequence, so that if we know only the first half of the sequence we can predict approximately the frequency of 0's in the second half, provided that we assume the whole sequence is random. The prediction of future frequency is just a tautology based on the assumption of randomness for the whole sequence.

It seems necessary to have a definition, or at least some rules, for deciding when a finite sequence is random to apply the von Mises theory. Given such a definition, it is possible to construct a logical probability distribution that will include the von Mises limiting frequencies: define the probability of the sequence $x_1, x_2, \dots, x_n$ as $\lim_k N_k(x)/N_k$ where $N_k(x)$ is the number of random sequences of length k beginning with $x_1, x_2, \dots, x_n$ and N_k is the number of random sequences of length k. In this way a probability is defined on events which are unions of finite sequences. A definition of randomness would not be acceptable unless $P[x_{n+1} = 1 \mid \text{proportion of 1's in } x_1, \dots, x_n = p_n] - p_n \to 0$ as $n \to \infty$, that is, unless the conditional probability of a 1 at the next trial converged to the limiting frequency of 1's.

True, definitions of randomness may vary, so that this is no unique solution—but the arbitrariness necessary to define finite randomness for applying frequency theory is the same arbitrariness which occurs in defining prior probabilities in the logical and subjective theories.

Asymptotically all theories agree; von Mises discusses only the asymptotic case; to apply a frequency theory to finite sequences, it is necessary to make the same kind of assumptions as Jeffreys makes on prior probabilities.

1.4. Empirical Theories: Kolmogorov

Kolmogorov (1933) formalized probability as measure: he *interpreted* probability as follows.

(1) There is assumed a complex of conditions C which allows any number of repetitions.
(2) A set of elementary events can occur on establishment of conditions C.
(3) The event A occurs if the elementary event which occurs lies in A.
(4) Under certain conditions, we may assume that the event A is assigned a probability $P(A)$ such that
   (a) one can be practically certain that if the complex of conditions C is repeated a large number of times n, then if m be the number of occurrences of event A, the ratio m/n will differ very slightly from $P(A)$.
   (b) if $P(A)$ is very small one can be practically certain that when conditions C are realized only once, the event A would not occur at all.

The axioms of finite probability will follow for $P(A)$, although the axiom of continuity will not.

As frequentists must, Kolmogorov is struggling to use Bernoulli's limit theorem for a sequence of independent identically distributed random variables without mentioning the word probability. Thus "the complex of conditions C which allows any number of repetitions"—how different must the conditions be between repetitions? Thus "practically certain" instead of "with high probability." Logical and subjective probabilists argue that a larger theory of probability is needed to make precise the rules of application of a frequency theory.

1.5. Empirical Theories: Falsifiable Models

Statisticians in general have followed Kolmogorov's prescription. They freely invent probability models, families of probability distributions that describe the results of an experiment. The models may be falsified by repeating the experiment often and noting that the observed results do not concur with the model; the falsification, using significance tests, is itself subject to uncertainty, which is described in terms of the original probability model. A direct interpretation of probability as frequency appears to need an informal extra theory of probability (matching the circularity in Laplace's equally possible cases), but the "falsifiable model" interpretation appears to avoid the circularity. We propose a probability model, and then reject it, or modify it, if the observed results seem improbable. We are using Kolmogorov's rule (4) (b) that "formally" improbable results are "practically" certain not to happen. If they do happen we doubt the formal probability.

The weaknesses in the model approach:

(1) The repetitions of the experiment are assumed to give independent, identically distributed results. Otherwise laws of large numbers will not apply. But you can't test that independence without taking some other series of experiments, requiring other assumptions of independence, and requiring other tests. In practice the assumption of independence is usually untested (often producing very poor estimates of empirical frequencies; for example, in predicting how often a complex piece of equipment will break, it is dangerous to assume the various components will break independently). The assumption of independence in the model theory is the analogue of the principle of insufficient reason in logical theories. We assume it unless there is evidence to the contrary, and we rarely collect evidence.

(2) Some parts of the model, such as countable additivity or continuity of a probability density, are not falsifiable by any finite number of observations.

(3) Arbitrary decisions about significance tests must be made; you must decide on an ordering of the possible observations on their degree of denial of the model—perhaps this ordering requires subjective judgment depending on past knowledge.

1.6. Subjective Theories: De Finetti

De Finetti (1930/1937) declares that the degree of probability attributed by an individual to a given event is revealed by the conditions under which he would be disposed to bet on that event. If an individual must bet on all events A which are unions of elementary events, he must bet according to some probability $P(A)$ defined by assigning non-negative probabilities to the elementary events, or else a dutch book can be made against him—a combination of bets is possible in which he will lose no matter which elementary event occurs. (This is only a little bit like von Mises's principle of the impossibility of a gambling system.) De Finetti calls such a system of bets coherent.

In the subjectivist view, probabilities are associated with an individual. Savage calls them "personal" probabilities; a person should be coherent, but any particular event may be assigned any probability without questioning from others. You cannot say that "my probability that it will rain this afternoon is .97" is wrong—it reports my willingness to bet at a certain rate. Bayes (1763) defines probability as "the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening." His probability describes how a person *ought* to bet, not how he *does* bet. It should be noted that the subjectivist theories insist that a person be coherent in his betting, so that they are not content to let a person bet how he pleases; psychological probability comes from the study of actual betting behavior, and indeed people are consistently incoherent (Wallsten (1974)).

There are numerous objections to the betting approach some technical (is it feasible?), others philosophical (is it useful?).

(i) *People don't wish to offer precise odds*—Smith (1961) and others have suggested ranges of probabilities for each event; this is not a very serious objection.

(ii) *A bet is a price, subject to market forces*—depending on the other actors; Borel (1924) considers the case of a poker player, who by betting high, increases his probability of winning the pot. Can you say to him "your probability of winning the pot is the amount you are willing to bet to win the pot divided by the amount in the pot."

Suppose you are in a room full of knowledgeable meteorologists, and you declare the probability it will rain tomorrow is .95. They all rush at you waving money. Don't you modify the probability? We may not be willing to bet at all if we feel others know more. Why should the presence of others be allowed to affect our probability?

(iii) *The utility of money is not linear*—You may bet \$1 to win \$500 when the chance of winning is only 1/1000; the gain of \$500 seems more than 500 times the loss of \$1. Ramsey (1926) and Savage (1954) advance theories of rational decision making, choosing among a range of available actions, that produce both utilities and probabilities for which the optimal decision is always that decision which maximizes expected utility.

The philosophical objection is that *I* don't particularly care how *you* (opinionated and uninformed as you are) wish to bet. To which the subjectivists will answer that subjective judgments are necessary in forming conclusions from observations; let us be explicit about them (Good (1976, p. 143)). To which the empiricists will reply, let us separate the "good" empirically verifiable probabilities, the likelihoods, from the "bad" subjective probabilities which vary from person to person. (Cox and Hinkley (1974, p. 389) "For the initial stages ... the approach is ... inapplicable because it treats information derived from data as on exactly equal footing with probabilities derived from vague and unspecified sources.")

1.7. Subjective Theories: Good

Good (1950) takes a degree of belief in a proposition E given a proposition H and a state of mind of a person M, to be a primitive notion allowing no precise definition. Comparisons are made between degrees of belief; a set of comparisons is called a *body of beliefs*. A *reasonable* body of beliefs contains no contradictory comparisons.

The usual axioms of probability are assumed to hold for a numerical probability which has the same orderings as a body of beliefs. Good recommends a number of rules for computing probabilities, including for example the device of imaginary results: consider a number of probability assignments to a certain event; in combination with other fixed probability judgments, each will lead through the axioms to further probability judgments; base your original choice for probabilities on the palatability of the overall probabilities which ensue. If an event of very small probability occurs, he suggests that the body of beliefs be modified.

Probability judgments can be sharpened by laying bets at suitable odds, but there is no attempt to define probability in terms of bets. Good (1976, p. 132) states that "since the degrees of belief, concerning events over which he has no control, of a person with ideally good judgment, should surely not depend on whether he uses his beliefs in any specific manner, it seems desirable to have justifications that do not mention preferences or utilities. But utilities necessarily come in whenever the beliefs are to be used in a practical problem involving action."

Good takes an attitude, similar to the empirical model theorists, that a probability system proposed is subject to change if errors are discovered through significance testing. In standard probability theory, changes in probability due to data take place according to the rules of conditional probability; in the model theory, some data may invalidate the whole probability system and so force changes not according to the laws of probability. There is no contradiction in following this practice because we separate the formal theory from the rules for its application.

1.8. All the Probabilities

An overview of the theories of probability may be taken from the stance of a subjective probabilist, since subjective probability includes all other theories. Let us begin with the assumption that an individual attaches to events numerical probabilities which satisfy the axioms of probability theory.

If no rules for constructing and interpreting probabilities are given, the probabilities are inapplicable—for all we know the person might be using length or mass or dollars or some other measure instead of probability. Thus the theories of Laplace and Keynes are not practicable for lack of rules to construct probability. Jeffreys provides rules for many situations (although the rules are inconsistent and somewhat arbitrary). Good takes a belief to be a primitive notion; although he gives numerous rules for refining and correcting sets of probabilities, I believe that different persons might give different probabilities under Good's system, on the same knowledge, simply because they make different formalizations of the primitive notion of degree of belief. Such disagreements are accepted in a subjective theory, but it seems undesirable that they are caused by confusion about meanings of probability. For example if you ask for the probability that it will rain tomorrow afternoon, one person might compute the relative frequency of rain on afternoons in the last month, another might compute the relative amount of today's rain that fell this afternoon; the axioms are satisfied. Are the differences in computation due to differences in beliefs about the world, or due to different interpretations of the word probability?

The obvious interpretation of a probability is as a betting ratio, the amount you bet over the amount you get. There are certainly some complications in this interpretation—if a probability is a price, it will be affected by the market in which the bet is made. But these difficulties are overcome by Savage's treatment of probability and utility in which an individual is asked to choose coherently between actions, and then must do so to maximize expected utility as measured by an implied personal probability and utility. The betting interpretation arises naturally out of the foundations of probability theory as a guide to gamblers, and is not particularly attached to any theory of probability. A logical probabilist, like Bayes, will say that a probability is what you *ought* to bet. A frequentist will say that a bet is justified only if it would be profitable in the long run—Fisher's evaluation of estimation procedures rests on which would be more profitable in the long run. A subjectivist will say that the probability is the amount *you* are willing to bet, although he will require coherence among your bets. It is therefore possible to adopt the betting interpretation without being committed to a particular theory of probability.

As Good has said, the frequency theory is neither necessary nor sufficient. Not sufficient because it is applicable to a single type of data. Not necessary because it is neatly contained in logical or subjectivist theories, either through Bernoulli's celebrated law of large numbers which originally generated the frequency theory, or through de Finetti's celebrated convergence of conditional probabilities on exchangeable sequences, which makes it clear what probability judgments are necessary to justify a frequency theory. (A sequence $x_1, x_2, \ldots, x_n, \ldots$ is exchangeable if its distribution is invariant under finite permutations of the indices, and then if the x_i have finite second moment, the expected value of x_{n+1} given $x_1, \ldots, x_n$ and $(1/n)\sum x_i$ converge to the same limiting random variable.) Thus the frequency theory gives an approximate value to conditional expectation for data of this type: the sequence of repeated experiments must be judged exchangeable.

The frequency theory does not assist with the practical problem of prediction from short sequences. Nor does it apply to other types of data. For example we might judge that the series is stationary rather than exchangeable: the assumption is weaker but limit results still apply under certain conditions. The frequency theory would be practicable if data consisted of long sequences of exchangeable random variables (the judgment of exchangeability being made informally, outside the theory); but too many important problems are not of this type.

The model theory of probability uses probability models that are "falsified" if they give very small probability to certain events. The only interpretation of probability required is that events of small probability are assumed "practically certain" not to occur. The advance over the frequency theory is that it is not necessary to explain what repeatable experiments are. The loss is that many probabilities must be assumed in order to compute the probabilities of the falsifying events, and so it is not clear which probabilities are false if one of the events occur. The interpretation of small probabilities as practically zero is not adequate to give meaning to probability. Consider for example the model that a sample of n observations is independently sampled from the normal: one of the observations is 20 standard deviations from the rest; we might conclude that the real distribution is not normal or that the sampled observations are not independent (for example the first $(n - 1)$ observations may be very highly correlated). Thus we cannot empirically test the normality unless we are sure of the independence; and assuming the independence is analogous to assuming exchangeability in de Finetti's theories.

Finally the subjective theory of probability is objectionable because probabilities are mere personal opinions: one can give a little advice; the probabilities should cohere, the set of probabilities should not combine to give unacceptable probabilities; but in the main the theory *describes* how ideally rational people act rather than *recommends* how they should act.

1.9. Infinite Axioms

Two questions arise when probabilities are defined on infinite numbers of events. These questions cannot be settled by reference to empirical facts, or by considering interpretations of probability, since in practice we do not deal with infinite numbers of events. Nevertheless it makes a considerable difference in the mathematics which choices are made.

In Kolmogorov's axioms, the axiom of countable additivity is assumed. This makes it possible to determine many useful limiting probabilities that would be unavailable if only finite additivity is assumed, but at the cost of limiting the application of probability to a subset of the family of all subsets of the line. Philosophers are reluctant to accept the axiom, but mathematicians are keen to accept it; de Finetti and others have developed a theory of finitely additive probability which differs in exotic ways from the regular theories—he will say "consider the uniform distribution on the line, carried by the rationals"; distribution functions do not determine probability distributions on the line. Here, the axiom of countable additivity is accepted as a mathematical convenience.

The second infinite axiom usually accepted is that the total probability should be one. This is inconvenient in Bayes theory because we frequently need uniform distributions on the line; countable additivity requires that total probability be infinite.