The Foundations of Statistical Inference

A Discussion

Opened by Professor L. J. Savage
at a meeting of the Joint Statistics Seminar,
Birkbeck and Imperial Colleges,
in the University of London

First published in 1962
© 1962 by G. A. Barnard and D. R. Cox
Printed in Great Britain by
Spottiswoode Ballantyne & Co Ltd
London & Colchester
Catalogue No. (Methuen) 2/5237/11
Contents

I Subjective probability and statistical practice  page 9
   L. J. Savage

II Prepared contributions
   M. S. Bartlett  36
   G. A. Barnard  39
   D. R. Cox  49
   E. S. Pearson  53
   C. A. B. Smith  58

III Discussion  62

   References  104

   Name Index  109

   Subject Index  110
Preface

When it became known that Professor L. J. Savage was visiting London in the summer of 1959 and was willing to speak on the applications of subjective probability to statistics, it was arranged that he should address the Joint Statistics Seminar of Birkbeck and Imperial Colleges. The present monograph is based on papers and discussion at that meeting which took place at Birkbeck College on July 27th and 28th.

The monograph is in three parts. Part I is a somewhat expanded form of Professor Savage's opening lecture. Part II gives five short invited contributions that had been prepared in advance of Professor Savage's lecture. A sixth contribution by Professor D. V. Lindley is not reproduced here, but has appeared in expanded form as a paper in the Proceedings of the Fourth Berkeley Symposium. The discussion recorded in Part III of the monograph is largely concerned with the issues raised in Professor Savage's lecture. In editing, the order in which the discussion took place has been slightly rearranged and one or two additional statements have been inserted.
PART I

Subjective Probability and Statistical Practice

LEONARD J. SAVAGE

1. Introduction

I am here to enlist your active participation in a movement with practical implications for statistical theory and applications at all levels, from the most elementary classroom to the most sophisticated research. Personal contact with so many competent and active statisticians in connection with issues that still seem liable to emotional misinterpretation when merely written is very auspicious. Nor could one possibly arrange better to stimulate and hear the criticisms and doubts that the subjectivistic contribution to statistics must answer.

My own attitude toward the movement has changed materially since I contributed to it a book called The Foundations of Statistics (Savage, 1954). Though this book emphasizes the merits of the concept of subjective (or personal) probability, it was not written in the anticipation of radical changes in statistical practice. The idea was, rather, that subjective probability would lead to a better justification of statistics as it was then taught and practiced, without having any urgent practical consequences. However, it has since become more and more clear that the concept of subjective probability is capable of suggesting and unifying important advances in statistical practice.

It helps to emphasize at the outset that the role of subjective probability in statistics is, in a sense, to make statistics less subjective. We all know how much the activity of one who uses statistics depends on judgement, both in the planning of experiments and in the analysis of them. For example, we are often counselled by statistical theory to choose among the many operating characteristic functions that reflect the choice of an experiment and an analysis, or the choice of an analysis alone. This choice among available operating characteristics
is recognized almost universally to be a subjective matter, depending
on the judgement of the person, or of each person, concerned. The
theory of subjective probability shows these necessarily subjective
judgements to be far less arbitrary or free than they have heretofore
superficially seemed, and therein lies much of the value of this concept
for statistics.

I know little of the early history of subjective probability, though
early references could surely be found. The earliest clear statement of
the concept of subjective probability known to me is due to Borel
(1924). A little more recent but more thoroughgoing was the formulat-
on of Ramsey (1931), which is in no way obsolete. Ramsey was fol-
lowed closely and independently by de Finetti (1937, 1949, 1958), who
continues to explore the foundations of probability with extraordinary
competence and thoroughness. Adequate formulation was also given
by Koopman (1940a, b; 1941). These pioneers in the concept of sub-
jective probability did not write as statisticians, and the application of
the concept to statistics raises many questions outside the scope of
their work.

There are doubtless many relatively early publications discussing
the application of subjective probability to statistics, for example,
those by Molina (1931) and Fry (1934). But the idea was much dis-
couraged for several decades. The book by I. J. Good (1950) is a land-
mark in its statistical reawakening; see also Good (1952). In recent
years, several qualified statisticians have been interested in more or
less explicit applications of subjective probability (Anscombe, 1958;
Hodges and Lehmann, 1952; Lindley, 1956; Wallace, 1959; Whittle,
1958). A most interesting textbook on statistics for students of business
that wholeheartedly embraces subjective probability has recently been

Though Sir Harold Jeffreys has not been a subjectivist, his work,
exemplified by two books (Jeffreys, 1948, 1957), in common with that
of subjectivists, makes serious use of Bayes's theorem. Moreover,
Jeffreys's belief in the existence of canonical initial distributions (for
certain situations) does not keep him from studying also arbitrary
initial distributions, which are just what subjectivists need. Anyone
wishing to explore subjective probability will find many valuable
lessons in the two books just mentioned that do not yet seem to be
available elsewhere, and much of what I shall say here is taken directly from Jeffreys.

Today's talk is not axiomatic, and mathematical rigour is not one of its objectives, though I shall of course not willfully make mathematical mistakes. It is mainly through examples that I hope to leave you more interested in, and more sanguine about, applications of subjective probability to statistical inference.

By inference I mean roughly how we find things out – whether with a view to using the new knowledge as a basis for explicit action or not – and how it comes to pass that we often acquire practically identical opinions in the light of evidence. Statistical inference is not the whole of inference but a special kind. The typical inference of the detective, historian, or conjecturing mathematician and the clever inferences of science are not statistical inferences. Still, it is hard to draw the line, and there seems to be nothing to lose and much to gain by keeping the more general concept in mind, provided we remember to give special attention to those aspects of inference that seem especially appropriate to the working statistician.

2. Subjective probability

Subjective probability refers to the opinion of a person as reflected by his real or potential behaviour. This person is idealized; unlike you and me, he never makes mistakes, never gives thirteen pence for a shilling, or makes such a combination of bets that he is sure to lose no matter what happens. Though we are not quite like that person, we wish we were, and it will be important for you to try to put yourself mentally in his place. To facilitate this identification, Good (1950) called him ‘you’, and I shall for the moment call him ‘thou’. The probability that refers to thee is basically a probability measure in the usual sense of modern mathematics. It is a function \( \Pr \) that assigns a real number to each of a reasonably large class of events \( A, B, \ldots \), including a universal event \( S \), in such a way that if \( A \) and \( B \) have nothing in common,

\[
\Pr (A \text{ or } B) = \Pr (A) + \Pr (B),
\]

\[
\Pr (A) \geq 0,
\]

\[
\Pr (S) = 1.
\]
The extra-mathematical thing, the thing of crucial importance, is that
Pr is entirely determined in a certain way by potential behaviour.
Specifically, Pr(A) is such that

\[ \frac{Pr(A)}{Pr(\text{not } A)} = \frac{Pr(A)}{1 - Pr(A)} \]

is the odds that thou wouldst barely be willing to offer for A against
not A.

The definition just given will not be altogether unfamiliar to you,
and you will see what it is driving at. Roughly speaking, it can be
shown that such a probability structure Pr, and one only, exists for
every person who behaves coherently in that he is not prepared to
make a combination of bets that is sure to lose (de Finetti, 1937,
pp. 6–9; Savage, 1954). This assertion is not quite correct in that a
coherent person may justifiably vary his odds with the size of the bet.
To use this definition effectively, you should try to think in terms of
bets that are rather small but worth considering. The great advantage
of this definition over more rigorous ones like the one borrowed from
de Finetti (1937, pp. 4–5) for use in my book (Savage, 1954) is that the
one in terms of odds seems much easier to apply introspectively.
Without insisting on an axiomatic exploration today, please believe
that there is considerable rationale behind the concept of subjective
probability in the various references cited, make an introspective
effort to apply the concept to yourself, and see with me what it leads
to in a few statistical examples.

The concept of subjective probability has serious defects. These can
be instructively appraised by exploring the close analogy between the
odds that you would offer on an event, and the price at which you
would buy or sell some valuable object. Both concepts are afflicted
with vagueness and temptation to dishonesty. It might be hard for you
to fix with precision the odds that you would offer that a particular keg
of nails meets some specified industrial standard or that the moon is
covered thickly with fine dust; in the same way, it might be hard for
you to specify the price at which you would sell your automobile or
buy a specific piece of information about the aurora borealis. Again,
if the facts about the nails or the moon should be disclosed to you, it
may become even harder to say honestly what you would have bet;
similarly, once highly satisfactory prices have been offered to you, it
is even harder than before to say honestly what prices would have been just satisfactory.

These difficulties are real, but they must not be allowed to frighten us out of trying to use the concepts at all. We can, if we try, do quite a bit with them as they are, and we can mitigate some of their inadequacies by using common sense and ingenuity. There is the hope that distinct improvements will be made on such concepts some day, but it seems to me that they are, each in its own line, the best that we have today.

Most people tacitly accept, and I think justifiably, that the concept of (equilibrium) price cannot be altogether escaped by anyone who would think of his own or other people’s economic behaviour. But statistical theory has for several decades been largely dedicated to trying, futilely, I would say, to escape altogether from the concept of acceptable odds, or subjective probability, at least where the analysis of data is concerned. In so far as we want to arrive at opinions on the basis of data, it seems inescapable that we should use, together with the data, the opinions that we had before it was gathered. And I believe that ‘opinion’, when analysed, is coterminous with ‘odds’. We have had a slogan about letting the data speak for themselves, but when they do, they tell us only how to modify our opinions, not what opinion is justifiable. If my statement of these general principles is somewhat dogmatic and abrupt, it is because I trust that examples will show you better than abstract arguments how the ingenious attempts to build a statistical theory without subjective probability have fallen short and how the concept of subjective probability leads to substantial improvements.

To my own mind, one of the most striking symptoms of the inadequacy of statistical theory without subjective probability is the lack of unity that such theory has had. I speak not only of such schisms as that between the adherents of R. A. Fisher and those of Neyman and Pearson, but also of the ununified, or opportunistic, structure of the theories proposed by both of these two schools.

For example, according to the Neyman-Pearson school there are many different virtues that a system of confidence intervals might have. A user of statistics is supposed to try to achieve as many of these as possible, and then to choose among them when there is conflict.
Typically, this theory leaves the user of statistics with a wide range of choices. He is supposed to survey the available operating characteristics and choose the one that he likes best among those that are at all reasonable (that is, admissible or nearly so). This choice is frankly subjective, but for certain historical reasons, the dominant school of statisticians has not seen that the idea of subjective probability makes the choice easier and more systematic than it would at first sight seem to be.

The recommendations of Fisher (1956) present a different kind of fragmentation. He counsels us to do one thing when we know 'nothing', a state of knowledge difficult, really impossible, to define, and to do another when initial probabilities, in some nonsubjectivist sense, permit application of Bayes's theorem. Even when we know 'nothing', Fisher does not present us with a unified method, but tells us to use fiducial probability, when the 'fiducial argument', the meaning of which is still a little hazy, applies, and to do something quite different when it does not apply.

The fragmentation presented by these theories suggests, though it does not prove, that something important is missing, and the application of subjective probability does make for unity. The subjectivist too will of course exploit the advantages of special situations, but he sees these not isolated as islands but as interesting regions merging with the rest of the mainland.

It is sometimes alleged as a criticism of the concept of subjective probability that science must be objective, that right reason must lead from given evidence to one and only one right conclusion. This line of thinking does not appear to be valid, fruitful or practical; see Bridgman (1940). Though it has, I think it fair to say, been expressed in some form by adherents of the school of Neyman and Pearson, it is certainly not essential to that school, which for the most part freely admits that the reaction to an experiment depends on subjective factors, like choice among operating characteristics. It seems to me that such 'objectivity' as we enjoy in practice stems from the tendency of diverse opinions to converge toward one another under the weight of evidence. It must be emphasized that this convergence is non-uniform, so it cannot be pretended that only a small sphere of opinions is left open after the accumulation of weighty evidence. In particular,
I believe R. A. Fisher is mistaken when he argues thus: 'This property of increasingly large samples has been sometimes put forward as a reason for accepting the postulate of knowledge a priori. It appears, however, more natural to infer from it that it should be possible to draw valid conclusions from the data alone, and without a priori assumptions' (Fisher, 1934, p. 287).

It has sometimes been contended that there are two different kinds of statistical theory, one appropriate to economic contexts and another to pure science; see, for example, Fisher (1955). In my own opinion, this dualistic view is incorrect. At any rate, the applications of subjective probability to be discussed today are equally available to economic and scientific applications of statistics. Except that subjective probability is defined as an economic concept, in terms of choice among gambles, this talk will scarcely refer explicitly to decision, loss, or other economic concepts. When subjective probability is taken seriously these other concepts, though they remain important, become relatively uninteresting, because in principle the solution of every decision problem is simply to maximize expected income with respect to the subjective probability that applies at the moment of making the decision. This leaves us free, at least in today's talk, to emphasize the calculation of posterior, or final, probabilities, though in more advanced applications the quality of certain approximations would have to be judged partly on the basis of possible losses.

3. Bayes's theorem and the likelihood principle

Write Bayes's theorem somewhat informally thus:

$$\Pr(\lambda | x) \propto \Pr(x | \lambda) \Pr(\lambda).$$

In words, the probability that the unknown parameter has the value \( \lambda \) given the datum \( x \) is proportional to the product of the probability of observing \( x \) given \( \lambda \) multiplied by the initial probability of \( \lambda \). 'Proportionality' here means proportionality in \( \lambda \) regarding \( x \) as fixed; if a different datum \( x' \) were observed, there would typically be a different constant of proportionality. As is familiar to you all, the various probabilities referred to can also be interpreted as probability densities where necessary; a more general formula can of course be written, but this one seems to have mnemonic value.
It is helpful to notice that in usual applications of statistics the probability of \( x \) given \( \lambda \) tends to have a quality that might cautiously be called 'objectivity'. For example, everyone concerned with the experiment may be agreed that \( x \) is normally distributed around \( \lambda \) with unit variance, or that \( x \) is a Poisson variable with mean \( \lambda \), or that \( x \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \), the parameter \( \lambda \) in the last example consisting of the pair \((\mu, \sigma^2)\). These probabilities are not really objective; they are expressions of opinions. In practice, they are seldom even taken seriously as realistic opinions by experienced statisticians, but are regarded as rather rough practical ways to get on with the problem until more realistic assumptions prove necessary. For example, if \( x \) is supposed to be 15 normally distributed physical measurements, there will probably be serious talk by all concerned about a spurious, or outlying, observation if the largest reading is separated from the second largest by as much as the second largest is from the least. This means that all concerned have (and therefore presumably always had latent) some doubts about the rigorous normality of the sample. Similarly, a sample in which the magnitude of the readings is nearly perfectly correlated with the order in which they were taken is sure to raise eyebrows. In short, simple models do not often fully represent our opinions about the possible outcomes of an experiment. Useful though such models are, the danger of accepting them literally cannot be overemphasized.

In contrast with the conditional probability of \( x \) given \( \lambda \), the probability of \( \lambda \) itself is usually conspicuously personal and vague. In principle, anyone can, by asking himself how he would bet, elicit his own subjective probability distribution for the velocity of neon light in beer. But no one is really prepared to do so with much precision, and still less is close agreement from person to person to be expected. It is largely because of these difficulties that Bayes's theorem has so long been regarded as useless by most modern statisticians. Not only are these difficulties often surmountable, but, in my experience, whenever an experiment justifies a conclusion the justification can always be given in terms of Bayes's theorem.

In view of (1), if the initial probability of \( \lambda \) is ill-defined or not agreed upon, the same must be true of the final probability of \( \lambda \), that is, the probability of \( \lambda \) given \( x \). Nonetheless, there is an important
practical sense in which the probability of $\lambda$ given $x$ may be much more precise and much better agreed upon than the initial probability of $\lambda$, as will be illustrated in the next section.

According to Bayes’s theorem, $Pr(x|\lambda)$, considered as a function of $\lambda$, constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and more precisely, if $y$ is the datum of some other experiment, and if it happens that $Pr(x|\lambda)$ and $Pr(y|\lambda)$ are proportional functions of $\lambda$ (that is, constant multiples of each other), then each of the two data $x$ and $y$ have exactly the same thing to say about the values of $\lambda$. For example, the probability of seeing 6 red-eyed flies in a randomly drawn sample of 100 is proportional to $\lambda^6(1-\lambda)^{94}$, where $\lambda$ is the frequency of red-eyed flies in the population, whether the experiment consisted in counting the number of red-eyed flies in a random sample of 100, or of sampling flies at random until 6 with red eyes are observed, or countless other sequential variations of these experiments. I, and others, call this important principle the likelihood principle. The function $Pr(x|\lambda)$—rather this function together with all others that result from it by multiplication by a positive constant—is called the likelihood.

The likelihood principle flows directly from Bayes’s theorem and the concept of subjective probability, and it is to my mind a good example of the fertility of these ideas. The principle was, however, first emphasized to statisticians by Barnard (1947b) and Fisher (1956), on other grounds. It seems to command more and more assent the more you think about it, criticize it, and seek counterexamples against it.

The likelihood principle is, however, in conflict with many historically important concepts of statistics. For example, whether a test (or an estimate) is unbiased depends not on the likelihood alone, but rather on $Pr(x|\lambda)$ considered as a function of $x$ as well as a function of $\lambda$. Similarly with the concepts of significance or confidence level. For instance, it has been widely believed that the import of such a datum as 6 red-eyed flies out of 100 depends on whether the experiment was designed to observe 100 flies or designed to observe 6 red-eyed flies. An estimate unbiased for either of these experiments is biased for the other, and there is a considerable literature on unbiased estimates for sequential observation of Bernoullian data, to which I
myself have contributed (Blackwell, 1947; DeGroot, 1959; Girshick et al., 1946; Savage, 1947). In view of the likelihood principle, all of these classical statistical ideas come under new scrutiny, and must, I believe, be abandoned or seriously modified.

The principle has important implications in connection with optional stopping. Suppose the experimenter admitted that he had seen 6 red-eyed flies in 100 and had then stopped because he felt that he had thereby accumulated enough data to overthrow some popular theory that there should be about 1 per cent red-eyed flies. Does this affect the interpretation of 6 out of 100? Statistical tradition emphasizes, in connection with this question, that if the sequential properties of his experimental programme are ignored, the persistent experimenter can arrive at data that nominally reject any null hypothesis at any significance level, when the null hypothesis is in fact true. Such a rejection is therefore no real evidence against the null hypothesis. These truths are usually misinterpreted to suggest that the data of such a persistent experimenter are worthless or at least need special interpretation; see, for example, Anscombe (1954), Feller (1940), Robbins (1952). The likelihood principle, however, affirms that the experimenter’s intention to persist does not change the import of his experience. The true moral of the facts about optional stopping is that significance level is not really a good guide to ‘level of significance’ in the sense of ‘degree of import’, for the degree of import does depend on the likelihood alone, a theme to which I must return later in the lecture.

There is a class of actuarial problems of considerable theoretical, and perhaps also practical, interest that promises to be greatly simplified by systematic application of the likelihood principle. Here is an example. Children come into a clinic for observation at various time intervals after the onset of a serious disease and remain under observation until they are either withdrawn from the clinic, say by arbitrary action of their parents, or until they die, or until they survive for five years after the onset of the disease.

It is desired to estimate the probability of surviving for five years. In practice this is messy data, because there is a more than justifiable suspicion that the moment when children are brought to the clinic or withdrawn from it has some correlation with the prognosis, but even
if this practical point is set aside, the formal problem that remains is
still a complicated one. There has been considerable study of such
problems based on objectivistc statistical ideas like unbiasedness and
confidence intervals, and depending on various stochastic models of
the mechanism that brings subjects into the study and withdraws
them from it. Recent key references are the papers of Elveback (1958)
and Kaplan and Meier (1958). According to the likelihood principle,
however, these mechanisms have nothing to do with the import of the
data for the pressure of mortality and, in particular, for the proba-
bility of surviving for five years. With this hint, new progress on the
problem is to be expected and many dead ends are recognized.

To see how the likelihood principle works on a problem of the type
envisioned in the preceding paragraphs, divide the five-year interval
into short intervals, and suppose it is known that of $x_i$ children who
had survived to the beginning of the $i$th interval, $y_i$ survived through-
out that interval. If $p_i$ is the probability that a child who survives to
the $i$th interval will survive through it, then $p = \prod p_i$ is the probability
of five-year survival. Irrespective of any model that might be proposed
to account for the $x_i$, say in terms of parameters $\xi,$

$$
\Pr(x_1, \ldots, y_1, \ldots | p_1, \ldots ; \xi)
= \Pr(x_1 | \xi) \Pr(y_1 | p_1, x_1) \Pr(x_2 | x_1, \xi) \Pr(y_2 | p_2, x_2) \ldots
= f(x_1, \ldots ; \xi) \prod p_i^{y_i} (1 - p_i)^{x_i - y_i}.
$$

The function $f$ is irrelevant to the likelihood, so that only the
observed $x_i,$ not the mechanism which produced them, enters into the
likelihood.

The concept of ancillary statistic, introduced by Fisher, has been
difficult to grasp and to define precisely – see, for example, Cox
(1958a) – but the likelihood principle seems to provide the key. A
typical important instance of an ancillary statistic is this. If $x_i$ and $y_i$
are both from a random source, say independent sample pairs from a
joint normal distribution, and the problem is to study the regression
of $y$ on $x,$ then it is, everyone seems to agree, legitimate to make all
inferences as though the $x_i$ were not random and the original experi-
mental plan had been to sample the $y_i$ exactly at these $x_i.$ The $x_i$ are
here called an ancillary statistic, and the likelihood principle does
indeed make it clear that the mechanism that happened to select the
\( x_i \) is of no relevance to inferences about the regression coefficient.
Actually, it is tacit in this argument that perfect knowledge of the
distribution from which the \( x_i \) were sampled would be of no use (or,
more realistically, of little use) in drawing conclusions about the repre-
sion coefficient. This is a subjective matter and can therefore vary from
problem to problem and person to person. One might, for example,
have some strong special reason to believe that if the \( x_i \) are broadly
(or narrowly) distributed the regression coefficient is likely to be large.

The likelihood principle always supports the appealing conclusions
that have been based on ancillary statistics. But the likelihood
principle leads to the simplifications almost automatically, whereas
ancillary statistics are discovered only by ingenuity and insight. An
important reason for this is that one problem can have many ancillary
statistics on a par with each other. For a striking, if academic, example,
suppose \( x \) and \( y \) are normal about 0 with variance 1 and correlation
\( \rho \). Then \( x \) and \( y \) are each by themselves irrelevant to \( \rho \), and each is an
ancillary statistic for the total observation \((x, y)\) by any criterion known
to me. Inclusion of several pairs \((x_i, y_i)\), rather than one only, makes
no essential difference.

4. Precise measurement

One of the most interesting and satisfying applications of subjective
probability is that of precise measurement. This is the kind of measure-
ment we have when the data are so incisive as to overwhelm the initial
opinion, thus bringing a great variety of realistic initial opinions to
practically the same conclusion. Put yourself – not an over-idealized
person – into a certain statistical situation of textbook simplicity. The
one I propose is humble but, for me, instructive. Whoever finds it
frivolous can supply something graver: the weight of a nougat or of
the earth, the melting point of a new compound, or the sex-ratio at
birth in post-war Germany – a good example, because sample fre-
quency here is so like a normal measurement of variance \( \frac{1}{2} N^{-1} \).

You are holding a potato, or some other irregular object, in your
hand and have need to know something about its weight. You can, in