D. G. MAYO AND M. KRUSE

PRINCIPLES OF INFERENCE AND THEIR CONSEQUENCES

> The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proved or disproven . . . [Edwards *et al.*, 1963, p. 193].

1 INTRODUCTION

What do data tell us about hypotheses or claims? When do data provide good evidence for or a good test of a hypothesis? These are key questions for a philosophical account of evidence and inference, and in answering them, philosophers of science have often appealed to formal accounts of probabilistic and statistical inference. In so doing, it is obvious that the answer will depend on the principles of inference embodied in one or another statistical account. If inference is by way of Bayes' theorem, then two data sets license different inferences only by registering differently in the Bayesian algorithm. If inference is by way of error statistical methods (e.g., Neyman and Pearson methods), as are commonly used in applications of statistics in science, then two data sets license different inferences or hypotheses if they register differences in the error probabilistic properties of the methods.

The principles embodied in Bayesian as opposed to error statistical methods lead to conflicting appraisals of the evidential import of data, and it is this conflict that is the pivot point around which the main disputes in the philosophy of statistics revolve. The differences between the consequences of these conflicting principles, we propose, are sufficiently serious as to justify supposing that one "cannot be just a little bit Bayesian" [Mayo, 1996], at least when it comes to a philosophical account of inference, but rather must choose between fundamentally incompatible packages of evidence, inference, and testing. In the remainder of this section we will sketch the set of issues that seems to us to serve most powerfully to bring out this incompatibility.

EXAMPLE 1 (ESP Cards). The conflict shows up most clearly with respect to the features of the *data generation process* that are regarded as relevant for assessing evidence. To jump right into the crux of the matter, we can consider a familiar type of example: To test a subject's ability, say, to predict draws from a deck of five ESP cards, he must demonstrate a success rate that would be very improbable if he were merely guessing. Supposing that after a long series of trials, our subject

D. Corfield and J. Williamson (eds.), Foundations of Bayesianism, 381–403.