1

Introduction

> We statisticians, with our specific concern for uncertainty, are even more liable than other practical men to encounter philosophy, whether we like it or not.
>
> (Savage 1981a)

1.1 Controversies

Statistics is a mature field but within it remain important controversies. These controversies stem from profoundly different perspectives on the meaning of learning and scientific inquiry, and often result in widely different ways of interpreting the same empirical observations.

For example, a controversy that is still very much alive involves how to evaluate the reliability of a prediction or guess. This is, of course, a fundamental issue for statistics, and has implications across a variety of practical activities. Many are captured by a case study on the evaluation of evidence from clinical trials (Ware 1989). We introduce the controversy with an example. You have to guess a secret number. You know it is an integer. You can perform an experiment that would yield either the number before it or the number after it, with equal probability. You know there is no ambiguity about the experimental result or about the experimental answer. You perform this type of experiment twice and get numbers 41 and 43. What is the secret number? Easy, it is 42. Now, how good an answer do you think this is? Are you tempted to say "It is a perfect answer, the secret number has to be 42"? It turns

Decision Theory: Principles and Approaches G. Parmigiani, L. Y. T. Inoue

out that not all statisticians think this is so easy. There are at least two opposed perspectives on how to go about figuring out how good our answer is:

> **Judge an answer by what it says**. *Compare the answer to other possible answers, in the light of the experimental evidence you have collected, which can now be taken as a given.*

versus

> **Judge an answer by how it was obtained**. *Specify the rule that led you to give the answer you gave. Compare your answer to the answers that your rule would have produced when faced with all possible alternative experimental results. How far your rule is from the truth in this collection of hypothetical answers will inform you about how good your rule is. Indirectly, this will tell you how well you can trust your specific answer.*

Let us go back to the secret number. From the first perspective you would compare the answer "42" to all possible alternative answers, realize that it is the only answer that is not ruled out by the observed data, and conclude that there is no ambiguity about the answer being right. From the second perspective, you ask how the answer was obtained. Let us consider a reasonable recipe for producing answers: average the two experimental results. This approach gets the correct answer half the time (when the two experimental results differ) and is 1 unit off the remainder of the time (when the two experiments yield the same number). Most measures of error that consider this entire collection of potential outcomes will result in a conclusion that will attribute some uncertainty to your reported answer. This is in sharp contrast with the conclusion reached following from the first perspective. For example, the standard error of your answer is $1/\sqrt{2}$. By this principle, you would write a paper reporting your discovery that "the secret number is 42 (s.e. 0.7)" irrespective of whether your data are 41 and 43, or 43 and 43. You can think of other recipes, but if they are to give you a single guess, they are all prone to making mistakes when the two experimental results are the same, and so the story will have the same flavor.

The reasons why this controversy exists are complicated and fascinating. When things are not as clear cut as in our example, and multiple answers are compatible with the experimental evidence, the first perspective requires weighing them in some way—a step that often involves judgment calls. On the other hand the second perspective only requires knowing the probabilities involved in describing how the experiments relate to the secret number. For this reason, the second approach is perceived by many to be more objective, and more appropriate for scientific inquiry. Objectivity, its essence, worthiness, and achievability, have been among the most divisive issues in statistics. In an extreme simplification the controversy can be captured by two views of probability:

> **Probability lives in the world**. *Probability is a physical property like mass or wavelength. We can use it to describe stochastic experimental*

> *mechanisms, generally repeatable ones, like the assignments of experimental units to different conditions, or the measurement error of a device. These are the only sorts of probabilistic considerations that should enter scientific investigations.*

versus

> **Probability lives in the mind**. *Probability, like most conceptual constructs in science, lives in the context of the system of values and theories of an individual scientist. There is no reason why its use should be restricted to repeatable physical events. Probability can for example be applied to scientific hypotheses, or the prediction of one-time events.*

Ramsey (1926) prefaced his fundamental paper on subjective probability with a quote from poet William Blake: "Truth can never be told so as to be understood, and not be believed."

These attitudes define a coordinate in the space of statisticians' personal philosophies and opinions, just like the poles of the previous controversy did. These two coordinates are not the same. For example, there are approaches to the secret number problem that give different answers depending on whether data are 41 and 43, or 43 and 43, but do not make use of "subjective" probability to weigh alternative answers. Conversely, it is common practice to evaluate answers obtained from subjective approaches, by considering how the same approaches would have fared in other experiments.

A key aspect that both these dimensions have in common is the use of a stochastic model as the basis for learning from data. In the secret number story, for example, the starting point was that the experimental results would fall to the left or right of the secret number *with equal probability*. The origin of the role of probability in interpreting experimental results is sampling. The archetype of many statistical theories is that experimental units are sampled from a larger population, and the goal of statistical inference is to draw conclusions about the whole population. A *statistical model* describes the stochastic mechanism based on which samples are selected from the population. Sometimes this is literally the case, but more often samples and populations are only metaphors to guide the construction of statistical procedures. While this has been the model of operation postulated in most statistical theory, in practical applications it is only one pole of yet another important controversy:

> **Learning requires models**. *To rigorously interpret data we need to understand and specify the stochastic mechanism that generated them. The archetype of statistical inference is the sample-population situation.*

versus

> **Learning requires algorithms**. *To efficiently learn from data, it is critical to have practical tools for exploring, summarizing, visualizing,*

clustering, classifying. These tools can be built with or without explicit consideration of a stochastic data-generating model.

The model-based approach has ancient roots. One of the relatively recent landmarks is Fisher's definition of the likelihood function (Fisher 1925). The algorithmic approach also goes back a long way in history: for example, most measures of dependence, such as the correlation coefficient, were born as descriptive, not inferential tools (Galton 1888). The increasing size and complexity of data, and the interface with computing, have stimulated much exploratory data analysis (Tukey 1977, Chambers *et al.* 1983) and statistical work at the interface with artificial intelligence (Nakhaeizadeh and Taylor 1997, Hastie *et al.* 2003). This controversy is well summarized in an article by Breiman (2001).

The premise of this book is that it is useful to think about these controversies, as well as others that are more technical in statistics, from first principles. The principles we will bring to bear are principles of rationality in action. Of course, this idea is in itself controversial. With this regard, the views of many statisticians distribute along another important dimension of controversy:

> **Statisticians produce knowledge**. The scope of statistics is to rigorously interpret experimental results, and present experimental evidence in an unbiased way to scientists, policy makers, the public, or whoever may be in charge of drawing conclusions or making decisions.

versus

> **Statisticians produce solutions to problems**. Understanding data requires placing them in the context of scientific theories, which allow us to sort important from ancillary information. One cannot answer the question "what is important?" without first considering the question "important for what?"

Naturally, producing knowledge helps solving problems, so these two positions are not in contrast from this standpoint. The controversy is on the extent to which the goals of an experiment should affect the learning approaches, and more broadly whether they should be part of our definition of learning.

The best known incarnation of this controversy is the debate between Fisher and Neyman about the meaning of hypothesis tests (Fienberg 1992). The Neyman–Fisher controversy is broader, but one of the key divides is that the Neyman and Pearson theory of hypothesis testing considers both the hypothesis of interest and at least one alternative, and involves an explicit quantification of the consequences of rejecting or accepting the hypothesis based on the data: the type I and type II errors. Ultimately, Neyman and Pearson's theory of hypothesis testing will be one of the key elements in the development of formal approaches to rationality-based statistical analysis. On the other hand Fisher's theory of significance test does not require considering an alternative and incarnates a view of science in which hypotheses represent working

approximations to natural laws, that serve to guide experimentation, until they are refuted with sufficient strength that new theories evolve.

A simple example challenges theories of inference that are based solely on the evidence provided by observation, regardless of scope and context of the theory. Let x_1 and x_2 be two Bernoulli trials. Suppose the experimenter's probabilities are such that $P(x_1 = 0) = P(x_2 = 0) = 0.5$ and $P(x_1 + x_2 = 0) = 0.05$. Then, $P(x_1 + x_2 = 1) = 0.9$ and $P(x_1 + x_2 = 2) = 0.05$. Let e be the new evidence that $x_1 = 1$, let h_1 be the hypothesis that $x_1 + x_2 = 2$, and h_2 be the hypothesis that $x_2 = 1$. Given e, the two hypotheses are equivalent. Yet, probability-wise, h_1 is corroborated by the data, whereas h_2 is not. So if one is to consider the change in probability as a measure of support for a theory, one would be left with either an inconsistent measure of evidence, or the need to defend the position that the two hypotheses are in some sense different even when faced with evidence that proves that they are the same. This and other similar examples seriously question the idea that inductive practice can be adequately represented by probabilities alone, without relation to their rational use in action.

There can be disagreements of principle about whether consideration of consequences and beliefs belongs to scientific inquiry. In reality, though, it is our observation that the vast majority of statistical inference approaches have an implicit or explicit set of goals and values that guide the various steps of the construction. When making a decision as simple as summarizing a set of numbers by their median (as opposed to, say, their mean) one is making judgments about the relative importance of the possible oversimplifications involved. These could be made formal, and in fact there are decision problems for which each of the two summaries is optimal. Our view is that scientific discussion is more productive when goals are laid out in the open, and perhaps formalized, than when they are hidden or unappreciated. As the old saying goes, "there are two types of statisticians: those who know what decision problem they are solving and those who don't."

Despite the draconian simplifications we have made in defining the dimensions along which these four controversies unfold, one would be hard pressed to find two statisticians that live on the same point in this four-dimensional space. One of the goals of this book is to help students find their own spot in a way that reflects their personal intellectual values, and serves them best in approaching the theoretical and applied problems that are important to them.

We definitely lean on the side of "judging answers by what they say" and believing that "probabilities live in the mind." We may be some distance away along the models versus algorithm dimension—at least judging by our approaches in applications. But we are both enthusiastic about the value of thinking about rationality as a guide, though sometimes admittedly a rough guide, to science, policy, and individual action. This guidance comes at two levels: it tells us how to formally connect the tools of an analysis with the goals of that analysis; and it tells us how to use rationality-based criteria to evaluate alternative statistical tools, approaches, and philosophies.

Overall, our book is an invitation to Bayesian decision-theoretic ideas. While we do not think they necessarily provide a solution to every statistical problem, we

find much to think about in this comment from Herman Chernoff (from a personal communication to Martin McIntosh):

> Frankly, I am not a Bayesian. I go under the following principle. If you don't understand a problem from a Bayesian decision theory point of view, you don't understand the problem and trying to solve it is like shooting at a target in the dark. Once you understand the problem, it is not necessary to attack it from a Bayesian point of view. Bayesian methods always had the difficulty that our approximations to our subjective beliefs could carry a lot more information than we thought or felt willing to assume.

1.2 A guided tour of decision theory

We think of this book as a "tour guide" to the key ideas in decision theory. The book grew out of graduate courses where we selected a set of exciting papers and book chapters, and developed a self-contained lecture around each one. We make no attempt at being comprehensive: our goal is to give you a tour that conveys an overall view of the fields and its controversies, and whets your appetite for more. Like the small pictures of great paintings that are distributed on leaflets at the entrance of a museum, our chapters may do little justice to the masterpiece but will hopefully entice you to enter, and could guide you to the good places.

As you read it, keep in mind this thought from R. A. Fisher:

> The History of Science has suffered greatly from the use by teachers of second-hand material, and the consequent obliteration of the circumstances and the intellectual atmosphere in which the great discoveries of the past were made. A first-hand study is always useful, and often . . . full of surprises. (Fisher 1965)

Our tour includes three parts: foundations (axioms of rationality); optimal data analysis (statistical decision theory); and optimal experimental design.

Coherence. We start with de Finetti's "Dutch Book Theorem" (de Finetti 1937) which provides a justification for the axioms of probability that is based on a simple and appealing rationality requirement called coherence. This work is the mathematical foundation of the "probabilities live in the mind" perspective. One of the implications is that new information is merged with the old via Bayes' formula, which gets promoted to the role of a universal inference rule—or Bayesian inference.

Utility. We introduce the axiomatic theory of utility, a theory on how to choose among actions whose consequences are uncertain. A rational decision maker proceeds by assigning numerical utilities to consequences, and scoring actions by their expected utility. We first visit the birthplace of quantitative utility: Daniel Bernoulli's St. Petersburg paradox (Bernoulli 1738). We then present in detail von Neumann and

Morgenstern's utility theory (von Neumann and Morgenstern 1944) and look at a criticism by Allais (1953).

Utility in action. We make a quick detour to take a look at practical matters of implementation of rational decision making in applied situations, and talk about how to measure the utility of money (Pratt 1964) and the utility of being in good health. For applications in health, we examine a general article (Torrance *et al*. 1972) and a medical article that has pioneered the utility approach in health, and set the standard for many similar analyses (McNeil *et al*. 1981).

Ramsey and Savage. We give a brief digest of the beautiful and imposing axiomatic system developed by Savage. We begin by tracing its roots to the work of Ramsey (1931) and then cover Chapters 2, 3, and 5 from Savage's *Foundations of statistics* (Savage 1954). Savage's theory integrates the coherence story with the utility story, to create a more general theory of individual decision making. When applied to statistical practice, this theory is the foundation of the "statisticians find solutions to problems" perspective. The general solution is to maximize expected utility, and expectations are computed by assigning personal probabilities to all unknowns. A corollary is that "answers are judged by what they say."

State independence. Savage's theory relies on the ability to separate judgment of probability from judgment of utility in evaluating the worthiness of actions. Here we study an alternative axiomatic justification of the use of subjective expected utility in decision making, due to Anscombe and Anmann (1963). Their theory highlights very nicely the conditions for this separation to take place. This is the last chapter on foundations.

Decision functions. We visit the birthplace of statistical decision theory: Wald's definition of a general statistical decision function (Wald 1949). Wald proposed a unifying framework for much of the existing statistical theory, based on treating statistical inference as a special case of game theory, in which the decision maker faces nature in a zero-sum game. This leads to maximizing the smallest utility, rather than a subjective expectation of utility. The contrast of these two perspectives will continue through the next two chapters.

Admissibility. Admissibility is the most basic and influential rationality requirement of Wald's classical statistical decision theory. A nice surprise for Savage's fans is that maximizing expected utility is a safe way, and often, at least approximately, the only way, to build admissible statistical decision rules. Nice. In this chapter we also reinterpret one of the milestones of statistical theory, the Neyman–Pearson lemma (Neyman and Pearson 1933), in the light of the far-reaching theory this lemma sparked.

Shrinkage. The second major surprise from the study of admissibility is the fact that $\bar{x}$—the motherhood and apple pie of the statistical world—is inadmissible in estimating the mean of a multidimensional normal vector of observations. Stein (1955) was the first to realize this. We explore some of the important research directions

that stemmed from Stein's paper, including shrinkage estimation, empirical Bayes estimation, and hierarchical modeling.

Scoring rules. We change our focus to prediction and explore the implications of holding forecasters accountable for their predictions. We study the incentive systems that must be set in place for the forecasters to reveal their information/beliefs rather than using them to game the system. This leads to the study of proper scoring rules (Brier 1950). We also define and investigate calibration and refinement of forecasters.

Choosing models We try to understand whether statistical decision theory can be applied successfully to the much more elusive tasks of constructing and assessing statistical models. The jury is still out. On this puzzling note we close our tour of statistical decision theory and move to experimental design.

Dynamic programming. We describe a general approach for making decisions dynamically, so that we can both learn from accruing knowledge and plan ahead to account for how present decisions will affect future decisions and future knowledge. This approach, called dynamic programming, was developed by Bellman (1957). We will try to understand why the problem is so hard (the "curse of dimensionality").

Changes in utility as information. In decision theory, the value of the information carried by a data set depends on what we intend to do with the data once we have collected them. We use decision trees to quantify this value (DeGroot 1984). We also explore in more detail a specific way of measuring the information in a data set, which tries to capture "generic learning" rather than specific usefulness in a given problem (Lindley 1956).

Sample size. We finally come to terms with the single most common decision statisticians make in their daily activities: how big should a data set be? We try to understand how all the machinery we have been setting in place can help us and give some examples. Our discussion is based on the first complete formalization of Bayesian decision-theoretic approaches to sample size determination (Raiffa and Schlaifer 1961).

Stopping. Lastly, we apply dynamic programming to sequential data collection, where we have the option to stop an experiment after each observation. We discuss the stopping rule principle, which states that within the expected utility paradigm, the rule used to arrive at the decision to stop at a certain stage is not informative about parameters controlling the data-generating mechanism. We also study whether it is possible to design stopping rules that will stop experimentation only when one's favorite conclusion is reached.

A terrific preparatory reading for this book is Lindley (2000) who lays out the philosophy of Bayesian statistics in simple, concise, and compelling terms. As you progress through the book you will find, generally in each chapter's preamble, alternative texts that dwell on individual topics in greater depth than we do. Some are also listed next. A large number of textbooks overlap with ours and we make no attempt at being comprehensive. An early treatment of statistical decision theory is Raiffa and Schlaifer

(1961), a text that contributed enormously to defining practical Bayesian statistics and decision making in the earliest days of the field. Their book was exploring new territory on almost every page and, even in describing the simplest practical ideas, is full of deep insight. Ferguson (1967) is one of the early influential texts on statistical decision theory, Bayes and frequentist. DeGroot (1970) has a more restricted coverage of Part Two (no admissibility) but a more extensive discussion of Part Three and an in-depth discussion of foundations, which gives a quite independent treatment of the material compared to the classical papers discussed in our book. A mainstay statistical decision theory book is Berger (1985) which covers topics throughout our tour. Several statistics books have good chapters on decision-theoretic topics. Excellent examples are Schervish (1995) and Robert (1994), both very rigorous and rich in insightful examples. Bernardo and Smith (1994) is also rich in foundational discussions presented in the context of both statistical inference and decision theory. French (1988), Smith (1987), and Bather (2000) cover decision-analytic topics very well. Kreps (1988) is an accessible and very insightful discussion of foundations, covered in good technical detail. A large number of texts in decision analysis, medical decision making, microeconomics, operations research, statistics, machine learning, and stochastic processes cover individual topics.