Part I

Statistical Modeling: Foundations and Limitations

1

Issues in the Foundations of Statistics: Probability and Statistical Models

"Son, no matter how far you travel, or how smart you get, always remember this: Someday, somewhere, a guy is going to show you a nice brand-new deck of cards on which the seal is never broken, and this guy is going to offer to bet you that the jack of spades will jump out of this deck and squirt cider in your ear. But, son, do not bet him, for as sure as you do you are going to get an ear full of cider."

— Damon Runyon[1]

ABSTRACT. *After sketching the conflict between objectivists and subjectivists on the foundations of statistics, this chapter discusses an issue facing statisticians of both schools, namely, model validation. Statistical models originate in the study of games of chance and have been successfully applied in the physical and life sciences. However, there are basic problems in applying the models to social phenomena; some of the difficulties will be pointed out. Hooke's law will be contrasted with regression models for salary discrimination, the latter being a fairly typical application in the social sciences.*

1.1 What is probability?

For a contemporary mathematician, probability is easy to define, as a countably additive set function on a σ-field, with a total mass of one. This definition, perhaps cryptic for non-mathematicians, was introduced by A. N. Kolmogorov around 1930, and has been extremely convenient for mathematical work; theorems can be stated with clarity, and proved with rigor.[2]

For applied workers, the definition is less useful; countable additivity and σ-fields are not observed in nature. The issue is of a familiar type—what objects in the world correspond to probabilities? This question divides statisticians into two camps:

  (i) the "objectivist" school, also called the "frequentists,"
  (ii) the "subjectivist" school, also called the "Bayesians," after the Reverend Thomas Bayes (England, c. 1701–61) (Bayes, 1764).

Other positions have now largely fallen into disfavor; for example, there were "fiducial" probabilities introduced by R. A. Fisher (England, 1890–1962). Fisher was one of the two great statisticians of the century; the other, Jerzy Neyman (b. Russia, 1894; d. U.S.A., 1981), turned to objectivism after a Bayesian start. Indeed, the objectivist position now seems to be the dominant one in the field, although the subjectivists are still a strong presence. Of course, the names are imperfect descriptors. Furthermore, statisticians agree amongst themselves about as well as philosophers; many shades of opinion will be represented in each school.

1.2 The objectivist position

Objectivists hold that probabilities are inherent properties of the systems being studied. For a simple example, like the toss of a coin, the idea seems quite clear at first. You toss the coin, it will land heads or tails, and the probability of heads is around 50%. A more exact value can be determined experimentally, by tossing the coin repeatedly and taking the long run relative frequency of heads. In one such experiment, John Kerrich (a South African mathematician interned by the Germans during World War II) tossed a coin 10,000 times and got 5067 heads: The relative frequency was $5067/10,000 = 50.67\%$. For an objectivist such as myself, the probability of Kerrich's coin landing heads has its own existence, separate from the data; the latter enable us to estimate the probability, or test hypotheses concerning it.

The objectivist position exposes one to certain difficulties. As Keynes said, "In the long run, we are all dead." Heraclitus (also out of context)

is even more severe: "You can't step into the same river twice." Still, the tosses of a coin, like the throws of a die and the results of other such chance processes, do exhibit remarkable statistical regularities. These regularities can be described, predicted, and analyzed by technical probability theory. Using Kolmogorov's axioms (or more primitive definitions), we can construct statistical models that correspond to empirical phenomena; although verification of the correspondence is not the easiest of tasks.

1.3 The subjectivist position

For the subjectivist, probabilities describe "degrees of belief." There are two camps within the subjectivist school, the "classical" and the "radical." For a "classical" subjectivist, like Bayes himself or Laplace—although such historical readings are quite tricky—there are objective "parameters" which are unknown and to be estimated from the data. (A parameter is a numerical characteristic of a statistical model for data—for instance, the probability of a coin landing heads; other examples will be given below.) Even before data collection, the classical subjectivist has information about the parameters, expressed in the form of a "prior probability distribution."

The crucial distinction between a classical subjectivist and an objectivist: The former will make probability statements about parameters—for example, in a certain coin-tossing experiment, there is a 25% chance that the probability of heads exceeds .67. However, objectivists usually do not find that such statements are meaningful; they view the probability of heads as an unknown constant, which either is—or is not—bigger than .67. In replications of the experiment, the probability of heads will always exceed .67, or never; 25% cannot be relevant. As a technical matter, if the parameter has a probability distribution given the data, it must have a "marginal" distribution—that is, a prior. On this point, objectivists and subjectivists agree; the hold-out was R. A. Fisher, whose fiducial probabilities come into existence only after data collection.

"Radical" subjectivists, like Bruno de Finetti or Jimmie Savage, differ from classical subjectivists and objectivists; radical subjectivists deny the very existence of unknown parameters. For such statisticians, probabilities express degrees of belief about observables. You pull a coin out of your pocket, and—Damon Runyon notwithstanding—they can assign a probability to the event that it will land heads when you toss it. The braver ones can even assign a probability to the event that you really will toss the coin. (These are "prior" probabilities, or "opinions.") Subjectivists can also "update" opinions in the light of the data; for example, if the coin is tossed ten times, landing heads six times and tails four times, what is the

chance that it will land heads on the eleventh toss? This involves computing a "conditional" probability using Kolmogorov's calculus, which applies whether the probabilities are subjective or objective.

Here is an example with a different flavor: What is the chance that a Republican will be president of the U.S. in the year 2025? For many subjectivists, this is a meaningful question, which can in principle be answered by introspection. For many objectivists, this question is beyond the scope of statistical theory. As best I can judge, however, complications will be found on both sides of the divide. Some subjectivists will not have quantifiable opinions about remote political events; likewise, there are objectivists who might develop statistical models for presidential elections, and compute probabilities on that basis.[3]

The difference between the radical and classical subjectivists rides on the distinction between parameters and observables; this distinction is made by objectivists too and is often quite helpful. (In some cases, of course, the issue may be rather subtle.) The radical subjectivist denial of parameters exposes members of this school to some rhetorical awkwardness; for example, they are required not to understand the idea of tossing a coin with an unknown probability of heads. Indeed, if they admit the coin, they will soon be stuck with all the unknown parameters that were previously banished.[4]

1.3.1 Probability and relative frequency

In ordinary language, "probabilities" are not distinguished at all sharply from empirical percentages—"relative frequencies." In statistics, the distinction may be more critical. With Kerrich's coin, the relative frequency of heads in 10,000 tosses, 50.67%, is unlikely to be the exact probability of heads; but it is unlikely to be very far off. For an example with a different texture, suppose you see the following sequence of ten heads and ten tails:

$$T\ H\ T\ H\ T\ H\ T\ H\ T\ H\ T\ H\ T\ H\ T\ H\ T\ H\ T\ H.$$

What is the probability that the next observation will be a head? In this case, relative frequency and probability are quite different.[5]

One more illustration along that line: United Airlines Flight 140 operates daily from San Francisco to Philadelphia. In 192 out of the last 365 days, Flight 140 landed on time. You are going to take this flight tomorrow. Is your probability of landing on time given by 192/365? For a radical subjectivist, the question is clear; not so for an objectivist or a classical subjectivist. Whatever the question really means, 192/365 is the wrong answer—if you are flying on the Friday before Christmas. This is Fisher's "relevant subset" issue; and he seems to have been anticipated

by von Mises. Of course, if you pick a day at random from the data set, the chance of getting one with an on-time landing is indeed 192/365; that would not be controversial. The difficulties come with (i) extrapolation and (ii) judging the exchangeability of the data, in a useful Bayesian phrase. Probability is a subtler idea than relative frequency.[6]

1.3.2 Labels do not settle the issue

Objectivists sometimes argue that they have the advantage, because science is objective. This is not serious; "objectivist" statistical analysis must often rely on judgment and experience: Subjective elements come in. Likewise, subjectivists may tell you that objectivists (i) use "prior information," and (ii) are therefore closet Bayesians. Point (i) may be granted. The issue for (ii) is how prior information enters the analysis, and whether this information can be quantified or updated the way subjectivists insist it must be. The real questions are not to be settled on the basis of labels.

1.4 A critique of the subjectivist position

The subjectivist position seems to be internally consistent, and fairly immune to logical attack from the outside. Perhaps as a result, scholars of that school have been quite energetic in pointing out the flaws in the objectivist position. From an applied perspective, however, the subjectivist position is not free of difficulties either. What are subjective degrees of belief, where do they come from, and why can they be quantified? No convincing answers have been produced. At a more practical level, a Bayesian's opinion may be of great interest to himself, and he is surely free to develop it in any way that pleases him; but why should the results carry any weight for others?

To answer the last question, Bayesians often cite theorems showing "inter-subjective agreement." Under certain circumstances, as more and more data become available, two Bayesians will come to agree: The data swamp the prior. Of course, other theorems show that the prior swamps the data, even when the size of the data set grows without bounds— particularly in complex, high-dimensional situations. (For a review, see Diaconis and Freedman 1986.) Theorems do not settle the issue, especially for those who are not Bayesians to start with.

My own experience suggests that neither decision-makers nor their statisticians do in fact have prior probabilities. A large part of Bayesian statistics is about what you would do *if* you had a prior.[7] For the rest, statisticians make up priors that are mathematically convenient or attractive. Once used, priors become familiar; therefore, they come to be accepted

as "natural" and are liable to be used again. Such priors may eventually generate their own technical literature.

1.4.1 Other arguments for the Bayesian position

Coherence. Well-known theorems, including one by Freedman and Purves (1969), show that stubborn non-Bayesian behavior has costs. Your opponents can make a "dutch book," and extract your last penny—if you are generous enough to cover all the bets needed to prove the results.[7] However, most of us don't bet at all; even the professionals bet on relatively few events. Thus, coherence has little practical relevance. (Its rhetorical power is undeniable—who wants to be incoherent?)

Rationality. It is often urged that to be rational is to be Bayesian. Indeed, there are elaborate axiom systems about preference orderings, acts, consequences, and states of nature, whose conclusion is—that you are a Bayesian. The empirical evidence shows, fairly clearly, that those axioms do not describe human behavior at all well. The theory is not descriptive; people do not have stable, coherent prior probabilities.

Now the argument shifts to the "normative": If you were rational, you would obey the axioms and be a Bayesian. This, however, assumes what must be proved. Why would a rational person obey those axioms? The axioms represent decision problems in schematic and highly stylized ways. Therefore, as I see it, the theory addresses only limited aspects of rationality. Some Bayesians have tried to win this argument on the cheap: To be rational is, by definition, to obey their axioms. (Objectivists do not always stay on the rhetorical high road either.)

Detailed examination of the flaws in the normative argument is a complicated task, beyond the scope of the present article. In brief, my position is this. Many of the axioms, on their own, have considerable normative force. For example, if I am found to be in violation of the "sure thing principle," I would probably reconsider.[9] On the other hand, taken as a whole, decision theory seems to have about the same connection to real decisions as war games do to real wars.

What are the main complications? For some events, I may have a rough idea of likelihood: One event is very likely, another is unlikely, a third is uncertain. However, I may not be able to quantify these likelihoods, even to one or two decimal places; and there will be many events whose probabilities are simply unknown—even if definable.[10] Likewise, there are some benefits that can be assessed with reasonable accuracy; others can be estimated only to rough orders of magnitude; in some cases, quantification may not be possible at all. Thus, utilities may be just as problematic as priors.

The theorems that derive probabilities and utilities from axioms push the difficulties back one step.[11] In real examples, the existence of many states of nature must remain unsuspected. Only some acts can be contemplated; others are not imaginable until the moment of truth arrives. Of the acts that can be imagined, the decision-maker will have preferences between some pairs but not others. Too, common knowledge suggests that consequences are often quite different in the foreseeing and in the experiencing.

Intransitivity would be an argument for revision, although not a decisive one; for example, a person choosing among several job offers might well have intransitive preferences, which it would be a mistake to ignore. By way of contrast, an arbitrageur who trades bonds intransitively is likely to lose a lot of money. (There is an active market in bonds, while the market in job offers—largely nontransferable—must be rather thin; the practical details make a difference.) The axioms do not capture the texture of real decision making. Therefore, the theory has little normative force.

The fallback defense. Some Bayesians will concede much of what I have said: The axioms are not binding; rational decision-makers may have neither priors nor utilities. Still, the following sorts of arguments can be heard. The decision-maker must have some ideas about relative likelihoods for a few events; a prior probability can be made up to capture such intuitions, at least in gross outline. The details (for instance, that distributions are normal) can be chosen on the basis of convenience. A utility function can be put together using similar logic: The decision-maker must perceive some consequences as very good, and big utility numbers can be assigned to these; he must perceive some other consequences as trivial, and small utilities can be assigned to those; and in between is in between. The Bayesian engine can now be put to work, using such approximate priors and utilities. Even with these fairly crude approximations, Bayesian analysis is held to dominate other forms of inference: That is the fallback defense.

Here is my reaction to such arguments. Approximate Bayesian analysis may in principle be useful. That this mode of analysis dominates other forms of inference, however, seems quite debatable. In a statistical decision problem, where the model and loss function are given, Bayes procedures are often hard to beat, as are objectivist likelihood procedures; with many of the familiar textbook models, objectivist and subjectivist procedures should give similar results if the data set is large. There are sharp mathematical theorems to back up such statements.[12] On the other hand, in real problems—where models and loss functions are mere

approximations—the optimality of Bayes procedures cannot be a mathematical proposition. And empirical proof is conspicuously absent.

If we could quantify breakdowns in model assumptions, or degrees of error in approximate priors and loss functions, the balance of argument might shift considerably. The rhetoric of "robustness" may suggest that such error analyses are routine. This is hardly the case even for the models. For priors and utilities, the position is even worse, since the entities being approximated do not have any independent existence—outside the Bayesian framework that has been imposed on the problem.

De Finetti's theorem. Suppose you are a radical subjectivist, watching a sequence of 0's and 1's. In your prior opinion, this sequence is exchangeable: Permuting the order of the variables will not change your opinion about them. A beautiful theorem of de Finetti's asserts that your opinion can be represented as coin tossing, the probability of heads being selected at random from a suitable prior distribution. This theorem is often said to "explain" subjective or objective probabilities, or justify one system in terms of the other.[13]

Such claims cannot be right. What the theorem does is this: It enables the subjectivist to discover features of his prior by mathematical proof, rather than introspection. For example, suppose you have an exchangeable prior about those 0's and 1's. Before data collection starts, de Finetti will prove to you by pure mathematics that in your own opinion the relative frequency of 1's among the first n observations will almost surely converge to a limit as $n \to \infty$. (Of course, the theorem has other consequences too, but all have the same logical texture.)

This notion of "almost surely," and the limiting relative frequency, are features of your opinion not of any external reality. ("Almost surely" means with probability 1, and the probability in question is your prior.) Indeed, if you had not noticed these consequences of your prior by introspection, and now do not like them, you are free to revise your opinion—which will have no impact outside your head. What the theorem does is to show how various aspects of your prior opinion are related to each other. That is all the theorem can do, because the conditions of the theorem are conditions on the prior alone.

To illustrate the difficulty, I cite an old friend rather than making a new enemy. According to Jeffrey (1983, p. 199), de Finetti's result proves "your subjective probability measure [is] a certain mixture or weighted average of the various possible objective probability measures"—an unusually clear statement of the interpretation that I deny. Each of Jeffrey's "objective" probability measures governs the tosses of a p-coin, where p is your limiting relative frequency of 1's. (Of course, p has a probability

distribution of its own, in your opinion.) Thus, p is a feature of your opinion, not of the real world: The mixands in de Finetti's theorem are "objective" only by terminological courtesy. In short, the "p-coins" that come out of de Finetti's theorem are just as subjective as the prior that went in.

1.4.2 To sum up

The theory—as developed by Ramsey, von Neumann and Morgenstern, de Finetti, and Savage, among others—is great work. They solved an important historical problem of interest to economists, mathematicians, statisticians, and philosophers alike. On a more practical level, the language of subjective probability is evocative. Some investigators find the consistency of Bayesian statistics to be a useful discipline; for some (including me), the Bayesian approach can suggest statistical procedures whose behavior is worth investigating. But the theory is not a complete account of rationality, or even close. Nor is it the prescribed solution for any large number of problems in applied statistics, at least as I see matters.

1.5 Statistical models

Of course, statistical models are applied not only to coin tossing but also to more complex systems. For example, "regression models" are widely used in the social sciences, as indicated below; such applications raise serious epistemological questions. (This idea will be developed from an objectivist perspective, but similar issues are felt in the other camp.)

The problem is not purely academic. The census suffers an undercount, more severe in some places than others; if certain statistical models are to be believed, the undercount can be corrected—moving seats in Congress and millions of dollars a year in entitlement funds (*Survey Methodology* (1992) 18(1); *Jurimetrics* (1993) 34(1); *Statistical Science* (1994) 9(4). If yet other statistical models are to be believed, the veil of secrecy can be lifted from the ballot box, enabling the experts to determine how racial or ethnic groups have voted—a crucial step in litigation to enforce minority voting rights (*Evaluation Review*, (1991) 1(6); Klein and Freedman, 1993).

1.5.1 Examples

Here, I begin with a noncontroversial example from physics, namely, Hooke's law: Strain is proportional to stress. We will have some number n of observations. For the ith observation, indicated by the subscript i, we hang weight$_i$ on a spring. The length of the spring is measured as length$_i$. The regression model says that[14]

(1) $$\text{length}_i = a + b \times \text{weight}_i + \epsilon_i.$$