Foundations of Science
1 (1995/96), 19-39

David Freedman
Statistics Department
University of California
Berkeley, CA 94720, USA

SOME ISSUES IN THE FOUNDATION OF STATISTICS

> "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]

Key Words: Statistics, Probability, Objectivist, Subjectivist, Bayes, de Finetti, Decision theory, Model validation, Regression.

Abstract. After sketching the conflict between objectivists and subjectivists on the foundations of statistics, this paper 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]From 'The Idyll of Miss Sarah Brown', *Collier's Magazine*, 1933. Reprinted in *Guys and Dolls: The Stories of Damon Runyon.* Penguin Books, New York, 1992, pp.14-26. The quote is edited slightly, for continuity.

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 1. 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:

- the "objectivist" school, also called the "frequentists";
- the "subjectivist" school, also called the "Bayesians," after the Reverend Thomas Bayes (England, c.1701-1761).

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 second, 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.

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

[2]This note will give a compact statement of Kolmogorov's axioms. Let Ω be a set. By definition, a σ-field $\mathcal{F}$ is a collection of subsets of Ω, which has Ω itself as a member. Furthermore,

- $\mathcal{F}$ is closed under complementation (if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$), and
- $\mathcal{F}$ is closed under the formation of countable unions: if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, then $\cup_i A_i \in \mathcal{F}$.

A probability P is a non-negative, real-valued function on Ω, such that $P(\Omega) = 1$ and P is countably additive: if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, and the sets are pairwise disjoint, in the sense that $A_i \cap A_j = \emptyset$ for $i \neq j$, then $P(\cup_i A_i) = \sum_i P(A_i)$. A random variable X is an $\mathcal{F}$-measurable function on Ω. Informally, probabilists might say that Nature chooses $\omega \in \Omega$ according to P, and shows you $X(\omega)$; the latter would be the "observed value" of X.

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 5,067 heads: the relative frequency was 5,067/10,000 = 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 hypothesis concerning it.

The objectivist position exposes one to certain famous difficulties. As Keynes said, "In the long run, we are all dead." Heraclitus' epigram (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, 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.

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 10 times, landing heads 6 times and tails 4 times, what is the chance that it will land heads on the 11th 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 them to certain rhetorical awkwardness; for example, they are required not to understand the idea of a 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]

Probability and relative frequency. In ordinary language, "probabilities"

[3]Models will be discussed in section 5. Those for presidential elections may not be compelling. For genetics, however, chance models are well established; and many statistical calculations are therefore on a secure footing. Much controversy remains, for example, in the area of DNA identification (*Jurimetrics*, vol. 34, no. 1, 1993).

[4]The distinction between classical and radical subjectivists made here is not often discussed in the statistical literature; the terminology is not standard. See, for instance, Diaconis and Freedman (1980a), Efron (1986), Jeffrey (1983, sec. 12.6).

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 10 heads and 10 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: 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]

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:

[5]Some readers may say to themselves that here, probability is just the relative frequency of transitions. However, a similar but slightly more complicated example can be rigged up for transition counts; an infinite regress lies just ahead. My point is only this: relative frequencies are not probabilities. Of course, if circumstances are favorable, the two are strongly connected – that is one reason why chance models are useful for applied work.

[6]To illustrate the objectivist way of handling probabilities and relative frequencies, I consider repeated tosses of a fair coin: the probability of heads is 50%. In a sequence of 10,000 tosses, the chance of getting between 49% and 51% heads is about 95%. In replications of this (large) experiment, about 95% of the time, there will be between 49% and 51% heads. On each replication, however, the probability of heads stays the same – namely, 50%.

The strong law of large numbers provides another illustration. Consider n repeated tosses of a fair coin. With probability 1, as $n \to \infty$, the relative frequency of heads in the first n tosses eventually gets trapped inside the interval from 49% to 51%; ditto, for the interval from 49.9% to 50.1%; ditto, for the interval from 49.99% to 50.01%; and so forth. No matter what the relative frequency of heads happens to be at any given moment, the probability of heads stays the same – namely, 50%. Probability is not relative frequency.

subjective elements come in. Likewise, subjectivists may tell you that (i) objectivists 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.

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. 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.

Other arguments for the Bayesian position. Coherence. There are well-

[7]Similarly, a large part of objectivist statistics is about what you would do if you had a model; and all of us spend enormous amounts of energy finding out what would happen if the data kept pouring in. I wish we could learn to look at the data more directly, without the fictional models and priors. On the same wish-list: we stop pretending to fix bad designs and inadequate measurements by modeling.

known theorems, including (Freedman and Purves, 1969), showing that stubborn non-Bayesian behavior has costs. They 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.[8] 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 played on a table 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

[8] A "dutch book" is a collection of bets on various events such that the bettor makes money, no matter what the outcome.

[9] According to the "sure thing principle," if I prefer A to B given that C occurs, and I also prefer A to B given that C does not occur, I must prefer A to B when I am in doubt as to the occurrence of C.

[10] Although one-sentence concessions in a book are not binding, Savage (1954, p.59) does say that his theory "is a code of consistency for the person applying it, not a system of

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 non-transferable – 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

---

predictions about the world"; and personal probabilities can be known "only roughly."

Another comment on this book may be in order. According to Savage (1954, pp.61-62), "on no ordinary objectivistic view would it be meaningful, let alone true, to say that on the basis of the available evidence it is very improbable, though not impossible, that France will become a monarchy within the next decade." As anthropology of science, this seems wrong. I make qualitative statements about likelihoods and possibilities, and expect to be understood; I find such statements meaningful when others make them. Only the quantification seems problematic: What would it mean to say that P(France will become a monarchy) = .0032? Many objectivists of my acquaintance share such views; although caution is in order when extrapolating from such a sample of convenience.

[11]The argument in the text is addressed to readers who have some familiarity with the axioms. This note gives a very brief review; Kreps (1988) has a chatty and sympathetic discussion (although some of the details are not quite in focus); Le Cam (1977) is more technical and critical.

In the axiomatic setup, there is a space of "states of nature," like the possible orders in which horses finish a race. There is another space of "consequences"; these can be pecuniary or non-pecuniary (win \$1,000, lose \$5,000, win a weekend in Philadelphia, etc.). Mathematically, an "act" is a function whose domain is the space of states of nature, and whose values are consequences. You have to choose an act: that is the decision problem. Informally, if you choose the act f, and the state of nature happens to be s, you enjoy (or suffer) the consequence $f(s)$. For example, if you bet on those horses, the payoff depends on the order in which they finish: the bet is an act, and the consequence depends on the state of nature. The set of possible states of nature, the set of possible consequences, and the set of possible acts are all viewed as fixed and known. You are supposed to have a transitive preference ordering on the acts, not just the consequences. The sure thing principle is an axiom in Savage's setup.

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; 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

[12]Wald's idea of a statistical decision problem can be sketched, as follows. There is an unobservable parameter. Corresponding to each parameter value θ, there is a known probability distribution P_θ for an observable random quantity X. (This family of probability distributions is a "statistical model" for X, with parameter θ.) There is a set of possible "decisions"; there is a "loss function" $L(d, \theta)$ which tells you how much is lost by making the decision d when the parameter is really θ. (For example, d might be an estimate of θ, and loss might be squared error.) You have to choose a "decision rule," which is a mapping from observed values of X to decisions. Your objective is to minimize "risk," that is, expected loss.

A comparison with the setup in note 11 may be useful. The "state of nature" seems to consist of the observable value of X, together with the unobservable value θ of the parameter. The "consequences" are the decisions, and "acts" are decision rules. (The conflict in terminology is regrettable, but there is no going back.) The utility function is replaced by L, which is given but depends on θ as well as d.

A Bayes' procedure is optimal in that its risk cannot be reduced for all values of θ; any such "admissible" procedure is a limit of Bayes' procedures ("the complete class theorem"). The maximum likelihood estimator is "efficient"; and its sampling distribution is close to the posterior distribution of θ by the "Bernstein-von Mises theorem," which is actually due to Laplace. More or less stringent regularity conditions must be imposed to prove any of these results, and some of the theorems must be read rather literally; Stein's paradox and Bahadur's example should at least be mentioned.

Standard monographs and texts include Berger (1985), Berger and Wolpert (1988), Bickel and Doksum (1977), Casella and Berger (1990), Ferguson (1967), Le Cam (1986),

– 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,

---

Lehmann (1983, 1986), and Rao (1973). The Bernstein-von Mises theorem is discussed in Le Cam and Yang (1990) and Prakasa Rao (1987).

Of course, in many contexts, Bayes procedures and frequentist procedures will go in opposite directions; for a review, see Diaconis and Freedman (1986). These references are all fairly technical.

[13]Diaconis and Freedman (1980ab, 1981) review the issues and the mathematics. The first-cited paper is relatively informal; the second gives a version of de Finetti's theorem applicable to a finite number of observations, with bounds; the last gives a fairly general mathematical treatment of partial exchangeability, with numerous examples and it is quite technical. More recent work is described in Diaconis and Freedman (1988, 1990).

The usual hyperbole can be sampled in Kreps (1988, p.145): de Finetti's theorem is "the fundamental theorem of statistical inference – the theorem that from a subjectivist point of view makes sense out of most statistical procedures." This interpretation of the theorem fails to distinguish between what is assumed and what is proved. It is the assumption of exchangeability that enables you to predict the future from the past, at least to your own satisfaction, not the conclusions of the theorem or the elegance of the proof. If have an exchangeable prior, the statistical world looks like your oyster, de Finetti or no de Finetti.

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 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.

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.

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*, vol. 18, no. 1, 1992; *Jurimetrics*, vol. 34, no. 1, 1993; *Statistical Science*, vol. 9, no. 4, 1994). 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*, vol. 15, no. 6, 1991; Klein and Freedman, 1993).

Here, I begin with a (relatively) non-controversial example from physics – Hooke's law: strain is proportional to stress. (This law is named after Robert Hooke, England, 1653-1703.) 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 (for quite a large range of weights [14]),

$$\text{length}_i = a + b \times \text{weight}_i + \varepsilon_i. \tag{1}$$

The "error" term ε_i is needed because measured length will not be exactly equal to $a + b \times$ weight. If nothing else, measurement error must be reckoned with. We model ε_i as a sequence of draws, made at random with replacement from a box of tickets; each ticket shows a potential error – the ε_i that will be realized if that ticket is the ith one drawn. The average of all the potential errors in the box is assumed to be 0. In more standard terminology, the ε_i are assumed to be "independent and identically distributed with mean 0." Such assumptions can present difficult scientific issues, because error terms are not observable.

In equation (1), a and b are parameters, unknown constants of nature that characterize the spring: a is the length of the spring under no load, and b is stretchiness – the increase in length per unit increase in weight. These parameters are not observable, but they can be estimated by "the method of least squares," developed by Adrien-Marie Legendre (France, 1752-1833) and Carl Friedrich Gauss (Germany, 1777-1855) to fit astronomical orbits. Basically, you choose the values of $\hat{a}$ and $\hat{b}$ to minimize the sum of the squared "prediction errors," $\sum_i e_i^2$, where e_i is the prediction error for the ith observation:[15]

[14] With large-enough weights, a quadratic term will be needed in equation (1). Moreover, beyond some point, the spring passes its "elastic limit" and snaps.

[15] The residual e_i is observable, but is only an approximation to the disturbance term ε_i in (1); that is because the estimates $\hat{a}$ and $\hat{b}$ are only approximations to the parameters a and b.

$$e_i = \text{length}_i - \hat{a} - \hat{b} \times \text{weight}_i. \quad (2)$$

These prediction errors are often called "residuals:" they measure the difference between the actual length and the predicted length, the latter being $\hat{a} + \hat{b} \times$ weight.

No one really imagines there to be a box of tickets hidden in the spring. However, the variability of physical measurements (under many but by no means all circumstances) does seem to be remarkably like the variability in draws from a box. This is Gauss' model for measurement error. In short, statistical models can be constructed that correspond rather closely to empirical phenomena.

I turn now to social-science applications. A case study would take us too far afield, but a stylized example – regression analysis used to demonstrate sex discrimination in salaries, adapted from (Kaye and Freedman, 1994) – may give the idea. We use a regression model to predict salaries (dollars per year) of employees in a firm from:

- education (years of schooling completed),
- experience (years with the firm),
- the dummy variable "man," which takes the value 1 for men and 0 for women.

Employees are indexed by the subscript i; for example, salary_i is the salary of the ith employee.

The equation is [16]

$$\text{salary}_i = a + b \times \text{education}_i + c \times \text{experience}_i + d \times \text{man}_i + \varepsilon_i. \quad (3)$$

Equation (3) is a statistical model for the data, with unknown parameters a, b, c, d; here, a is the "intercept" and the others are "regression coefficients"; ε_i is an unobservable error term. This is a formal analog of Hooke's law (1); the same assumptions are made about the errors. In other words, an employee's salary is determined as if by computing

$$a + b \times \text{education} + c \times \text{experience} + d \times \text{man}, \quad (4)$$

[16] Such equations are suggested, somewhat loosely, by "human capital theory." However, there remains considerable uncertainty about which variables to put into the equation, what functional form to assume, and how error terms are supposed to behave. Adding more variables is no panacea: Freedman (1983), Clogg and Haritou (1994).

then adding an error drawn at random from a box of tickets. The display (4) is the expected value for salary given the explanatory variables (education, experience, man); the error term in (3) represents deviations from the expected.

The parameters in (3) are estimated from the data using least squares. If the estimated coefficient d for the dummy variable turns out to be positive and "statistically significant" (by a "t-test"), that would be taken as evidence of disparate impact: men earn more than women, even after adjusting for differences in background factors that might affect productivity. Education and experience are entered into equation (3) as "statistical controls," precisely in order to claim that adjustment has been made for differences in backgrounds.

Suppose the estimated equation turns out as follows:

$$\text{predicted salary} = \$7,100 + \$1,300 \times \text{education} + \$2,200 \times \text{experience} + \$700 \times \text{man}. \tag{5}$$

That is, $\hat{a} = \$7,100$, $\hat{b} = \$1,300$, and so forth. According to equation (5), every extra year of education is worth on average \$1,300; similarly, every extra year of experience is worth on average \$2,200; and, most important, men get an premium of \$700 over women with the same education and experience, on average.

A numerical example will illustrate (5). A male employee with 12 years of education (high school) and 10 years of experience would have a predicted salary of

$$\$7,100 + \$1,300 \times 12 + \$2,200 \times 10 + \$700 \times 1 = \$7,100 + \$15,600 + \$22,000 + \$700 = \$45,400. \tag{6}$$

A similarly situated female employee has a predicted salary of only

$$\$7,100 + \$1,300 \times 12 + \$2,200 \times 10 + \$700 \times 0 = \$7,100 + \$15,600 + \$22,000 + \$0 = \$44,700. \tag{7}$$

Notice the impact of the dummy variable: \$700 is added to (6), but not to (7).

A major step in the argument is establishing that the estimated coefficient of the dummy variable in (3) is "statistically significant." This step

turns out to depend on the statistical assumptions built into the model. For instance, each extra year of education is assumed to be worth the same (on average) across all levels of experience, both for men and women. Similarly, each extra year of experience is worth the same across all levels of education, both for men and women. Furthermore, the premium paid to men does not depend systematically on education or experience. Ability, quality of education, or quality of experience are assumed not to make any systematic difference to the predictions of the model.

The story about the error term – that the ε's are independent and identically distributed from person to person in the data set – turns out to be critical for computing statistical significance. Discrimination cannot be proved by regression modeling unless statistical significance can be established, and statistical significance cannot be established unless conventional presuppositions are made about unobservable error terms.

Lurking behind the typical regression model will be found a host of such assumptions; without them, legitimate inferences cannot be drawn from the model. There are statistical procedures for testing some of these assumptions. However, the tests often lack the power to detect substantial failures. Furthermore, model testing may become circular; breakdowns in assumptions are detected, and the model is redefined to accommodate. In short, hiding the problems can become a major goal of model building.

Using models to make predictions of the future, or the results of interventions, would be a valuable corrective. Testing the model on a variety of data sets – rather than fitting refinements over and over again to the same data set – might be a good second-best (Ehrenberg and Bound, 1993). With Hooke's law (1), the model makes predictions that are relatively easy to test experimentally. For the salary discrimination model (3), validation seems much more difficult. Thus, built into the equation is a model for non-discriminatory behavior: the coefficient d vanishes. If the company discriminates, that part of the model cannot be validated at all.

Regression models like (3) are widely used by social scientists to make causal inferences; such models are now almost a routine way of demonstrating counter-factuals. However, the "demonstrations" generally turn out to be depend on a series of untested, even unarticulated, technical assumptions. Under the circumstances, reliance on model outputs may be quite unjustified. Making the ideas of validation somewhat more precise is a serious problem in the philosophy of science. That models should correspond to reality is, after all, a useful but not totally straightforward idea – with some history to it. Developing models, and testing their connection to the

phenomena, is a serious problem in statistics.[17]

Standard errors, t-statistics, and statistical significance. The "standard error" of $\hat{d}$ measures the likely difference between $\hat{d}$ and d, due to the action of the error terms in equation (3). The "t-statistic" is $\hat{d}$ divided by its standard error. Under the "null hypothesis" that $d = 0$, there is only about a 5% chance that $|t| > 2$. Such a large value of t would demonstrate "statistical significance." Of course, the parameter d is only a construct in a model. If the model is wrong, the standard error, t-statistic, and significance level are rather difficult to interpret.

Even if the model is granted, there is a further issue: the 5% is a probability for the data given the model, namely, $P\{|t| > 2 \parallel d = 0\}$. However, the 5% is often misinterpreted as $P\{d = 0 \mid \text{data}\}$. Indeed, this misinterpretation is a commonplace in the social-science literature, and seems to have been picked up by the courts from expert testimony.[18] For an objectivist, $P\{d = 0 \mid \text{data}\}$ makes no sense: parameters do not exhibit chance variation. For a subjectivist, $P\{d = 0 \mid \text{data}\}$ makes good sense, but its computation via the t-test is grossly wrong, because the prior probability that $d = 0$ has not been taken into account: the calculation exemplifies the "base rate fallacy." Power matters too.

(The single vertical bar "|" is standard notation for conditional proba-

[17] For more discussion in the context of real examples, with citations to the literature of model validation, see Freedman (1985, 1987, 1991, 1994). Many recent issues of *Sociological Methodology* have essays on this topic. Also see Oakes (1990), who discusses modeling issues, significance tests, and the objectivist-subjectivist divide.

[18] Some legal citations may be of interest (Kaye and Freedman, 1994): Waisome v. Port Authority, 948 F.2d 1370, 1376 (2d Cir. 1991) ("Social scientists consider a finding of two standard deviations significant, meaning there is about 1 chance in 20 that the explanation for a deviation could be random"); Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (5th Cir. 1982) ("A variation of two standard deviations would indicate that the probability of the observed outcome occurring purely by chance would be approximately five out of 100; that is, it could be said with a 95% certainty that the outcome was not merely a fluke."); Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, 271 (N.D. Tex. 1980), vacated and remanded, 723 F.2d 1195 (5th Cir. 1984) ("if a 5% level of significance is used, a sufficiently large t-statistic for the coefficient indicates that the chances are less than one in 20 that the true coefficient is actually zero.").

An example from the underlying technical literature may also be of interest. According to (Fisher, 1980, p.717), "in large samples, a t-statistic of approximately two means that the chances are less than one in twenty that the true coefficient is actually zero and that we are observing a larger coefficient just by chance ... A t-statistic of approximately two and one half means the chances are only one in one hundred that the true coefficient is zero ..." No. If the true coefficient is zero, there is only one chance in one hundred that $|t| > 2.5$. (Frank Fisher is a well known econometrician who often testifies as an expert witness, although I do not believe he figures in any of the cases cited above.)

bility. The double vertical bar "||" is not standard; Bayesians might want to read this as a conditional probability; for an objectivist, || is intended to mean "computed on the assumption that ...")

Statistical models and the problem of induction. How do we learn from experience? What makes us think that the future will be like the past? With contemporary modeling techniques, such questions are easily answered – in form if not in substance.

- The objectivist invents a regression model for the data, and assumes the error terms to be independent and identically distributed; "iid" is the conventional abbreviation. It is this assumption of iid-ness that enables us to predict data we have not seen from a training sample – without doing the hard work of validating the model.

- The classical subjectivist invents a regression model for the data, assumes iid errors, and then makes up a prior for unknown parameters.

- The radical subjectivist adopts an exchangeable or partially exchangeable prior, and calls you irrational or incoherent (or both) for not following suit.

In our days, serious arguments have been made from data. Beautiful, delicate theorems have been proved; although the connection with data analysis often remains to be established. And an enormous amount of fiction has been produced, masquerading as rigorous science.

6. Conclusions

I have sketched two main positions in contemporary statistics, objectivist and subjectivist, and tried to indicate the difficulties. Some questions confront statisticians from both camps: How do statistical models connect with reality? What areas lend themselves to investigation by statistical modeling? When are such investigations likely to be sterile?

These questions have philosophical components as well as technical ones. I believe model validation to be a central issue. Of course, many of my colleagues will be found to disagree. For them, fitting models to data, computing standard errors, and performing significance tests is "informative," even though the basic statistical assumptions (linearity, independence of errors, etc.) cannot be validated. This position seems indefensible, nor are the consequences trivial. Perhaps it is time to reconsider.

Acknowledgments. I would like to thank Dick Berk, Cliff Clogg, Persi Diaconis, Joe Eaton, Neil Henry, Paul Humphreys, Lucien Le Cam, Diana Petitti, Brian Skyrms, Terry Speed, Steve Turner, Amos Tversky, Ken Wachter and Don Ylvisaker for many helpful suggestions – some of which I could implement.

References

Bayes, Thomas (1764), An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London* **53**, 370-418.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. New York, Springer-Verlag.

Berger, J. and Wolpert, R. (1988), *The Likelihood Principle.* 2nd ed. Hayward, Calif., Institute of Mathematical Statistics.

Bickel, P. J. and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco, Holden-Day.

Box, G. E. P. and Tiao, G. C. (1992), *Bayesian Inference in Statistical Analysis.* New York, Wiley.

Casella, G. and Berger, R. L. (1990), *Statistical Inference.* Pacific Grove, Calif., Wadsworth & Brooks/Cole.

Clogg, C. C. and Haritou, A. (1994), The Regression Method of Causal Inference and a Dilemma with this Method. Technical report, Department of Sociology, Pennsylvania State University. To appear in V. McKim and S. Turner (eds), *Proceedings of the Notre Dame Conference on Causality in Crisis.*

de Finetti, B. (1959), *La probabilità, la statistica, nei rapporti con l'induzione, secondo diversi punti di vista.* Rome, Centro Internazionale Matematica Estivo Cremonese. English translation in de Finetti (1972).

de Finetti, B. (1972), *Probability, Induction, and Statistics.* New York, Wiley.

Diaconis, P. and Freedman, D. (1980a), de Finetti's Generalizations of Exchangeability, pp.233-50 in Richard C. Jeffrey (ed), *Studies in Inductive Logic and Probability.* Vol. 2. Berkeley, University of California Press.

Diaconis, P. and Freedman, D. (1980b), Finite Exchangeable Sequences, *Annals of Probability* **8**, 745-64.

Diaconis, P. and Freedman, D. (1981), Partial Exchangeability and Sufficiency, pp.205-36 in *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions. Sankhya.* Calcutta, Indian Statistical Institute.

Diaconis, P. and Freedman. D. (1986), On the Consistency of Bayes' Estimates, *Annals of Statistics* **14**, 1-87, with discussion.

Diaconis, P. and Freedman. D. (1988), Conditional Limit Theorems for Exponential Families and Finite Versions of de Finetti's Theorem, *Journal of Theoretical Probability* **1**, 381-410.

Diaconis, P. and Freedman. D. (1990), Cauchy's Equation and de Finetti's Theorem, *Scandinavian Journal of Statistics* **17**, 235-50.

Efron, B. (1986), Why Isn't Everyone a Bayesian? *The American Statistician* **40**, 1-11, with discussion.

Ehrenberg, A. S. C. and Bound, J. A. (1993), Predictability and Prediction, *Journal of the Royal Statistical Society,* **Series A, Part 2, 156**, 167-206.

Ferguson, T. (1967), *Mathematical Statistics: a Decision Theoretic Approach.* New York, Academic Press.

Fisher, F. M. (1980), Multiple Regression in Legal Proceedings, *Columbia Law Review* **80**, 702-36.

Fisher, R. A. (1959), *Smoking: The Cancer Controversy.* Edinburgh, Oliver & Boyd. see pp.25-29 on relevant subsets.

Freedman, D. (1983), A Note on Screening Regression Equations, *The American Statistician* **37**, 152-55.

Freedman, D. (1985), Statistics and the Scientific Method, pp.343-90 in W. M. Mason and S. E. Fienberg (eds), *Cohort Analysis in Social Research: Beyond the Identification Problem.* New York, Springer.

Freedman, D. (1987), As Others See Us: A Case Study in Path Analysis, *Journal of Educational Statistics* **12**, no. 2, 101-223, with discussion.

Freedman, D. (1991), Statistical Models and Shoe Leather, chapter 10 in Peter Marsden (ed), *Sociological Methodology 1991*, with discussion.

Freedman, D. (1994), From Association to Causation Via Regression. Technical report no. 408, Statistics Department, University of California, Berkeley. To appear in V. McKim and S. Turner (eds), *Proceedings of the Notre Dame Conference on Causality in Crisis.*

Freedman, D. and Purves, R. (1969), Bayes Method for Bookies, *Annals of Mathematical Statistics* **40**, 1177-86.

Freedman, D., Pisani, R., Purves, R. and Adhikari, A. (1991), *Statistics*, 2nd ed. New York, Norton.

Gatsonis, C. et al., eds. (1993), *Case Studies in Bayesian Statistics.* New York, Springer-Verlag, Lecture Notes in Statistics, vol. 83.

Gauss, C. F. (1809), *Theoria Motus Corporum Coelestium.* Hamburg, Perthes et Besser. Reprinted in 1963 by Dover, New York.

Jeffrey, R. C. (1983), *The Logic of Decision.* 2nd ed. University of Chicago

Press.

Kahneman, D., Slovic, P. and Tversky, A., eds. (1982), *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press.

Kaye, D. and Freedman, D. (1994), *Reference Manual on Statistics.* Washington, D.C., Federal Judicial Center.

Klein, S. and Freedman, D. (1993), Ecological Regression in Voting Rights Cases, *Chance Magazine* **6**, 38-43.

Kolmogorov, A. N. (1933), Grundbegriffe der Wahrscheinlichkeitstheorie, *Ergebnisse Mathematische* **2** no. 3.

Kreps, D. (1988), *Notes on the Theory of Choice.* Boulder, Westview Press.

Laplace, P. S. (1774), Memoire sur la probabilité des causes par les évenements, *Memoires de mathématique et de physique presentés a l'académie royale des sciences, par divers savants, et lûs dans ses assemblées* **6**. Reprinted in Laplace's *Oeuvres Complètes* **8**, 27-65. English translation by S. Stigler (1986), *Statistical Science* **1**, 359-378.

Le Cam, Lucien M. (1977), A Note on Metastatistics or "An Essay Toward Stating a Problem in the Doctrine of Chances," *Synthese* **36**, 133-60.

Le Cam, Lucien M. (1986), *Asymptotic Methods in Statistical Decision Theory.* New York, Springer-Verlag.

Le Cam, Lucien M. and Yang, Grace Lo (1990), *Asymptotics in Statistics: Some Basic Concepts.* New York, Springer-Verlag.

Lehmann, E. (1986), *Testing Statistical Hypotheses.* 2nd ed. Pacific Grove, Calif., Wadsworth & Brooks/Cole.

Lehmann, E. (1983), *Theory of Point Estimation.* Pacific Grove, Calif., Wadsworth & Brooks/Cole.

McNeil, B., Pauker, S., Sox, H. Jr., and Tversky, A. (1982), On the Elicitation of Preferences for Alternative Therapies, *New England Journal of Medicine* **306**, 1259-62.

Oakes, M. (1990), *Statistical Inference.* Chestnut Hill, Mass., Epidemiology Rescurces, Inc.

O'Hagan, A. (1988), *Probability: Methods and Measurement.* London, Chapman and Hall.

Peirce, C. S. (1878), The Doctrine of Chances, *Popular Science Monthly* **12** (March 1878), pp.604-615. Reprinted as pp.142-54 in N. Houser and C. Kloesel (eds) (1992), *The Essential Peirce.* Indiana University Press.

Popper, K. (1983), *Realism and the Aim of Science.* Totowa, N.J., Rowman and Littlefield.

Prakasa Rao, B. L. S. (1987), *Asymptotic Theory of Statistical Inference.* New York, Wiley.

Ramsey, F. P. (1926), in R. B. Braithwaite (1931), *The Foundations of Mathematics and other Logical Essays.* London, Routledge and Kegan Paul.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications.* 2nd ed. New York, Wiley.

Savage, Leonard J. (1972), *The Foundations of Statistics.* 2d rev. ed. New York, Dover Publications.

Stigler, S. (1986), *The History of Statistics.* Harvard University Press.

Tversky, A. and Kahneman, D. (1986), Rational Choice and the Framing of Decisions, *Journal of Business* **59**, no. 4, part 2, pp.S251-78.

Tversky, A. and Kahneman, D. (1983), Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment, *Psychological Review* **90**, 293-315.

von Mises, R. (1964), *Mathematical Theory of Probability and Statistics.* H. Geiringer (ed). New York, Academic Press.

von Neumann, J. and Morgenstern, O. (1944), *Theory of Games and Economic Behavior.* Princeton University Press.