*Statistical Science*
1998, Vol. 13, No. 3, 248–276

# Logicist Statistics I. Models and Modeling

**A. P. Dempster**

*Abstract.* Arguments are presented to support increased emphasis on logical aspects of formal methods of analysis, depending on probability in the sense of R. A. Fisher. Formulating probabilistic models that convey uncertain knowledge of objective phenomena and using such models for inductive reasoning are central activities of individuals that introduce limited but necessary subjectivity into science. Statistical models are classified into overlapping types called here *empirical*, *stochastic* and *predictive*, all drawing on a common mathematical theory of probability, and all facilitating statements with logical and epistemic content. Contexts in which these ideas are intended to apply are discussed via three major examples.

*Key words and phrases:* Logicism and proceduralism; specificity of analysis; formal subjective probability; complementarity; subjective and objective; formal and informal; empirical, stochastic and predictive models; U.S. national census; screening for chronic disease; global climate change.

## 1. FOUNDATIONS: "WHAT WE CAN SAY"

The aim of this paper and a planned sequel on statistical inference is to direct attention to an alternative account of the language and logic of applied statistics. The focus of the present paper on models and modeling derives from a perception that these topics are too little emphasized in expositions of statistics, despite their centrality. On one side, data analysis is directed at exposing structure whose formal content can be described as mathematical models, while on the other side formal statistical inference depends essentially upon assumed formal probability models. Statistical inference consists by definition of the interpretation of uncertain reality through models.

My account synthesizes subjective and objective aspects of practice. Another theme emphasizes an ever-present interplay of formal and informal elements in scientific analysis. My most basic aim, however, is to distinguish the "logic" of reasoning about a specific situation under analysis, from "procedures" that I suggest are best understood as packaged templates for logic. Whereas textbook descriptions of practice often appear to suggest that "what statisticians do is choose and apply procedures," I argue that statistics is equally importantly about "what we can say"[1] beyond reporting the products of applying defined procedures. This is especially true of statistical inference. (Superscript numerals refer to the notes collected in Section 6.)

The "proceduralist" versus "logicist" theme goes to the heart of the protracted controversies[2] that pitted Ronald Fisher against Jerzy Neyman and his school. Neyman followed Fisher by placing sampling models at the epicenter of statistical inference, all but eliminating the widespread earlier acceptance of Bayesian logic.[3] The mathematical properties of sampling distributions that Fisher derived with great skill were fundamental to Neyman's "frequentist" theory of choosing procedures, so that Fisher has often been adopted by Neymanians as a distinguished predecessor. Whereas many statisticians are puzzled that Fisher was so adamantly opposed to Neyman's transparent and often elegant theory,[4] my perception is that Fisher was engaged in insisting that the logic of interpreting each individual application is primary.[5] I believe it is also transparent that logicism is more inclusive than proceduralism, and despite Neyman's objections[6] deserves rehabilitation.

The formal outline of an inferential procedure, say for parameter estimation or for testing a null hypothesis, is much the same between Fisher and Neyman, but their conceptions of how to choose a procedure are fundamentally different. Fisher in-

*A. P. Dempster is Professor of Theoretical Statistics, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138* (*e-mail: dempster@stat.harvard.edu*).

stinctively thought in terms of the logic of specific inferences and in terms of choosing a procedure that captures all the information in a data set, as, for example, by sufficiency. There is no necessary role for long run operating characteristics in this program. Fisher undercut his own influence, however, through intemperate refusals to allow that frequentist properties of procedures are theoretically informative and have interesting connections with his own conceptions of information.[7] The bitterness of the dispute, together with the fact that most mathematical statisticians sided with Neyman, resulted in Fisher's logicist position being virtually suppressed in many academic circles.

Subjective elements are clearly present in the competing methodologies of Fisher and Neyman for understanding procedures. Both are dependent on assumed models, and neither has a clear operational analysis of sources and justifications for model choice, thus leaving much up to informal subjective judgment by practitioners. Both theories have further ambiguities about choosing a procedure even given a model, and neither copes well with the demands of large contemporary data bases that require ever more complex models. Logicist inference does, however, pay for its more inclusive purview by introducing a further type of subjectivity, namely, a formal logic of uncertainty that applies to specific analyses. Whereas Fisher's logicist statistics implies formal subjective analysis in aid of ordinary informal scientific discourse, Neyman rejects this limited type of subjectivity and believes that frequentist and behavioral theory can substitute. This constitutes the underlying issue of substance in their disputes.

I argue that "formal subjective probability" ought to be recognized and accepted as part of normal statistical science. Not only are Bayesian prior and posterior probabilities formal subjective probabilities, but sampling probabilities are equally interpretable as formal subjective probabilities that inform users about the uncertainties in prospective sampling processes. Fisher's writing is consistent with the perception that he thought about sampling distributions in these terms. Correspondingly, for him, Bayesian inference is not a distinct methodology, but something to be used when the requisite prior model assumptions can be given empirical sources, similar to those traditionally understood to underlie sampling models. While not rejecting Bayesian logic, Fisher along with many of his contemporaries thought it desirable and important to develop alternative forms of uncertain inference based directly on sampling distributions. It appears therefore that, while Fisher might be described as frequentist or Bayesian in limited ways, his understanding of roles for statistical theory in science was evidently broader than either narrowly conceived school allows.

Neyman rejected formal subjective probability as used by Fisher in his fiducial argument for assigning limits to an unknown parameter value in a specified sampling model. Because parameter values are conceived as "fixed" in frequentist theory, if one then takes observations as also fixed, one is led to "absurd" statements such as $P(1 < 2 < 3.2) = 0.95$ (Neyman, 1977).[8] It is questionable, however, to reject the formalization of subjective uncertainties about parameter values on principle while logically parallel formal subjective certainties are implicit in scientific applications of formal deterministic models (i.e.,"equations"), and in particular are fundamental in frequentist theory to connect model-based asymptotic properties of mathematical structures with objective facts about real-world frequencies. By concentrating on mathematical theory, frequentists tend to ignore implications for practice of their own criterion of long run validity. For example, I believe it is rare to find a frequentist attempting empirical verification of an assumed long run property associated with a reported frequentist application. Moreover, to claim that a frequentist inference statement has objective validity in practice is to assume that an adopted model holds for a meaningful long run, hence applies to each member of the long run, and most especially to the particular member defining a specific application. These requirements put a strong responsibility on users to know what is meant by the validity of a model in a unique practical situation. However desirable complete scientific objectivity may seem in principle, refusal to admit formal quantification of specific subjective uncertainties leaves a hole too big to ignore in the foundations of statistical science.[9]

The term "subjective" covers diverse activities of problem-formulation and problem-solving carried out by individuals, drawing on memory, making choices and, most especially, attempting to reason about specific phenomena and issues under analysis. Most such subjective activities consist largely of an undocumented and "informal" stream of consciousness. By contrast, the true subject matter of science is conventionally perceived to be the external or "objective" world. Direct cognition of the objective world comes to us through visual or auditory or even tactile mechanisms. Our minds attempt to assimilate and integrate such data with other complexes of knowledge and understanding. As mathematical scientists, we translate parts of our

informal data bases into precise formal elements. In applied statistics, these formal elements consist of mathematical models associated with statistical phenomena. We then apply formal procedures for data selection, experimental intervention and data analysis to quantities defined in the models. Finally, the results are interpreted in the language of the motivating sphere of human inquiry, which in statistics may be as diverse as deciphering genetic codes or analyzing economic policy. The reasoning behind such interpretations seeks to lay bare the essential story emerging from the numbers in relation to the problem under analysis. Such informal "logic" forms a part of the basis of "what we can say."

Formal logic implies a calculus for obtaining certain or uncertain knowledge about facts given knowledge of other facts. It occupies an important corner of the $2 \times 2$ cross-classification of components of science defined by the pairs subjective–objective and informal–formal, namely, formal subjective logic. A special case, deterministic Boolean logic, is routinely accepted as the basis for computing from input facts and a truth table whether complex statements are "true" or "false" with certainty. Boolean principles are in fact ubiquitous, since every equation embedded in a formal model is interpreted to mean that the equality is "true" with certainty in the formal small world described by the model. When several equations are assumed, then the conjunction of the statements is concluded to be "true" with certainty. Regarding statements of uncertainty, however, there has been strong reluctance to use the straightforward extension to probabilistic logic, which is the interpretation of results of formal model-based probability calculations to assign a numerical measure of uncertainty to "truth" or "falsity" of an objective circumstance. Boole himself had no such inhibitions and devoted six chapters of his book to the topic (Boole, 1854). The position taken in this paper is that every probability computed from an accepted formal model of a specific situation has a corresponding interpretation as a formal expression of uncertainty of an idealized hypothetical individual, sometimes referred to as "you," in implied formally represented circumstances. The model is assumed to separate formally known from formally unknown, therefore formally certain from formally uncertain.

Subjective elements are kept separate from substance in most scientific discourse, creating a misleading appearance of objectivity untainted by opinion. It is accepted that progress is guided in part by personal insight and skilled execution, and that referees weigh such qualities when passing subjective judgments for recommending publications and awards. More basic, however, is substantive uncertainty that accompanies all scientific analysis, especially applied statistical analysis. Scientific writing tends to mention uncertainty explicitly only in passing, and rarely makes educated use of technical concepts like chance, or probability, or likelihood. I believe that when statistical scientists in particular fail to make appropriate uses of these tools, including pointing out their formal subjective interpretations, they are evading an important part of their professional responsibilities. The uncertainties encountered in statistical practice are neither wholly subjective nor wholly objective; that is, neither are they entirely contained in the minds of scientists faced with limited information, nor are they universal uncertainties put in place by an all-seeing Creator. All scientists, and especially statistical scientists, need quantitative analysis of specific uncertainties that by nature depend on balancing complementary roles for formal subjective and formal objective elements.

Behind the proceduralist–logicist distinction lies a broader complementarity[10] familiar in professions such as law and medicine between problem-solving by following rules and problem-solving by reasoning from givens to conclusions. Frequentist theory rests on a fiction that the design and analysis of statistical studies can be implemented by selecting and carefully following protocols set down in advance after contemplating long run properties under assumed stochastic models. The dominance of such thinking over Fisherian logicism emerged circa 1930 from the frequentist theories of Neyman and his collaborator Egon Pearson. Throughout this presentation I am insisting that both elements need to be involved simultaneously. Rules cannot be set down in advance, as is especially obvious during the processes of model choice, on which frequentism depends.

That model assumptions are reflective of objective reality yet are not fully part of that reality is at the core of the explanation of modeling that I advocate, and moreover fits well with Bohr's famous statement, "It is wrong to think that the task of physics is to find out how nature is. Physics concerns what we can say about nature."[11] The quote implies agnosticism about whether any model assumptions made in practice represent objective truths or laws, and instead advocates focusing on how such laws define "what we can say" in the form of defensible statements about objective reality. Statistical logicism includes here uncertain statements computed from probability models. How to choose and justify formal models in practice is not solved by this at-

titude, but is placed in a context of real science that fairly balances objective and subjective, and puts aside an operationally spurious concept of true model.

The complementarity of formal and informal elements of science is deeply but often invisibly embedded in modeling practice. Formal representations of the external world are so basic to statistical sciences, as to all quantitative sciences, that routine technical discourse rarely separates them from their manifest informal counterparts. In statistics, we objectify formal "units" and "variables" that give rise to fundamental data "arrays," whether observed or unobserved. Similarly, formal representations of "knowing" or "inferring" values of postulated quantities are implicitly accepted by practitioners of all philosophical persuasions. If a deterministic "law" is represented by an equation, as in a typical linear or nonlinear model, then applying the equation in a specific context assumes a relation among variable values that holds exactly and supports a precise model-based statement. By contrast, Neymanian proceduralism advocates exclusion of statements of probabilistic logic, such as an assertion that the mean of specific unknown population values is, say, less than 100 with probability 0.7 given specified evidence. On this contentious issue, logicism follows Laplace's conception of what we can say. While nature may be deterministic and therefore perfectly predictable to an all-seeing intelligence, fallible humans, including scientists, need formal probabilities for quantifying uncertainty.[12] A frequentist prohibition on such subjective interpretations of probability is neither credible nor necessary.

Logic itself has both formal and informal sides. Ultimately, scientific discourse is presented in terms of natural language that defies in its complexity reduction to formal mathematical laws, so is informal. Nevertheless, in objective situations that are simple enough to be fully captured by precise forms, logic also has a familiar representation in terms of formal Boolean logic that is generally accepted as noncontroversial. Textbook computations regarding chance phenomena would seem to be equally acceptable instances of probabilistic logic. Indeed, every use of a formal probability model in specific circumstances carries with it connotations of probabilistic reasoning, implying statements of uncertain prediction given understood circumstances. Since the models are formal, so is the logic. It is evident that R. A. Fisher thought of statistical techniques as conveying uncertain knowledge through several varieties of formal probability statements,[13] some of them quite new in form such as the direct interpretation of likelihood.

Implicit in this viewpoint is an idealized formal robot "you," who produces formal inferences for the statistician (Good, 1950, or Savage, 1962). "You" is of course only a tool of an actual flesh-and-blood scientist who must learn to translate statements from "you" into the natural language of ordinary science. These translations are at the core of statistical practice and are implicit in every report. But being subjective, they are mostly passed by in silence, coalescing the formal and informal, and hiding an important aspect of statistical analysis.

A rather crude obstacle to granting subjectivity a rightful place at the table derives from confusion of scientific subjectivity with radical informal subjectivity of an "anything goes" variety. The latter is no more than a straw man waiting to be knocked down, since the word "can" in the phrase "what we can say about nature" implies controls inherent in the scientific method. In particular, it is difficult to accept the political incorrectness of formal subjective probability, wrapped as its applied statistical realizations must be in formal rationality and scientific responsibility. The fundamental reason for recognizing subjective elements is simply their pervasiveness in all sciences, including especially statistical sciences. Every statistical analysis requires a lengthy list of choices by the practitioner, such as the choice of a small world to be formalized, and choices among procedures for data selection, experimental manipulations, data analysis, model construction and inference. The very term choice implies a subjective role.

The task of questioning Neymanian orthodoxy passed from Fisher to the neo-Bayesian movement that gained a minority foothold in the anglophone statistical world around 1950. In actuality, all mathematically trained statisticians, including Fisher, Neyman and both Pearsons,[14] recognize the logical force of the Bayesian paradigm, and only question its applicability in specific situations. Fisher in particular had a broader understanding of the role of probability in applied statistics than typifies his more ideological Bayesian successors in opposition to Neyman's decision-theoretic frequentism. For Fisher, Bayesian posteriors were simple and natural expressions of formal subjective uncertainty, as were $p$-values based on sampling distributions, with the latter involving the interpretation of subjective probability after the fact (Dempster, 1964, 1971).

The scientific world that statistics faces has changed since Fisher's time, drastically and permanently, especially regarding the complexity of the statistical phenomena we are both able and asked to address. In the astonishing worlds of present

and future computing technologies, there should emerge a broader range and synthesis of statistical technologies for modeling and inference than either Bayesians or Fisher appear to contemplate. I suggest that frequentist theory, as the primary evaluator of procedures, should and will gradually recede into history, as practice leads to experience and confidence with probabilistic inference and related modeling techniques that are becoming technically feasible.

A recurrent theme of my presentation is that formal model creation is a pivotal component of practice lying between data analysis and inference. Although the term "model-based statistics" suggests the existence of a variety that is not model-based, the formality of all statistical analysis implies assumed formal structures (in effect, models). Only the extent of probabilistic overlay is optional. That assumptions may be wrong and hence misleading is a half-truth obscuring the need for trading off, usually by informal judgment, between benefits and costs of assumptions. Scientific development is cumulative and depends on a pyramid of assumptions, tentative, approximate or even unrecognized. At a minimum, no statistical formalization can proceed without choices among objective features to represent or to omit, implying an assumption that elements left out are unimportant for purposes at hand and do not excessively bias conclusions and become costly errors of practice.

The many contemporary statisticians who place "data analysis" at the center of their working universe[15] are content to leave statistical modeling at the level of empirical fits to data (Tukey, 1977). In doing so, they in effect opt for a primitive logic of asserting that a simplifying empirical model is an adequate representation for purposes at hand, implying that formal estimation and formal assessment of prediction errors are secondary. This circumstance often holds, leading many scientists to do their own statistics, often well. It is nevertheless wise for statisticians to assume a need to assess differences between quantities computed from observations and unavailable quantities that the computed quantities are thought to approximate. When the latter are essential objects of investigation, formal procedures for assessing how much they differ from empirical quantities are inescapable, creating an opening for formal logic based on formal models incorporating unknown quantities of interest.

When formal inference is attempted, formal probability models are inescapable. Model choice involves Fisherian inference, with much backing and filling (Box, 1980). When tentatively adopted, models become an essential formal ingredient for connecting what is observed to what is not and may never be observed. Thus models are the basis for getting beyond limited information directly computable from observations.

## 2. WHAT "IS" A MODEL?

"Model" is used here interchangeably with the awkwardly long "mathematical model." The long form draws attention to abstract or purely mathematical content, while the short form suggests a type of replica, here a formal representation of objective reality through a corresponding mathematical structure. The term model implies, in addition to the abstract structure, a defined set of connections of the structure to the objective world, conveyed in part by names given to entities in the model, together with descriptive material on time, place, sampling and experimental methods and common understanding of scientific situations as expressed through informal language and thought.

There are differences in how the term model is used, between statistics and parallel mathematical disciplines, or between statistics and many sciences and professions where statistical phenomena receive formal analysis. In mathematical statistics, usage is typically generic, as when a sequence of Bernoulli trials is described as a model for coin tossing, with no specific sequence of coin tosses defined, and typically specifying a generic number $n$ of tosses. This happens with good reason, since it is part of the power of mathematics to make assertions of wide generality and potential applicability. On the other hand, a model in applied statistics would often be regarded as intended for specific objective circumstances, as when a model is constructed to represent the survival times of a defined sample of patients in a specific medical study. Generic usage is appropriate when contemplating the power of mathematical abstraction to analyze many different instances of a type of phenomenon through a single formal structure, while specific usage brings with it the considerations required to evaluate complementary processes of judging acceptability in a specific instance. "Model" in this paper generally implies an entity applied with specificity.

Computers are more and more connected with mathematical models. A computer model is a translation of a mathematical model into a computer program that mimics the application of an abstractly identical model to an objective situation. Computers are often used to create repeated artificial realizations that simulate multiple copies of the same

abstract model, where the instantiated values assigned to free variables vary across realizations, as, for example, in various resampling methods current in statistics. The term "model run" usually signifies a simulated copy of a model that mimics the behavior of the modeled process over a period of objective time. It is often of interest to compare actual values taken from a real-world scientific application with values from one or more model runs. Alternatively, such simulated copies can be used to study features of a model, either to assess its abstract properties or to understand its wider implications. Computer models are gaining in importance as the need to work with complex mathematical structures grows and far outruns the ability of traditional mathematical analysis to provide adequate qualitative and quantitative summaries of model properties.

In applied statistics, saying what a model "is" means describing the constituent parts and how they fit together. The major parts can be described as "small world descriptors" and "relations." More familiar terms for the same pair are "data structures" and "laws." Examples of laws are "equations" for deterministic relations, and "distributions" or "measures" for probabilistic laws. Traditionally in statistics, the concept of model has been almost synonymous with probabilistic relations, especially assumed families of probability distributions, while the categorization of the types of data structures over which the relations are defined receives less explicit discussion, usually being perceived as simple enough in any application to be left as understood from context. Given the steadily increasing complexity of routinely encountered statistical applications, choices of both appropriate descriptors and relations can only assume an ever expanding place in practice.[16]

In some fields the term "data" can mean any input to a computer, such as a program or empirical data, whereas in statistics "data" typically is restricted to the latter. Moreover, statistical data generally implies the presence of one or more varieties of repetition. The most common statistical data structure is the $n \times p$ "data matrix" whose $n$ rows are associated with repeated examples of the same type of object or statistical "unit," such as people, stars or corporations, and whose $p$ columns are associated with "variables" or attributes of the specific unit type, such as gender, weight or wealth of a person. Each of the $np$ elements of a data matrix is associated with a possible real number that codes the value of the attribute for that unit. I say possible because the hypothesized value of any such matrix entry may be objectively determined, or undetermined at the time of analysis, and may be subjectively known or unknown to the idealized formal statistical analyst "you" hypothesized to be adopting the model in a specified situation. Often regular arrays do not exist, as when certain attributes are defined only on units of a special type, thus necessitating the use of conditioned data structures, and so on to ever more complex data types.

In statistics, whether a datum, say, the value of a specific attribute for a specific individual, is observed and/or known, or unobserved and therefore unknown, is fundamental. In fact, the use of the term datum (something given) for an unknown quantity is contradictory, but nevertheless widely used, as in the phrase "missing data." I use "statistical data structure" to cover both available and unavailable "data," as in representing a sample and the population from which it is extracted. Other structure-enriching and situation-dependent complexities of data type include the simultaneous use of several kinds of replication, such as repeated measurements in time, or repetitions across 1D, 2D or 3D physical space, or such as hierarchies of unit types, like people as employees of firms, and firms organized by categories of business. The immediate goal is not to present a systematic catalog of statistical data types, but only to raise consciousness of the importance of a necessary foundation for the logic of applied statistics that remains underemphasized. Topics more commonly taught in computer science, such as the theory of relational data bases, are needed by statisticians as well.

Three varieties of relation appear separately or jointly under the common heading of statistical model: (a) an empirical relation, (b) a stochastic relation and (c) a predictive (deterministic and/or probabilistic) relation. The typical instance of variety (a) is a smooth or regular mathematical form fitted to a representation of some aspect of an empirical data set and regarded as a substitute representation that approximates the original in the sense of mathematical closeness, while summarizing and simplifying the meaning of the original through a form characterized by only a few parameter values. Most commonly, both the data representation and its approximating empirical model counterpart are presented graphically, often in color, so that the analyst can visualize both representations simultaneously and make an informal assessment of goodness-of-fit of the model to data. Fitting a curve such as a line or a spline to a 2D scatter of points, or a surface to a 3D or 4D scatter, are prevalent forms of empirical modeling, with color and perspective often used to add extra dimensions to 2D pictures. An older staple of statistical practice is an empirical fitted "probability" distri-

bution that approximates an empirical distribution, such as a normal density function superposed on an empirical histogram, or a Weibull survival function fitted to a Kaplan–Meier survival curve. Empirical modeling is a subtype of exploratory data analysis, intended to provide suggestive insights into objective phenomena, such as suggestions about what formal models to adopt as bases for formal reasoning, but in itself only able to feed directly into informal arguments and judgments.

Stochastic relations are products of judgments that observable values generated by objective statistical phenomena are analogous to the outcomes of a specified game of chance. For example, given a specified population of a million objects, any subset of 1,000 may be called a sample of size $n = 1{,}000$, and if I take certain precautions designed to validate in advance a prior belief that every possible subset is equally likely, in the sense of an informal judgment to that effect, then statistical modelers will often concur that the sample is a "simple random sample from the population." This terminology conveys a precise formal probability model with a sample space consisting of all possible subsets and a probability measure assigning equal numerical probabilities summing to unity across all possible subsets. The game analogy refers to an urn model where one draws 1,000 balls from an urn containing one million balls, the drawing being fair in the sense that every subset of 1,000 "has" (read, "can be said to have") as good a chance of being selected as any other.

The standard mathematical term for a function defined over a probability measure space of possible realizations is "random variable." The term is also natural and appropriate for hypothetical applications set up as vehicles for the mathematics. For example, if an urn contains red balls and black balls, or a population consists of persons infected and uninfected with HIV, then the count of red, or infected, in a random sample is properly termed a random variable in the mainline sense of credibly obeying a stochastic relation while having a mixed subjective–objective interpretation for the logicist applied statistician. A random variable has an associated probability distribution that describes implied uncertainty about the unknown value of the variable. In a dichotomous data structure such as red–black or infected–uninfected, the simple random sample assumption (or hypothesis from a model-critiquing standpoint) implies a hypergeometric or approximately binomial distribution for the random count. In the sampling context, the distribution is called a sampling distribution. Excepting small minorities of statisticians who would restrict applied statistical uses of the formal mathematics of probability to empirical modeling, stochastic models, and in particular random variables and associated distributions, are virtual hallmarks of applied statistics, and with good reason, since they are bases for all probabilistic inference from known to unknown, such as from an observed fraction of HIV-infected persons in a random sample to the corresponding unknown fraction in the sampled population.

In mathematical communities that study advanced mathematical properties of probability measures ("probabilists") and corresponding probabilistic properties of statistical procedures ("mathematical statisticians"), statistical applications and the associated awkwardnesses of matching abstract entities to objective reality commonly recede into implicit and subconscious formulaic patterns. It is understood that probabilities have connections to scientific uncertainty, but the nature and specification of the connections are easily left to nonspecific understanding because they are not directly involved in the hard work of the mathematics. In statistical practice, by contrast, one needs to convey either implicitly or explicitly the uncertain situations that "you" associate with each formal subjective probability in the model.

For example, probabilities associated with a random sampling hypothesis or with an associated sampling distribution, such as the distribution of the number of red balls in a random sample of 1,000, are interpretable as quantifying uncertainty about the outcome of a prospective sampling process. The interpretation of these probabilities after the sampling has taken place, and in particular after the number of red balls in the sample has been observed to be, say, 531, is rarely addressed, and when addressed is usually a source of confusion and controversy. Students of elementary statistics learn by a combination of formula and computer inquiry that the probability of 531 or more red balls from an urn containing 50% red balls is approximately 0.027, and given the observation 531 many are taught to report the value 0.027 as a "$p$-value." There is little misunderstanding about the interpretation of 0.027 before the sampling takes place, assuming that it had occurred to someone to make the computation before an observation was made. But almost continually since Jakob Bernoulli late in the 17th century started to compute sampling probabilities and implicitly connected them to inferring the unknown population fraction of red balls in the urn, or in his explanatory example the fraction of survivors in a population, there have been questions raised about the logical relevance of sam-

pling probabilities after the data are collected and recorded. I return to this question below.

There are close parallels between empirical models and stochastic models, in the sense that the same formal mathematical structures can appear in both guises in a single application. For example, when a normal density function is fitted to an empirical histogram, it is sometimes a natural accompaniment to regard the empirical data as an attribute of a random sample from some population, and thence to regard the observations underlying the histogram to be observed values of independently and identically distributed (i.i.d.) normal random variables. The empirical model is determined by fitted values of parameters called mean and variance, while the same quantities can be reinterpreted as estimates of the population mean and population variance under a more elaborate data structure that identifies a defined population from which the sample units are randomly drawn. There are many other examples of empirical models associated with stochastic models, such as regression models, ANOVA models and stationary time series models associated with normal (or "Gaussian") distributions, and a very large vista of non-Gaussian and/or nonlinear models, both extant and remaining to be developed.

Before proceeding to the more general concept of predictive models, I digress to discuss senses in which empirical and stochastic models are objective and subjective. In the background is a widespread value-judgment that science ought to be objective and that subjective elements are to be viewed with suspicion if not outright rejection. A parallel theme in the mathematical sciences is that mathematical representations, when carefully defined and communicated, are largely objective in their intrinsic abstract properties, a claim in which I concur, although the objectivity of purely mathematical objects is distinct from the objectivity of real-world natural and social entities, and like all forms of objectivity exists only as perceived through the subjective minds of corresponding literati. Of course, the applied scientific validity of a specific mathematical model based on a copy of an objective abstract mathematical structure is something else again, and begs serious questions about meaning, to be taken up later. A comment here is that the data structure parts of most mathematical models that are accepted as valid statistical models are relatively uncontroversial, whereas associated assumptions of formal deterministic and probabilistic relations often raise disagreements over acceptability, for example, over whether the subjective logical consequences of the model are sufficiently consistent with the available objective evidence. Both the data structure and the relational parts of a model are deliberate human constructions, and therefore have essential subjective aspects.

The recognition that all aspects of statistical practice, including empirical and stochastic modeling, depend on sequential choices by human analysts coexists uneasily with beliefs that science is devoted to the discovery of objective truth. The latter may be a worthy goal, but it is unreachable, as is especially obvious in statistical studies. Our models come and go more frequently than in most sciences and are often controversial. An explanation of the impermanence of formal models is the impossibility of fully coping with complex reality. Empirical data analysis evidently makes only modest claims regarding scientific truth, mainly in the form of suggestions about good ways to look at empirical facts in support of tentative steps along a tortuous path of discovery. By contrast, stochastic modeling enters science through traditions much closer to the physicist's ideal that nature is governed by true laws with precise mathematical expressions. In many branches of physical science, it is traditional to believe that nature is governed by precise deterministic laws and that the random or irregular appearance of many phenomena is due to the presence of many simultaneously interacting individually unobservable mechanisms, or sometimes to instabilities of chaotic nonlinear systems. Stochastic modelers often declare an unknown quantity, such as the next day's stock market, to be a "random variable," nominally signifying not only complex unpredictability but actual belief in the objectivity of chance mechanisms. Clearly, however, random phenomena cannot be both objectively stochastic and objectively chaotic.

Philosophical questions concerning the existence or nonexistence of stochastic mechanisms in the objective world have important interactions with applied statistics. If stochastic mechanisms truly exist, then it would be appropriate in practice to seek true stochastic models and to regard everyday imperfect models as approximations to corresponding unknown true models. In this case, the concept of a model error would mean what it says, namely, a difference between an adopted model and its associated true model. An alternative view is that, try as we may in our daily experience of statistical applications, we are unable to locate in the objective world meaningful evidence of the alleged stochastic mechanisms that produce assumed random outputs. Even the motivating examples of gamblers' games of chance whose simplicity led in the first place to the mathematics of probability cannot be supported with evidence that truly random mecha-

nisms of tossing and shuffling exist in the objective world. I believe that this alternative view is almost always required for a viable working philosophy in practice, and hence that a statement that a model is wrong is operationally meaningless and should be avoided. In place of positing model error, one needs to compare models as being more or less successful for their intended purposes, including accurate representation and trustworthy inference.

The language of objective randomness, whether stochastic or chaotic or both together, appears to render superfluous the complementary language of subjective probability, specifically, the interpretation of probability as quantifying the subjective uncertainty of the hypothetical formal analyst "you" that accompanies like a shadow the small world data structure of a formal model and performs the logic implied by the relations assumed in the model. Again, my position is that proscription of a subjective component is a mistake, even as I maintain that complementary objective elements are essential to balanced understanding, in particular regarding the importance of objective inputs to the determination of scientifically acceptable stochastic or chaotic models. Of course, one can choose to offer only the recipe and not the execution, as much of the procedure-laden literature of statistics appears to do, but at a cost of ducking responsibility for the quality of formal model choices and interpretations, and thence responsibility for the quality of translations of formal elements into informal interpretations.

Given that models cannot be reliably identified with objective reality, a complementary backup anchor in "what we can say" is suggested here as a necessary feature of formal models. Reflecting this position, Adrian Smith has commented that "a model is essentially a predictive machine for observable quantities" (Smith, 1995). I would add a friendly amendment to include quantities that may not be observable in practice but nevertheless are defined within a credible data structure. Also, the machine analogy is apt because a machine is made from objective parts and is intended to interact with elements of an objective external world, so encourages recognition of complementary objective aspects of what a model is.

Predictive relations, both deterministic and probabilistic, are meant for interpretation given specific states of knowledge. In the case of deterministic logic, say, of the simple Boolean variety, one is given certain facts and can deduce whether other facts are true or false by straightforward computation. Similarly, in Bayesian statistics, the device is to compute marginal posterior probabilities by computation from model and data. In Dempster (1998b) I argue that both cases illustrate a single methodology of model formulation followed by logical computation. An important remark here is that interpretation of these predictions assumes that at the time of prediction "you" has precisely the evidence postulated by a predictive model and observed statistical data.

A stochastic model is formally a particular subvariety of predictive model, in the sense that a presumed formal analyst possessing precise values of free parameters in an assumed model is empowered by the model to make uncertain probabilistic inferences about assumed outcomes of the game of chance or its scientific analog. The "parameters" of a stochastic model typically have an ambiguous status. Given the universality of subjective interpretation, parameters are always instruments of logic, as are the probabilities they help to specify. Just as contemplated numerical probabilities may have direct counterparts in the objective world, for example, as empirical frequencies, adopted parameter formalizations may be similarly related to empirical quantities, like population means. Also like probabilities, however, parameters need not have an objective status. Modern Bayesian theory (Bernardo and Smith, 1994) teaches that a basic source of parametric models is to be found in the assumption of exchangeability, so that parameters need only be technical constructs in a stochastic model, having a role in logical computation, but not assumed to have a value locatable in an objective data structure. Of course, it helps both to interpret and assess values for parameters to have sources in long run behavior of actual systems under study, but this is not always possible in complex situations where repetition is limited.

The predictive value of stochastic models is weakened by the reversal of roles of parameter values and outcomes, since in the context of statistical analysis the former are unknowns, while at least some of the latter are observables, thus confounding the normal logic of probabilistic prediction. Fisher with his postdictive interpretation and Neyman with effective denial of interpretation represent alternative attempts to finesse the difficulty. In any case, once fixed by observation, a random variable is in no sense variable, so might better be called a quantity. The problem of how to make predictive use of an assumed parametric family of sampling distributions, or more generally of parameterized stochastic models, is deferred to later writing.

In summary, a statistical model attempts to represent objective reality through a data structure and an associated set of relations. While mathematically

similar, the relations are of three different types, with each type providing a different version of "what we can say" about an empirical case study. The conceptual richness of statistical science derives in no small part from the variety of types and associated uses of relations.

## 3. REASONS AND REASONING BEHIND MODELS

Activities undertaken by a statistical modeler are associated with goals of the modeler. Three such goals are (1) approximation, (2) explanation or understanding and (3) prediction. There are three major traditions of statistical modeling: (i) the data analyst's approach of fitting empirical models to data, (ii) the applied probabilist's approach of associating stochastic models with classes of phenomena and (iii) the subjectivist statistician's approach of constructing formal relations that represent the uncertainty of the idealized formal analyst "you." There is a natural tendency among proponents of one or other tradition to emphasize a preferred goal over the others, and to link that goal to an associated modeling tradition, sometimes viewing alternative approaches as misguided. Violations of Bohr's principle of complementarity inevitably follow.

Adherents to schools (i) and (ii) see themselves by and large as objectivists, but differ in fundamental ways. For example, the first group tend to assume that modeling is applied to a primary set of data and is aimed at representations that exhibit simple recognizable structures, and thence are seen as teasing out messages in the data through deliberate data reduction and approximation. The second group are more phenomenon-driven than data-driven. They pay homage to objectivity by speaking as though stochastic models actually underlie observable phenomena in the external world. Under (ii) it is acceptable to say that a constructed model approximates the true model, adopting language that provides cover for those wishing to avoid the appearance of subjectivism. Modeling is a de facto subjective skill practiced in both schools behind screens of informality. By contrast, logicist modelers accept the centrality of formally representing uncertainty via formal subjective probability. Unfortunately, a subclass of Bayesian subjectivists courts ridicule by ignoring the intrusions of objective evidence into model construction, and have long raised the ire of objectivists who accuse them of using prior distributions to approximate mental states of uncertainty that are plainly poorly determined in the objective world. Meanwhile, objectivists themselves often deserve to be called to account for unfounded assumptions of objective stochastic mechanisms that equally plainly lack external existence. Modeling is sometimes regarded as primarily a task for subject matter specialists, but in most fields requisite knowledge and understanding of statistics remains thinly spread.

The situation is complicated because at an informal level we certainly do have mental states of uncertainty, and there should be no bar to attempting to formalize these states via precise mathematical representations, even if consistency of opinions, by itself and without supporting empirical evidence, lacks validity for interpersonal use, hence can be inappropriate when used for public decisions. Equally, there are important mental illuminations from informal contemplation of formal stochastic mechanisms even when their objective existence is a fiction, because they provide suggestive explanations of empirical phenomena through analogies to familiar games of chance. My own view is that all three schools have virtues that coexist easily under the principle of complementarity, whereas insisting on the exclusiveness of a single approach leads to untenable distortions of normal processes of scientific discovery and analysis. Complementarity implies life with a more complex outlook. Benefits include putting aside unnecessary and often foolish controversies.

Modeling takes place in a medium of informal reasoning, much of it invisible to readers of scientific reports. Accordingly, students and other observers of the statistical scene sometimes have the impression that "the model came in the door with the data."[17] The relative invisibility of processes of model choice contributes also to a perception that clever choice of a robust procedure can reduce or eliminate model dependence. While suspicion, criticism and improvement of specific models are parts of healthy science, in the end, estimates and consequent decisions must be made despite model dependence that remains nonremovable because the objective evidence required to remove it is not available. Statistical analysis must involve defining what is meant by model error and its associated risks, and do so in the face of objective circumstances of increasing complexity.

It may be helpful to separate modeling processes into "early" and "late," meaning those controlling the initiation of formal relations into a study, and those controlling the processes of modifying and refining model choices. The "early" strategies of category (i) empirical modelers appear to be dominated by the use of descriptive or exploratory data analysis procedures whose outputs may suggest smooth approximately mathematizable forms for data sum-

maries or data decompositions (Tukey, 1962). There is a danger in this approach of overinterpretation, or attaching substantive meaning to accidental and therefore misleading appearances. The term confirmatory data analysis has been coined (Tukey, 1977) for Fisherian tests of significance aimed at reducing false appearances. The reasoning here is not simply one of applying off-the-shelf tests and automatically rejecting when test statistics deviate from null expectations too far to be consistent with chance. The use of tests of significance as a part of "late" model evaluation requires sensitivity to processes by which interesting features were noticed in the course of data manipulation, often by an informal reaction to a display. Subsequently, there needs to be another informal judgment as to how selection of the thing noticed is acceptably reflected in the null model.

Being primarily phenomenon-driven, the applied probabilist's approach to modeling is sharply different from that of the empirical data analyst. In early stages, it may be driven scarcely at all by examination of specific data sets. The emphasis is on identifying important features of complex systems, such as biological units or trajectories of physical objects in space, then deciding from scientific understanding what features, both hidden and observable, merit formal representation by stochastic relations and finally using stochastic models, such as point processes or linear systems with random perturbations, to construct tentative relations among the quantities set out in the data structure.[18] Stochastic modeling is often practiced in association with a frequentist viewpoint that seeks to avoid the interpretation of specific probabilities.[19] Frequentism is a seductive theory with valid aspects that proponents support with enthusiasm, apparently untroubled by equally obvious limitations. Under frequentism, realizations of a system subject to stochastic modeling are viewed as randomly selected from an ensemble of possible realizations, whence probabilities in the model adopted for a specific realization encountered in practice are interpretable as relative frequencies calculated over the ensemble. The upside of the half-truth embodied in frequentist ideology is that quantities interpretable as probabilities are ordinary objective quantities obtainable in principle from simple counting. In practical terms, many probabilities and expectations appearing in stochastic models can be approximately obtained from empirical data to an acceptable degree of accuracy, and in some situations can be obtained exactly from symmetry assumptions accepted a priori. The downside is that when a frequency-based probability is subjectively interpreted relative to a specific unknown, the assumption of an ensemble of objective situations implies exchangeability of the specific instance with those hypothesized or identified as the basis of frequency probabilities. Like all assumptions, this one involves risks. For example, frequency probabilities involve dangers of substantial losses if used uncritically for bets, lest opponents have other information that can be exploited by choosing sides to bet on at quoted odds. Repetitions and associated relative frequencies are defining characteristics of statistics, not of particular statistical philosophies.

Rather than accept that scientists must pick and choose among different varieties of formal probability, I prefer to operate with a unitary concept that integrates the basic features of the major candidates for separate theories. While empirical frequency is unquestionably a leading source of approximate numerical probabilities, it is equally true that no empirical frequency can be acceptably linked to a situation-specific probability without a subjective judgment that symmetry among the elements counted is an appropriate feature in the objective situation where probability has the meaning of predictive uncertainty. Only by suppressing a natural symbiosis can one neatly assert that objective probability "is" frequency while subjective probability "is" degree of predictive certainty and therefore something quite separate.

In the sphere of modeling for statistical analysis, Bayesian practice deviates at least rhetorically from the path mapped by a unified theory. The formulation of Bayes separates the modeling tasks of constructing a stochastic model for observables to be used in computing the likelihood factor in the posterior density, and constructing a prior density that constitutes the closing factor. Much effort among Bayesian statisticians is devoted to "eliciting" prior probabilities from the heads of experts or teams of experts. The unified theory suggests that empirical sources are as important for prior probabilities elicited from experts as they are for stochastic models. It is arguable that the strong non-Bayesian inference schools of 20th century statistics would not have developed had frequency sources of prior probabilities been typically available. For science, the use of experts is valid, but only if the sources of objective evidence that gives meaning to the experts are also available to the modeler. The Bayesian algorithm should be regarded as a means of combining separate and independent (Dempster, 1998b) sources of evidence. Unless the sources are informally identified, the formalizing modeler cannot propose formal representations, nor judge whether the sources overlap or interfere.

An account of the origins of predictive models begins with informal identification of features of a data structure judged essential to motivating real-world questions. Since the identification process begins at a nonformal level, it uses the descriptive language of ordinary nonmathematical sciences that is translated by the modeling process into limited mathematical representations. The data structure of a model so conceived is designed to accommodate formal representations of the selected features, and the set of predictive relations erected on the structure reflects what is formalized as known about the features, in both deterministic and probabilistic senses. The standard features mentioned already are units, variables, times and places. There are also specialized feature types that appear repeatedly in many statistical models. One of these is "measurement error." Many observable quantities are routinely represented as true values of corresponding objective quantities identified in a data structure, but it is almost never the case that the correspondence is exact. Either we are virtually sure, as in the case of continuous-valued variables, that some differential, however small, exists between true and measured quantities, or, as in the case of discrete measures such as counts, even when complete precision can occur, a strong possibility of error remains. The basic modeling decisions are then whether to increase complexity by including a formal representation of error in the formal data structure, and, if so, what formal representation in terms of a stochastic model can be made to quantify prior uncertainty about the "errors." These choices are guided by judgments about the consequences of the model error arising because something perceived to exist is ignored in the formal model. If the measurement error is small enough that practical differences between including it or not are insignificant relative to ultimate uses of the model, as shown by comparing analyses with and without, then the analyst is justified in omitting formal representation. Invoking this principle is rarely easy or automatic, since information about the size of the measurement errors is involved, and the objective bases of such information may be strong or weak, and in either case may come both from immediate data and from external sources. The necessity of tentative judgments explains why a model choice can never provide more than a tentative prediction machine with a subjective operator.

Repetition is the lifeblood of statistical modeling.[20] For example, while the error made by a measuring instrument in a particular instance may be crucial in a specific context, the concept of measurement error associated with an instrument cannot be modeled without consideration of repeated use of the instrument, or a population of similar instruments. That is, the population of errors from repeated use of the instruments, both realized errors and hypothetical future errors, is fundamental to thinking about model-building. Historically, the first scientific applications of stochastic models arose in the two areas of modeling measurement errors and modeling sampling hypotheses, the canonical example of the latter being Jakob Bernoulli's binomial sampling model for human survival, where again repeated sampling is of the essence. In modern statistics, two or more levels of repeated sampling are commonly built into a single model. In hierarchical models, inference about a population distribution is generally based on observations of variation across repeated draws from the population, and on observed variation across repeated draws of populations from a superpopulation of populations, the latter making possible adjustments of within population inferences, such as shrinking toward superpopulation means. Through time series models it is possible to achieve predictive power, such as for predicting next month's unemployment rate from current and historical rates using repeated observations of how similar time configurations have fared historically as bases for predictions. Assumed "exchangeability" among repetitions is the most fundamental originator of statistical models.

A second major force behind model origination is the recognition of mechanisms. Much informal understanding of scientific phenomena, or of the state of the world more broadly, derives from beliefs that things are the way they are due to the operation of "causal processes" that describe how things work (Dempster, 1990). Opinions differ on whether uncertain causal effects require a version of probability theory all its own.[21] My own view is that a separate theory of causal logic is superfluous. Causation per se seems to have no explicit representation in formal statistical models, but causation is manifested empirically through informal understanding of mechanisms whose features and workings should be incorporated whenever possible into formal models. To model causal processes it is necessary that the data structure be rich enough to capture the variables through which the processes operate, and then relations among causes and effects can be captured through deterministic and probabilistic relations.[22] For example, in specific contexts, we can understand why observational errors arise, such as through vagueness of a questionnaire or malfunctioning of a piece of equipment, or through inattention of a human data collector, and we can in principle build models of small worlds that

formally incorporate the features of such mechanisms. We may choose not to construct such models, however, due to a sense that we have too little knowledge to formalize relations in a way that can contribute to improved inferences beyond what is implied by a simpler stochastic model. But often we can have enough knowledge of physical, biological or social mechanisms to justify explicit representations in models. Or we may be sufficiently convinced of the importance of certain mechanisms that we will provide them in the data structure even when available relations are too weak to lead to effective logical inferences. In such situations, the outcome of analysis is a conclusion that the objective bases of knowledge representation are too weak to sustain formal assessments. Examples are discussed in Section 4.

Since the finished set of assumptions that make up a model are, or ought to be, accompanied by a history of features considered essential, such as empirical sources and alternative forms considered and sometimes discarded, there is more to model criticism and revision than checking fit to data. More fundamentally, the issue is whether information embedded in the totality of evidence considered supports or contradicts answers to questions that were, or ought to have been, considered along the way. Were all essential aspects of the phenomenon, and especially of known operant mechanisms, adequately represented? Do the assumed formal deterministic and probabilistic relations of the model adequately reflect the evidence available?

Outputs of statistical procedures that assess fit of data to models are also useful inputs to the informal judgments that addressing such questions demands (Box, 1980). The question of fit is essentially one of prediction. Does the predictive machine of the model when applied to relevant features of internal or external data provide predictions sufficiently close to observed features? Testing fit of models is one of the issues addressed by statistical inference (Dempster, 1998b). Several related issues are involved. What makes an appropriate measure of fit, and how many different measures can be safely assayed in a specific example? Should the evaluation be through interpretation of tail-areas of sampling distributions ($p$-values), or interpretation of likelihood ratios or interpretation of Bayesian relative odds of competing models embedded in a formal supermodel? What does it mean to make a test sensitive to failure of particular features of a model, and is it necessary to have "alternative hypotheses" waiting in the wings before implementing a formal logic of testing? Such questions are addressed at length in the research literature.[23]

## 4. EXAMPLES

### 4.1 Preliminaries

A discussion of three statistical applications of importance both to statistics as a discipline and to society at large may serve to illustrate concrete situations where a greater emphasis on formal modeling could reduce confusion and controversy. The U.S. decennial census presents challenges that are special to U.S. laws and mores. The census engages many professional statisticians within the Bureau of the Census and is an ongoing concern of many users as well as research statisticians in academia, including some who have participated in adversarial legal proceedings on opposite sides of the "adjustment" question. Second, statistical issues that concern screening for life-threatening diseases are largely the province of biostatisticians and epidemiologists, representing technical strengths in both statistics and medicine. Lastly, statistical analysis is key to judging whether global surface temperature will continue to rise on the time scale of a few future generations, and whether observed climate changes on a range of spatial scales are "caused" in part by a systematic "greenhouse" effect from the buildup of atmospheric $CO_2$ and other trace gases. Statistical questions concerning global climate change are actively pursued mainly by a few small groups of full-time researchers, mostly with primary expertise in the physical sciences, especially atmospheric and oceanographic sciences. On the other hand, empirical modeling techniques such as multivariate principal components analysis and time series spectral analysis are heavily used in many kinds of climate investigations. Opportunities for interactions between quantitatively sophisticated specialists (such as demographers, medical researchers and climate scientists) and academic statisticians are numerous and widespread, especially for explicit and disciplined development of stochastic models and associated formal inference procedures.

In each of the three very different areas of science, involving different subspecialties of statistical expertise, the dominating issues of statistical practice are traditionally posed as choosing effective statistical techniques. The ultimate goal of framing logical statements about the specific objects of investigation, such as the specific true numbers of people in various geographical locations and demographic categories on Census Day, is often buried in a welter of differing opinions about the relative merits of procedural choices. When scientific conclusions involve uncertainty in a fundamental way, however, we need formal probabilistic reasoning to inform and defend

subsequent informal uncertainty judgments. Probabilistic reasoning in turn depends on model assumptions. To ensure good statistical conclusions, there is no substitute for inference-wise statisticians in active roles, not simply as technical advisors for procedural choices, but also to connect formal uncertainty assessments with meaningful answers to substantive problems.

Another feature of the illustrations to follow, as with much socially relevant applied statistics, is that the science does not take place in isolation from often conflicting political, societal and economic interests. All the more need exists, therefore, to create safe havens where considered and unprejudiced discussions can take place, free from opportunistic and one-sided argumentation of adversarial proceedings. Responsible professionals need to debate contentious issues among themselves, aiming to put a consensus before the public, where the consensus should include a formulation of what remains unknown, and why, as well as what has been resolved. For progress in this realm, the statistical profession requires greater consensus on broad principles, as well as involvement in the science.

The two basic elements of a model defined in Section 2 are a specification of data structure for a suitably inclusive small world, and a logical structure of hypothesized formal deterministic and probabilistic relations among unknowns. Formal modeling of objective phenomena should be carried out in a spirit of attempting to gain scientific consensus on these choices. Subsequent processes of formal logical deduction, such as Bayesian inference about relevant unknowns, are primarily problems of computation. At first encounter, the computations may appear forbidding, but they are mathematically straightforward and can be expected to yield to technical advances in logical and numerical computing. A more demanding requisite for changing professional attitudes is willingness to face up to reexamining deeply ingrained habits of thought, including open-mindedness about recognizing formal subjectivity as a guide to accurate informal understanding and reporting of uncertainty.

Clarity in the presence of both objective and subjective elements requires precise usage of technical language in relation to specific applications. A symptomatic illustration is provided by the terms "estimate" and "estimator," where the former is simply a number that represents a guess at some quantity, while the latter refers to a procedure that is designed to be used repeatedly and is conventionally evaluated under various sampling distribution assumptions, leading to theoretical properties such as "bias" or "standard error" or other measures of performance. It is fundamental to ask how such "operating characteristics" of a procedure relate to the specific result of applying the procedure. Failure to separate inhibits critical review of the key issue of interpretation of corresponding specific data.

The widely used term "bias" provides an interesting case study. "Bias" is used by statisticians in quite different technical senses. One sense concerns largely informal understanding of phenomena rooted in the objective world, whereas another concerns formal mathematical analysis that addresses the average value of an estimator under many repeated hypothetical applications given a specified sampling model. Thus a sample can be biased in the first sense due to faulty real-world sample selection processes, or an estimator can be biased in the second sense because the assumptions in a probabilistic sampling model imply the existence of mathematical bias. When the former type of bias is present, it can usually be shown to have a reflection in the mathematical concept, but not necessarily vice versa. Thus the standard mathematicization of "bias" can mislead if a statistician identifies "bias" with the sampling properties of proposed estimators. As is generally the case with studies of sample-theoretic operating characteristics, the connection of mathematical results with specific applications is problematic. In a specific application, for example, alternative sampling properties that condition on selected features of the specific application could well be thought to have relevance comparable to that of more marginal sampling probabilities, if not more. Not only is there no uniquely applicable operating characteristic in a specific situation, but statisticians who use Bayesian posterior distributions as the preferred estimation principle are led to a quite different formal prescription for assessing the effects of real-world biasing mechanisms, namely, the difference between a posterior estimate from a model that allows for the mechanism and a posterior estimate from a model that ignores the mechanism.

The moral of the story is that a primary emphasis in any specific application should be on formal modeling, or representation of real-world mechanisms, including sampling mechanisms, after which logicist principles of reasoning about the specifics of the case study, including reasoning from given data, largely prescribe appropriate inferential methods. In particular, discussions of "bias" should start from the identification of objective mechanisms, which are then represented formally via modeling processes that deliberately connect objective phenomena with mathematical idealizations.

To foreshadow the topic of Section 4.2, consider the widely recognized concept of "correlation bias" that was central to the adjustment controversies following both the 1980 and 1990 U.S. population censuses. At issue here is whether a hypothesized propensity to avoid being counted in a "capture" sample, here the original attempt at a full-count census, varies across the population in a way that is correlated with a corresponding propensity of individuals to avoid being counted in a "recapture" sample, here an attempted "Post Enumeration Survey" (PES). Once the selection mechanisms behind the correlation bias issue are recognized, the connection to the fundamental problem of census undercount is apparent, so must be addressed when seeking consensus on a model. The tendency of theoretical statisticians, however, has been to focus on the simple capture–recapture estimator that goes with the assumption of zero correlation, and thus to deflect attention away from model formulation and over to assessing hypothetical systematic long run average error from misuse of the naive estimator, with the ultimate goal of devising yet another procedure that "corrects" the naive estimator for "bias," which in turn requires estimating further unknowns that appear in the theoretical expression for "bias"—and so on. The practical goal should be different, namely, to reach consensus on a model that reflects what is regarded as current knowledge, whence a good posterior inference procedure will be nearly automatic. The statistical "bias" that matters is the difference between a quoted estimate, say computed from a standard formula, and a more considered estimate resulting from consensus on an acceptable model that captures relevant features of the objective phenomenon.[24]

The long run averages that underlie measures of performance of procedures rarely have an objective origin in the real world, so mostly can provide only imagined supports for choices among alternative estimators. It is in any case paradoxical that the rationalization of procedural choices advocated under the inductive behavior theory of Neyman relies on the same principle of minimizing expected risk as does the Bayesian decision-theoretic principle that aims to minimize expected risk a posteriori.[25] What matters is the actual error of an actual estimate, and this can only be assessed formally through a logic of uncertainty that applies directly to model-based assessment of the "true" value of the estimand as postulated in a formal data structure.

### 4.2 Example—U.S. Decennial Census

The decennial census is revered by some virtually as a ritual in the civil religion surrounding the U.S. Constitution. Absent the mystique, the census is more likely to be regarded as a mundane task: to devise an adequate way to count people, as required for allocating citizens to voting units or distributing funds accurately according to legislative mandates. On the contrary, as professionals know well, the counting process is a nontrivial scientific process of social measurement. My goal is not to offer advice on detailed protocols for future censuses. I have little expertise on the practical management of what is a large, complex and exceedingly frustrating operation. My goal is rather to illustrate the concepts of logicist statistics and statistical modeling that were laid out in abstract terms in preceding sections.

To this academic observer, it is surely wrong to be constrained by 18th century constitutional language indicating that the census will consist of complete enumeration. Statistical technologies have moved on, as have technologies for transportation, communication and computation that no one would propose limiting to 18th century standards. A more relevant concern is that standards of statistical science employed in the last few decades of the 20th century may need the overhaul and modernizing entailed by logicist methodology, in order to fit with the rapid changes in parallel technologies. There should of course be formal random sampling designs introduced where they promise acceptable accuracy at a price that cannot be matched otherwise. In parallel, randomization needs to be complemented by formal stochastic modeling and subsequent probabilistic logic specifically directed at the targets of sampling.

The first task of formal model construction is deciding what features of a complex objective world are to make up the data structure of mathematically represented elements. Choices here are controlled by the endpoints of the exercise. The difficulty for the census designer becomes quickly apparent when it is realized that very large variations in spatial scales are involved. For assigning numbers of congressional seats to each state, only total populations within state boundaries are needed, while for redrawing congressional districts within states more localized counts are important. Commercial census users may expect composition and characteristics of households down to small urban tracts. Measurement and sampling strategies that are suitable for one purpose may be inadequate or cost-ineffective for another. For the most detailed demands, households within a hierarchy of mapped political jurisdictions are the most common basis of a frame for locating the residents, although other living arrangements identified by point or areal lo-

cation are needed for parts of the population. One must assume that the stock of such unit-households on Census Day is accepted as a defined concept and is approximately on record to a satisfactory degree of objective accuracy and specificity for the task at hand. Similarly, an inventory of characteristics of individual persons, including age, sex and other variables may be assumed identified and given precise objective definitions within acceptable tolerances of ambiguity.

The data structure needs units and variables that represent processes of observation or measurement, so that individual units have both recorded and true values for observed variables. Similarly, where sampling is used, there are sampled units and unsampled units to consider. The fact of subjective judgmental choices of exactly what to represent formally and what to omit is very apparent. For example, when information is recorded by interview, one can in principle include both observed and observers as related units with recorded identities and characteristics for purposes of representing variation among measurers as well as measured. But such a degree of detail might be deemed impractical except for substudies. The fundamental processes of measurement and sample selection come associated with concepts of "error" that are ubiquitous and multifaceted, as in errors in lists of occupied housing units, errors of omission and multiple counting of persons and errors of omission or mistaken records on individual items. That much serious effort has gone into identifying, understanding and analyzing such errors is clear from available literature, as reviewed in the discussion articles in the August 1994 issue of *Statistical Science* (**9** 458–537). Much of the discussion draws on a huge body of detailed knowledge giving rise to dozens of informal interpretations and judgments. But an overall formal data structure that specifies the limited small world representation of the objective situation under analysis is hard to discern.

The logicist approach of this paper differs most seriously from current practice at the stage of specifying formal relations. Whereas formal data structures are implicit in the conduct of the census, even if not separated out and inscribed in a book of mathematically defined elements, the need for probabilistic relations, especially stochastic models, is typically demoted to the role of tools for the study of procedures for data collection and analysis that are evaluated under various loss functions that are themselves in dispute. Stochastic models are thus allowed a secondary role in evaluating choices among procedures, whereas inferential interpretation of formal probabilities of specific unknown target quantities is largely suppressed. Protocols for data collection are extended to protocols for reporting and are regarded as necessary for ensuring objectivity, as though science can escape subjective choices both before and after data collection. On the contrary, assumed models having a part to play in data-specific logic must be open to question and revision through processes of empirical data analysis and modeling, including scrutiny after data are in hand. The only way I see to control and contain fratricidal statistical controversies like those accompanying the postmortems on the 1980 and 1990 censuses is to prescribe a period of intense examination after the data are collected, including reviews of specific stochastic and predictive models, followed by a consensus conference in which the status of what is known and what is not known is set forth, including quantifiable assessments of uncertainty.

What I am advocating needs both a reorientation of methodological approach and a massive modeling effort. Since models provide the logical glue binding what is known and what is unknown, stochastic models need to be created for each error type. The key properties of the probabilities in these models are that they must be interpretable as defining the subjective uncertainty of the analyst before the error is committed, and that they can be converted into posterior measures of uncertainty by Bayesian or other formal probabilistic inference methods. The groundwork for stochastic modeling of measurement and sampling processes was laid in the 18th and 19th centuries, and much developed in the 20th. I believe that the current statistical leadership needs to engage in serious debate and reconsideration of fundamentals of probabilistic inference, so that a valuable intellectual heritage may be put to its natural uses.

The dispute over what to report as estimated population counts from the 1990 census came down to choosing between a pair of procedures, one yielding "unadjusted" estimates and the other "adjusted" estimates. An influential tradition in statistics sees virtue in unadjusted estimates that reflects the logical simplicity of reporting only direct counts. A large part of the appeal here is mistrust of technology, especially fear that technology will be manipulated to serve political ends. Biasing mechanisms that cause counting errors are well understood, however, and are known to induce substantial errors in raw counts. Stochastic and predictive modeling are basic statistical methodologies for assessing such errors, and when based on empirical studies have in my opinion much better credentials than either purely subjective guesses at the quantitative effects

of biasing mechanisms or the effective denial that goes with raw counts. As for political manipulation, it occurs because politicians can exploit differences among scientists, and rarely occurs because scientists themselves choose analyses on the basis of political positions.

Adoption of formal uncertainty analysis for error assessment adds complexity, but not without necessity or justification. According to Ericksen, Fienberg and Kadane (1994), the overall national population total estimate by the unadjusted method omits roughly 20 million people and wrongly includes about 16 million, implying a net undercount of about 4 million.[26] Error rates affecting subgroups of the population are evidently of economic and political significance. These are created by differential error rates of both overcounted and undercounted persons among geographical and political regions, together with differential sizes of subgroups, such as minorities, among different regions.

As recounted by Fienberg (1993), the issue of enshrining an official count was settled by a 1993 judicial decision[27] that put to rest legal opposition to a federal administrative decision not to adjust that was originally made by the Secretary of Commerce in 1987. Legal maneuvering had led to a hearing in 1992 that consisted largely of testimony by opposing teams of statisticians. But, for legal reasons, the case turned on a nonscientific issue, namely, the Secretary's decision having been "neither arbitrary nor capricious." The judge opined that the pro-adjustment party had the stronger scientific case, but that with respected experts on both sides he could not find the Secretary arbitrary or capricious.

I agree with the judge's lay opinion on the scientific merits, which means that I favor the positions of Belin and Rolph (1994) (BR) and Ericksen, Fienberg and Kadane (1994) (EFW) over those of the Berkeley-centered anti-adjusters Breiman (1994) (B) or Freedman and Wachter (1994) (FW). In fact, my position may be as far to one side of BR–EKF as B–FW is to the other. While detailed and cogent by their own lights, the B–FW analyses are undercut by an implicit frequentist philosophy that provides no logical basis for formal analysis of specific errors. No actual model can be the objectivists' true model, so they are restricted to formally analyzing hypothetical consequences of hypothetical false model assumptions, while never achieving the construction of formal statements about actual consequences in a specific situation. Their position provides a sure-fire platform for attacking any model-based or loss-function-based analysis, but they have little constructive to say.

The weakness of the case made by the defenders of the Census Bureau's adjustment proposal is reflected in the confusion of principles they adopt. A fundamental problem is that the census staff does not go nearly far enough in the direction of detailed model construction and logical inference. In the end, the thinking behind an adjustment procedure is likely to be Bayesian, but formal Bayesian logic is kept shrouded in mists, one suspects because the essential concept of formal subjective probability is taboo. Another factor is that the academic intellectuals themselves remain partly in the grip of Neyman's philosophy. To be fair, it is appropriate to tread carefully in a census environment where even the parametric model-based foundation common to Neyman and Fisher is viewed with suspicion. Consequently, the prevailing ideology among statisticians is reliance on hopelessly inadequate "design-based" tools that lead in complex situations to an incomprehensible maze of ad hoc primary estimators and estimators of variance components associated with sampling error analysis of the primary estimators.

One of the key features of the adjustment procedure is the device of smoothing used to control the estimation error coming from the separate use of small PES subsamples in each of a large number of poststrata. The smoothing idea has been prominent for 30 years at least, originally going under odd names such as "ridge regression" or "James–Stein estimation" associated with particular formulations or justifications that by now should seem archaic. The simplest and, I believe, the only practically available justification relevant to specific applications is Bayesian, and of course in a Bayesian framework the correct way to smooth is implicit in the model, so the issue of choosing a good smoothing procedure is moot.

Belin and Rolph (1994) and Rolph (1993) make a strong case within the prevailing theoretical ethos to explain and support the official smoothing procedure. This is done by invoking a loss function, conceived as in Wald's (1950) frequentist decision theory. FW criticize the use of loss functions on the narrow grounds that no single subjective choice of a loss function could justify a smoothing procedure. I agree, but the introduction of loss brings with it several more basic confusions. One is the Neymanian conflation of expected loss associated with the long run behavior of a statistical procedure with an actual loss sustained by accepting a specific numerical statistical estimate as a true value. A more basic question is whether the concept of loss has any place in the task of framing logical uncertainty. The negative answer to which I subscribe is that deci-

sion analysis is important in its place, but that the inputs to uncertainty analysis and decision analysis come from logically separate sources. Stochastic models and subsequent posterior probability assessments draw on knowledge and evidence, whereas utilities or losses draw on assessment of values. Only after each is separately formulated is it appropriate to mix them in formal computations. The dominance of proceduralism over logic in statistical thinking has promoted a fundamental misperception that permits values to spill over and partially control uncertainty analysis.

Then Census Bureau Director Bryant (1993) proposed the concept of a "one number" census, as a means to reduce the adjustment controversy and hence save the drain of intellectual resources to legal activities. If based on scientific consensus, the one number census would indeed effectively remove the problem, but there is little likelihood of consensus in the year 2000, even if political pressures were to abate. Bureau planners have recently envisaged partially replacing direct enumeration with a promising form of counting by random sample (U.S. Bureau of the Census, 1996). By allowing a controllable and quantifiable degree of sampling error, one can gain substantially on total error. For example, hypothetically one might plan a 1/9 sample, where by tripling the cost per household one could reduce measurement error sufficiently to obtain acceptable overall accuracy, and still reduce data collection costs by 2/3. At the time of writing, however, a series of news reports[28] details political opposition to any use of sampling. One subheading in large type reads "Money, politics and law, figure in a fight over census methods." Statistical science apparently does not rate recognition. The profession needs to consider how it can establish an image of scientific credentials that are less easily ignored.

Randomized sampling, both for a limited part of an initial count and for postenumeration surveys, seems a modest first step to a cost-effective census in the U.S. context. From my standpoint to the left of BR–EFK, however, it is clear that the profession is far from moving toward logicist interpretation of the products of randomized surveys. No current research effort on a meaningful scale exists that is directly aimed at formal modeling of a full system, including data structures that fully reflect error mechanisms, the specification of stochastic models representing empirical data collection processes, and especially carrying through to detailed inference computations. Only then will we possess well-articulated assessments of what the effects of sampling and measurement errors are at the various spatial scales for which estimated counts are needed. Logicist studies should be directed at comparison of expectations under alternative census designs. Then quite different postdata development is needed to assess errors in raw counts. Evidently, however, the profession needs much rethinking of fundamentals before a serious start on what needs to be done will occur.

### 4.3 Screening for Early Detection and Treatment of Cancer

The heuristic behind screening is an obvious phenomenon, namely, that cures or at least delayed advances of disease can be achieved through diagnosing and treating malignant cancers while the disease remains asymptomatic, thus preventing or delaying more advanced stages that are difficult or impossible to reverse. Interestingly, it is usually difficult to establish the presence of these benefits through statistical studies, including carefully designed and executed randomized intervention studies that follow several tens of thousands of subjects. The methodology of choice is to look for "statistically significant" differences in the end point of disease-specific population mortality between systematically screened treatment groups and a normal care control group. On reflection, given relatively low population incidence rates and the length of followup needed to reach most relevant cases of mortality, it is not surprising that tests based on feasible randomized population trials (RPTs) have low power for assessing group differences.

There are various sensible innovations in the literature, some being tried and others just advocated. A characteristic feature of these methods is that they require enriching the data structure, for example, to include surrogate outcome measures that can shorten followup time, to allow more detailed tracing of the natural history of a disease and its component subtypes and to permit more precise detailing of risk-relevant background characteristics of individuals in the study.[29] Introducing such complexity leads to informal analysis of how, and how much (or typically how little), universal screening might affect subclasses of individuals. So far there appears to have been minimal effort to develop such complex representations into formal models. The practice of avoiding complexity in formal models goes hand-in-glove with the advocacy of keeping analysis simple by reducing the output to a simple test of significance of an overall difference between treated and control groups. I believe that strong arguments deserve to be made to the effect that complexity should be retained and incorporated into formal models and that correspondingly more complex statistical infer-

ence procedures need to be developed and implemented. For example, formal data structures should record which of both treated and control patients were actually screened, when, by what procedures and what the results were. For cases developing the disease, questions of whether diagnosis was from a screen or from symptoms, and at what disease stage, are surely key variables to include in a complex model.

Two ongoing controversies concern whether there should be routine mammogram screening of women ages 40–49 for breast cancer and whether there should be routine chest X-ray (CXR) screening of smokers over age 45 for lung cancer. In the former case, the accepted wisdom appears to be edging toward support for the "statistical significance" of mortality improvements.[30] A consensus conference in early 1997 heard a wide range of opinions on potential benefits, risks and costs of screening women in their 40s, and on the problems and defects of currently available studies such as quality of experimental procedures, choices of screening intervals, length of followup, post hoc age subgrouping and so on. The consensus report advocated leaving the decision to women and their doctors, which produced an uproar, duly followed by political pressure, and then positive screening recommendations from several expert bodies.[31] In the lung cancer case, the situation is more problematic because the major source of evidence comes from three coordinated RPTs started in the 1970s that were designed primarily to test whether "sputum cytology" at four month intervals plus annual CXR might be more effective as a screening tool than CXR alone. In particular, the ambiguous nature of the statistical evidence regarding the value of regular CXR for diagnosing early lung cancer resulted in "no screen" recommendations since 1980 from several professional bodies, even for high-risk smokers. The statistical evidence relating to other cancers, several of which are commonly screened for in routine medical care, ranges from convincing to similarly ambiguous, whether supported by RPTs or not (Strauss, 1997, 1998; Strauss et al., 1997).

Two sets of stakeholders are foreground and background in screening debates. Most prominent in the literature are public health concerns about what is happening or could be happening to populations, and about screening programs that are or could be recommended and implemented, with what benefits and with what associated costs and collateral risks. More implicit are the concerns of individuals who decide to undergo or not to undergo specific medical procedures, and of their physicians who inform, advise and recommend. The causal reasoning and causal inference questions of public health officials and of clinicians have different emphases. The former are directly concerned with causal effects on groups of individuals of a decision to recommend specific screening policies and practices, whereas the latter are more directly interested in behavioral and biological mechanisms operating at the individual level. The contrasting causal processes do not operate in isolation. For example, the screened patient learns the outcome and then takes or declines actions that affect his or her unfolding biological situation, which in turn influences characteristics of screened vis-à-vis unscreened populations. A reason for the often confusing and inconsistent results from social experiments like RPTs is that they do not take place in isolation from individual decisions, such as decisions by individuals in control groups to avail themselves of screening practices nominally enforced in treated groups, thus lessening the contrast between groups.[32]

Several statistical complexities and difficulties of screening RPTs, have been described and named in the epidemiological literature (Morrison, 1992). Suppose that screening begins by inviting "treated" subjects to follow a program of regular screening, while the "control" group is left to normal care with or without advice to seek a standard screening test on their own. It is then anticipated that "cases" will start to turn up in the treated group more rapidly than in the control group, as a result of prescribed screens, while members of the control group who might have been diagnosed had they been screened will typically not be diagnosed until later, whether through a later screening test obtained outside the study or from physical symptoms. As long as the screening program remains operative, the number of positive diagnoses in the treated group can be expected to remain larger than in the control group. This expectation might at first appear paradoxical since it implies a nonnull empirical difference between treatment and control groups even when a relevant null hypothesis holds, namely, that long-term disease progression for an individual patient is the same whether the patient is in the treated group and is diagnosed early as a result of a screen, or is in the control group and diagnosed later. The empirical difference statistic is therefore regarded as subject to a "bias" dubbed in the literature "lead-time bias."

Two other "biases" of a similar character are "length bias" and "overdiagnosis bias." A naive comparison of survival times after diagnosis may be expected to show longer survival for those in a treated group than in a control group, even if one adopts a null hypothesis that early discov-

ery and treatment has no differential beneficial effect, because under the null hypothesis earlier discovery by $x$ days simply adds $x$ days to survival. Such "length bias" is often associated with the phenomenon of the screening tool turning up "pseudotumors" or very slowly growing lesions that would rarely lead to symptoms or serious illness, in which case "length bias" is enhanced by "overdiagnosis bias." Traditional statistical analyses often compute treatment–control differences on standard statistics like survival curves, and then attempt to allow informally or sometimes to correct formally for the effects of these "biases." Often such corrections are questionable, being based on awkward theoretical leaps, for example, ad hoc replacement of unknowns by estimates without accounting for their estimation error.

Modeling is not an easy task. The phenomena being modeled are extremely complex, and even models that appear complex can only begin to capture the main points. For example, techniques of performing and reading mammograms are in flux and are subject to errors and distortions, and these in turn are interlinked to variations in tissue density, type and location of tumor and so forth. Clearly, however, a model will as a matter of course create a time line for each individual, including the natural history of disease (if any) together with medical events including detectability by screen, detectability by symptoms, treatment and posttreatment assessment and disease-specific death. The literature on formal model representations is small and primitive, especially for breast cancer screens.[33] A more sustained modeling effort has been attempted in the lung cancer–CXR case (Flehinger and Melamed, 1994) leading to some simulation-based plausible estimates of averages bearing on the value of screening, to which I return below.

One major argument favoring formal models is that by design they take account of the real phenomena behind each phenomenological "bias." For example, a minimal model would surely include representation of both screen-detected and symptom-detected times of diagnosis, with at most one of these actualized for each individual, whence "lead-time" is formally specified in the model for every study individual. In a sense, there is no such thing as a "bias" here, only an anticipated aspect of the process under study, and not something to be corrected for, but something to be assessed through the model which gives it formal meaning. A second major argument in favor of detailed modeling is that it opens the way to logically satisfying statistical inferences.[34] Here, I believe, the weight of statistical orthodoxy in favor of simple null models and associated $p$-values for testing treatment–control differences has gone sadly wrong. The acceptability of the logic of using $p$-values for scientific judgments, as compared, for example, to adaptations of Bayesian logic, depends very much on context. I believe that the use of $p$-values for the more common randomized clinical trials (RCTs) of treatment efficacy in specified diseased populations has been unthinkingly carried over to the case of randomized screening trials denoted here RPTs, without allowing for a critical difference that affects the logic. Since the history of medicine is replete with scores of examples of treatments greeted optimistically only to be discarded later, a null hypothesis of no effect of treatment is often plausible for an RCT. An orthodoxy that demands strong evidence against the null hypothesis then provides societal protection against false positives. The situation is different with screening trials. Whereas RCTs can provide evidence of effectiveness even when the biological mechanism is unknown, as often happens in clinical medicine, we do understand the principal mechanism in the case of screening trials, namely, early detection and treatment. There is little interest in whether the mechanism is operating or not, since at least in some instances it is almost certain that cures take place that would not have occurred otherwise. The inference task is not conventional significance testing, but rather estimating with uncertainty how many extra years of life can be expected for different classes of persons.

In this regard, the modeling technology of Flehinger and Melamed appears to be very promising. Although the model is admitted to be simple, and the statistical estimation procedures are also quite primitive, the model has been used to obtain statements that are radically different both in nature and in content from anything otherwise available in the huge medical literature on screening. The following is a direct quote of their final paragraph:

> Based on this model and the data collected by the Cooperative Early Lung Cancer group, it was estimated (1) that the mean duration of stage I non-small cell lung cancer is at least 4 years, (2) that the probability of detecting stage I lung cancer by chest radiography is 16% or less, and (3) that the probability of curing stage I lung cancer is 50% or less. Further calculation indicates that if a high-risk population were examined annually with chest radiography from age 45 to 80 years, the mortality reduc-

> tion would be 18% or less. Although this is far from optimum, it could prevent as many as 25,000 lung cancer deaths each year in the United States.

If these numbers are reasonably credible, they have important implications, beside which the debate over proof by significance is a distraction. The authors comment that the epidemic of lung cancer among smokers cannot be ended by present methods of early detection and treatment, implying that better methods of both are needed. As an outsider, my sense is that the numbers do strongly support a finding of benefit from routine screening of high-risk populations. Of course, society also needs to decide what it can afford, whence modeling and estimation that demonstrate positive results cannot be the end of the story. But high priority needs to go toward appropriate statistical research and development in the context of actual RPTs.

### 4.4 Anthropogenic Factors in Climate Change

Scientists are being pressed to resolve questions concerning whether and to what degree anthropogenic factors, especially the burning of fossil fuels, are affecting "natural variability" in space and time. Possible systematic changes include most notably an upward trend in global average surface temperatures, perhaps accompanied by disruptive changes in local temperatures and precipitation amounts, or by alterations in other major variables such as large-scale circulation patterns of the atmosphere and oceans. Another possibility is severe coastal flooding as sea levels rise along with the melting of polar ice caps.

Climatology is a large scientific field supported by institutes and programs in many nations. Climate change research in particular is monitored by a major supranational body, the Intergovernmental Panel on Climate Change (IPCC), whose reports (Houghton et al., 1991, 1992, 1996) survey current scientific knowledge and opinion. Controlling emissions, especially of carbon dioxide, to an extent that might be needed soon to ward off major dislocations 50 to 100 years hence introduces large economic issues and associated political choices, affecting equity across many economic entities both within and among nations. As with the examples of Sections 4.2 and 4.3, slow scientific progress fosters a wide range of opinion among scientists, leading to difficult policy-making, and at times to accusations that scientists themselves are in thrall to special interests.[35]

Since variability and uncertainty characterize the study of climate, statistical technologies are integral to scientific conclusions. Countless papers in atmospheric and oceanographic research journals draw on established statistical techniques, from simple and standard graphical displays and data summaries to more complex methods such as spectral analysis of time series, and principal components analysis, the latter under the name EOF ("empirical orthogonal function") analysis. Other methods less familiar to statisticians are also widely used, such as methods developed by engineers for signal detection, and inverse methods more generally for reconstruction of images and other complex structures from limited data. The use of such techniques in climatology is mainly descriptive, with few references to formal stochastic modeling and inference, whence opportunities and challenges are extensive. Factors working against a change in the statistical status quo are a research leadership in climatology largely uninformed about possibilities, and a limited supply of trained and capable professional statisticians, most of whom seek careers in fields where applied statistics is better established.

The following brief overview starts with a summary of the science of climate change, then outlines statistical needs and strategies for a particular approach called the "fingerprint" method and finally discusses the current state of fingerprint analyses in the climate change literature. Popular views of the global warming threat rest on the "greenhouse" analogy, the idea being that carbon dioxide and other trace gases trap and reradiate some of the heat energy that the earth's surface would otherwise send back to space. As greenhouse gases build up in the atmosphere, so does the reradiation, part of which returns to the surface, leading to global warming. As detailed, for example, by Lindzen (1994, 1995, 1997), the analogy is superficial in its treatment of important complicating mechanisms. Much heat transfer in the atmosphere is due to convective flows and to infrared radiation from clouds. By far the largest component of a predicted greenhouse effect comes from water vapor whose amounts and distribution are controlled by uncertain feedbacks from other atmospheric changes, including sensitive effects from small amounts of water vapor at high altitudes. Interactions and heat transfers between oceans and atmosphere are large, and since ocean currents are massive and much slower than atmospheric movements they can delay the arrival of detectable climate change. Effects of ice, snow and rainfall on heat flows are large, as are effects of plant life on carbon dioxide budgets. Such factors do not imply that global warming is not a threat, but do suggest that the effects of the 30% increase in carbon dioxide that has

occurred since the start of the industrial revolution, and that seems unstoppable before reaching 100% in the next century, may not be easily quantifiable, or even separable at all from natural variability, given the present state of knowledge.

The science of climate change is physical science. As an outsider, my perception is that there are three available types of evidence and associated reasoning. The first draws on description of small worlds extracted from the large complex total system, such as limited aspects of atmospheric circulation, or zonally aggregated energy balances. After description comes the task of understanding and explanation in terms of causal processes. Discovery of phenomena often depends on empirical analysis of small data sets, such as the use of bivariate time series spectral analysis in the case of the Julian–Madden oscillation (Madden and Julian, 1972) or correlation analyses demonstrating teleconnections between El Niño and distant climates (Glantz, Katz and Nicholls, 1991). The Lindzen papers illustrate this first tradition applied to the question of a possible greenhouse effect.

The second type of analysis is represented in climatology by "general circulation models," or GCMs. These are massive deterministic models that advance in time in perhaps half-hour steps, and may be run to create a simulated climate of order of 1,000 years' duration. GCMs aim to provide a meaningful representation of a reasonably complete climate system. While always pushing the limits of available computing power, they can in fact only achieve representations that in many ways remain crude. The most advanced GCMs currently used to study climate change are CGCMs that jointly model atmosphere and ocean at a dozen or more levels and have spatial resolutions down to 150–300 km. The algorithms that generate GCMs are based on equations of classical physics and are close cousins of the numerical models used for forecasting weather. Weather models differ in assimilating extensive new data every few hours, and in attempting to forecast only up to a week or two. CGCMs have difficulty matching slow-moving oceans with faster atmospheres, and generally require ad hoc flux corrections between ocean and atmosphere, engineered to insure simulations with long-term stability. Also, since much climatological variability, such as that of cloud and storm systems, takes place on smaller spatial scales than the models can resolve, further ad hoc parameterizations of the effects of these variables are introduced. The importance of GCMs to the study of climate change comes from runs with experimentally controlled forcings that, for example, match increases in greenhouse gases estimated from the past and expected in the future. These experiments indicate a global warming threat that is real (Houghton et al., 1996). Many traditional scientists believe, however, that GCMs do not capture enough of the relevant physical science to merit being taken seriously. Critical analysis as in Lindzen (1994, 1995, 1997) points to specific model errors, as in the distribution of water vapor, that call into question GCM predictions of greenhouse warming.

A third type of scientific evidence and reasoning consists of statistical modeling and inference, and offers scope for much future development. Statistical analysis of empirical data provides a piece of the puzzle. After all, if the generally accepted annual time series of global average temperatures had not shown an upward trend since 1970 with no sign of abating, then neither greenhouse theories nor GCM experiments in themselves would have prompted a major international effort. Statistical analysis alone fails to provide a definitive answer, and not only for the familiar reason that statistical relations by themselves are inadequate to establish causation. If a statistical model could reliably forecast a global average temperature increase of, say, 3°C over the next 50 years, then it would matter little that underlying physical mechanisms were poorly understood. In fact there are simple statistical analyses (Bloomfield, 1992; Bloomfield and Nychka, 1992; Dempster and Liu, 1995) that show fairly convincingly that plausible stochastic variation can be separated from a positive linear trend over periods of roughly 1860 to 1990, and in the case of Dempster and Liu the stochastic component is even allowed to have nonstationary long memory. The main weakness of these analyses, as with inferences from stand-alone GCM experiments, is model error. The stochastic models in the examples are linear systems. Nonlinear time series models for use with plausible nonlinear physical systems have not yet been developed, and when they are developed we will need to check whether their forecast limits will be sufficiently wide as to cast doubt on positive messages from linear statistical analyses.

In a mature science of climate change as I envisage it, the three methodologies will interact and reinforce each other. For example, traditional physical reasoning suggests features that can profitably be studied empirically in small world representations, and still other features that must be accurately tracked in credible system-wide models, whether deterministic GCMs or multivariate space–time stochastic processes. While statisticians are accustomed to small applications, new stochastic models are needed for the vastly larger and more complex data structures that GCMs routinely

track. Such models are needed in particular for "fingerprint" analysis, whose aim is to match spatial patterns of trend in empirical data with spatial patterns predicted from GCM experiments (Schneider, 1994). How the concept is implemented in the current literature of climate change is sketched below, but first I attempt an idealistic look at how fingerprint analysis might look in mid-21st century, after stochastic models and associated inference techniques demonstrate their potential through wider development and application.

One assumption is that measurement processes will improve and their errors will be better understood. Similarly, GCMs will improve along with understanding of their strengths and limitations, leading to greater convergence of opinion on their value. Statistics will contribute stochastic models both for the unknown true values underlying observed data, and for errors defined as the differences of true and observed, leading to Bayesian posteriors for the unknown true space–time processes whose properties are to be compared to outputs of model experiments. A necessary feature of the models is that they incorporate a clear understanding of what is meant by the effects of forcing factors such as changes in greenhouse gases, and other explicitly introduced experimental conditions such as those representing anthropogenic or volcanic aerosol patterns, or natural variations in radiation from the sun. Since natural climate processes have sensitive dependence on initial conditions, the simple concept of additive models postulating a decomposition into signal plus noise, or similar linear models common in statistics, will very likely need to be replaced, along with the Gaussian assumptions that are the best currently available and computationally tractable stochastic forms. The continuing advance of computing power will make these developments possible.

On a more conceptual level, causal analysis will be strengthened by introducing higher dimensionality into the fingerprint patterns that are matched. A basic principle of causal inference is to render alternative explanations more difficult as the list of congruences between observation and theory increases. One possibility is to include precipitation along with temperatures, and another is to include three spatial dimensions in place of two. In the case of surface temperature, it could be important to include day–night differences, and seasonal patterns, alongside annual averages. Measures of change will be decomposed into changes at various time scales, as in Fourier or wavelet analyses. Similarly, spatial patterns will be represented in terms of a priori coordinate systems such as spherical harmonics whose characteristics can be estimated by pooling over neighboring frequency ranges. Different spatial characteristics over land, sea and ice are likely to be needed.

The current focus of the fingerprint literature addresses development of procedures for "detection" and "attribution" of climate change (Houghton et al., 1996). Detection is defined by climatologists to mean obtaining a statistically significant relation between an observed spatial pattern of change and a predicted pattern obtained from a GCM experiment.[36] The methodology and language here descend from the language used by statisticians who treat statistically significant "effects" from the analysis of variance as scientifically important, even as it is understood that detecting effects in this manner does not imply a causal interpretation associated with the detected correlate. The further step of causal attribution of a significant fingerprint is recognized to require a more stringent test. A recent suggestion of Hasselmann (1997) is equivalent to adding alternative explanatory variables to a regression model and checking for consistency of the nominated causal effect across alternative linear models, an approach familiar in social science statistics. A popular competing explanation concerns effects due to an increasing presence of atmospheric aerosol particles which are generally thought to have a cooling effect that may delay recognition of global changes due to greenhouse gases, a conclusion that is tentatively supported by recent papers. Skeptics, including myself, regard the use of added variables in regression as reflecting too narrow a range of alternative causal explanations to be convincing. There is also the nagging worry over the appropriateness of linear statistical models.

Evaluating a fingerprint study requires close contact with details. In Tett et al. (1996), for example, the observation is a spatial pattern of trends, where trend is defined as a mean over years 1986–1995 less the mean over 1961–1980 for a radiosonde temperature data set covering the years 1961–1995, and projected from 3D to 2D defined by latitude and altitude. A control run of 700 years of a CGCM is used to characterize natural variability by taking 129 staggered 35-year segments and computing spatial patterns as for the observations. Also four equally spaced time points on the control run are used as starting values for replications of an experiment that uses three different forcings, G, GS and GSO, where G denotes GHG, S denotes aerosols and O denotes stratospheric ozone. Signal patterns for each of the three forcings are computed by averaging estimates from the four replications. Visual comparison of the observed pattern with each of the three signals is suggestive of basic similarities.

A computation resembling a nonparametric significance test is carried out showing that a standard correlation between the observed trend and each signal is substantially greater than any of the analogous 129 correlations of control trends with the same signals.

On first encounter, the high pattern correlations are impressive, as are the placements of these correlations relative to a bootstrapped "null" distribution. Note, however, the assumption that the hypothetical world of the CGCM control run resembles how the actual world would behave over a 700-year period with stable forcing. In the actual world, warm and cold centuries over large regions could plausibly yield more interesting trend statistics than the CGCM can provide. Moreover, the bootstrap sample of 129 comes from only 20 nonoverlapping 35-year periods, and even 20 may be large as an indicator of equivalent independent sample size in the presence of long temporal waves, let alone nonlinear regime shifts. A more deliberate and principled approach to modeling and inference is appropriate.

A concept of "optimal fingerprint" has been developed by Hasselmann (1979, 1993) and implemented in recent Hamburg reports (Hegerl et al., 1996, 1997). An equivalent optimization principle is used by Gerald North (Hegerl and North, 1996; North and Stevens, 1998) in Texas A & M reports. Although couched in signal detection terms, the proposal is familiar to statisticians as "generalized least squares." In place of using the correlation coefficient between observed and signal patterns, one performs an equivalent linear regression of observed on signal. Then the identification of a nondiagonal spatial covariance for the error term can lead to a more sensitive measure of association. While such efficiency gains are worth attempting, the important issue is credible formulation of regression models, including specification and estimation of several covariance structures, on which optimal linear statistical analysis depends. One feature of the climatologists's approach that will immediately puzzle mainstream applied statisticians is the use of a control GCM run to characterize the statistical properties of the error term in the regression, in place of the standard statistical practice of modeling the residuals of the dependent variable. I believe that this difference is symptomatic of a need to carry out separate statistical modeling exercises on the empirical data and on the "pseudodata" outputs of the GCM experiments. The data structure with one signal term and one noise term is misleading. There should be an estimated signal from the observations alone, with an error covariance that reflects measurement error and perhaps deliberate exclusion of high-frequency and high-wave-number effects, and an estimated "pseudosignal" obtained from a multilevel factorial analysis of the pseudodata from designed GCM experiments, also with an error covariance deduced from carefully thought out time–space modeling. The task then would be to isolate major components of signal and pseudosignal, such as low order spherical harmonics that capture the spatial variation, and to ask whether these bear similarities that shine through their respective error variances.

I have emphasized the ultimate problem of causal inference. There are more limited aspects of the study of climate change that are worthy of sustained efforts by a statistical project. A prominent area concerns modeling and inference related to basic empirical measures. This is a very large field, including at one end paleoclimatology (Crowley and North, 1991), where proxy measures for temperature create histories as far back as a million years (DeMenocal et al., 1993), and at the other extreme the evolving tasks of coping with huge flows from satellite instruments in real time of many characteristics of atmosphere and oceans, including surface phenomena. In relation to fingerprint analyses of temperature records back to the mid-19th century, a large body of empirical work has been done by dedicated climatologists (e.g., Jones, 1994; Jones, Osborn and Briffa, 1997; Parker, Folland and Jackson, 1995) to recreate worldwide monthly average surface temperature fields from historical records of land stations and reports from ships at sea. Much of this work is concerned with expert assessment of measurement biases. Methods for passing from station data to gridded data involve ad hoc rules for coping with inevitable widespread missingness. A major opportunity exists for the development of stochastic representation and Bayesian reconstruction of unknown true historical fields, including meaningful standard errors.

## 5. DISCUSSION

Over a lengthy career it has been my conviction that the practice of statistical inference needs to transcend the limitations of both frequentist and Bayesian formulations. But it has been difficult to attract attention to the logicist ideas that I ascribe to R. A. Fisher, largely, I believe, because discussions of inference do not peel back the skin hiding basic elements of practice, such as the facts that practice starts from the explicit creation of formal data structures, and that formal subjective probability is a basic tool whose varied uses give statistical inference its unique flavor and power. Hence the

present paper is in part a necessary prelude to an exposition of the present state of inference (Dempster, 1998b). Beyond this, however, statistical modeling is the critical meeting ground for substantive science and statistical technologies, so merits in itself enhanced recognition. One consequence of such recognition may be reorientation of the directions of theoretical enquiry to support the modeling needs of examples like those I have sought to dramatize as major challenges and opportunities.

## 6. NOTES

1. See note 11.

2. See, for example, the early exchanges among Fisher, Neyman and Pearson following Neyman (1935) and the reprise, Neyman (1961).

3. My position like that of Box (1976, 1980) is that sampling distribution inferences and Bayesian posterior inferences have different logical roles in practice and do not imply conflicts of principle.

4. In a coda to Neyman (1967), Bill Cochran writes, "I have often wondered, as I suppose does Neyman, why Fisher seems not to have regarded the power of the test as relevant, although he developed the power functions of most of the common tests of significance" (Cochran, 1967). See also Fisher (1955), where Fisher opines that long run average theory is all right for the quality control needs of the likes of the Royal Navy, but surely not for science.

5. Fisher (1958) describes a specific probability as providing a "well-specified state of logical uncertainty" and when using it "we are making a statement of uncertainty."

6. Neyman maintained that Fisher's interpretations have "nothing in common with reasoning" (Neyman, 1957). For the behaviorist Neyman, inference is "taking a calculated risk." Fisher would have none of this. See Fisher (1956, pages 100–101 and 108–109).

7. I plan to write further about the differences between Fisher and Neyman (Dempster, 1998b). Many statisticians think of Fisher as a "frequentist," presumably due to the natural association between sampling distributions and long run frequencies. But even the most Bayesian among us is likely to accept that frequencies are valued as sources of probabilities in assumed models that express specific real-world uncertainties. The "frequentism" of Neyman is quite another thing, however, being a theory of evaluating statistical procedures from their long run properties. Nor is it correct that Fisher was unambiguously anti-Bayesian. Although some quotes such as the long passage from Fisher (1925) reproduced by Neyman (1957) are quite strong (e.g., "the theory of inverse probability is founded upon an error, and must be wholly rejected"), there is also a caveat in the quoted passage ("except in the trivial case when the population is itself a sample of a super-population the specification of which is known with accuracy"). In a late paper, Fisher (1959) explicitly discusses situations where the use of Bayesian priors is valid, drawing contrasts with other cases where alternatives to Bayes are needed. Similar attitudes on the necessity of an empirical basis for priors can be found in Edgeworth's statistical writing, beginning with the astute article, Edgeworth (1884), that expresses positions very close to those of my paper. See the lengthy discussions of Edgeworth in Stigler (1986, 1989). Edgeworth forms a bridge from the patchwork discussions of the 19th century to the more organized, dare I say ideological, theories of the 20th century.

8. Efron (1998) states that Fisher's fiducial methodology is "generally considered" his "biggest blunder" although a "fertile" one. I disagree about the "blunder" characterization (Dempster, 1998a). It is presumptuous in my opinion to believe that what has lasting value among Fisher's contributions to inference can be folded into Neyman's theories, and that Fisher was otherwise prone to mistakes that he would not admit. See also Dempster (1964, 1971).

9. Erich Lehmann wrote to me recently, "Neyman realized (fairly late) that his notion of frequentist probability was too narrow and accordingly he widened it. There are some interesting questions here which I don't understand well enough to write about at this time." I hope that Erich or others will clarify Neyman's late views.

10. The term "complementarity" was used in a specific semitechnical sense by Niels Bohr, who published about 30 philosophical papers on his concept among a total of more than 100 published works subsequent to his introduction of the concept in 1929 at age 43, originally in relation to the complementarity of wave and particle theories in physics, and later in many areas of science and policy where opposing positions can lead to controversy, and even violence, and yet on careful consideration are reconcilable as different aspects of a deeper unity (Pais, 1991). For example, Laplace taught that objectively nature is governed by deterministic laws while at the same time our subjective knowledge of nature in given circumstances is probabilistic. See note 12.

11. An interesting discussion on the sources of this quote questioning who actually wrote the original is to be found in Whitaker (1996).

12. Theodore Porter points out that while the concept of an all-seeing intelligence is "beyond doubt the most famous of Laplace's ideas," he "had need of this hypothesis . . . to make clear his conception of probability . . . " (Porter, 1986). As Laplace puts it, "The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance. Probability is relative, in part to this ignorance, in part to our knowledge" (Laplace, 1814). Science has moved on, but Laplace's formalization of how to quantify "our" uncertainty remains apropos.

13. See Chapter 3 of Fisher (1956), entitled "Forms of quantitative inference."

14. See note 5. Bayesian estimation was considered part of normal practice by leading statisticians in the early decades of this century, for example, "till better [methods] are forthcoming" (Pearson, 1920). Neyman (1977) credits Churchill Eisenhart, who was a student in London in the mid-1930s, with suggesting that the theory in his lectures "would look nicer if it were built from the start without any reference to Bayesianism or priors." Egon Pearson (1962) suggests that he and Neyman made a conscious decision to move away from the classical Bayes–Laplace formulation because the required priors were so rarely available.

15. In his Royal Statistical Society presidential address, Nelder (1986) referred to "the cult of the isolated data set" and commented that "much statistical expertise is displayed to make inferences from a single isolated data set." This together with the "lack of emphasis on problems of combining information" constitutes "an unsatisfactory feature of much statistical writing."

16. "In order that the applied statistician be in a position to cooperate effectively with the modern experimental scientist, the theoretical equipment of the statistician must include familiarity and capability of dealing with stochastic processes" (Neyman, 1960). Neyman practiced what he preached, putting major effort into modeling and analysis of such phenomena as physical particles (Neyman, 1955) and carcinogenesis (Neyman, 1960).

17. Personal communication from David Draper. There may be an original source unknown to us.

18. The mathematics of the main types of stochastic processes developed in the 20th century in parallel with that of statistical methods. Able statisticians descended from a British tradition have been adept at combining the two fields, as for example in the work of M. S. Bartlett, D. R. Cox, E. J. Hannan and P. A. P. Moran, among others.

19. Cox (1995) questions whether it is "sensible" to regard a model such as "the Maxwell–Boltzmann distribution" as "measuring uncertain knowledge." Granted, the motivation for much stochastic modeling is the description of aggregates. Still, the richness of the theory of probability rests on analogies between complex systems and simple games of chance. Because the theory can be applied to the play of a specific deck of cards, it can by analogy be applied to the positions of specific unobserved molecules, although I agree we might rarely wish to so apply it in that context. See note 12.

20. Repetition is also the lifeblood of physical modeling, as when Navier–Stokes equations governing fluid dynamics are repeatedly invoked at all points in space–time.

21. The banner for a separate probabilistic logic of causation is at present carried by Judea Pearl (e.g., Pearl, 1997). Earlier attempts can be found in Good (1961–62) and Suppes (1970). Pearl's effort is associated with the use of "graphical models," a natural connection since causes rarely operate alone and often have series and parallel structures. On the other hand, "chain graphs" and similar structures are also advocated for empirical modeling of data (Cox and Wermuth, 1996). Theoretical and computational aspects of graphical models are currently in a period of rapid development.

22. Cox (1990) uses the term "directly substantive" for models that "aim to explain what is observed in terms of processes (mechanisms), usually via quantities that are not directly observed and some theoretical notions as to how the system under study 'works'." His example concerned random bursts of rain from cloud cells.

23. A recent symposium features discussion papers by Draper (1995) and O'Hagan (1995). Bayes factors are reviewed by Kass and Raftery (1995). Aitkin (1997) and Dempster (1997) propose a different approach based on posterior distributions of likelihood ratios.

24. Similar advocacy of Bayesian modeling of selection mechanisms has been advocated by Zelen (1986).

25. The paradox is that the applicability of Neyman's principle of choice to a specific user in a specific situation evidently depends on the relevance of expectations computed from the hypothesized sampling model to the specific situation, while on the contrary his theory of inductive behavior was designed to avoid the case-specific interpretations of probability that were natural to Fisher.

26. A news report in *The New York Times* May 27, 1997, quotes unnamed census officials saying that "studies found an estimated 10 million went

uncounted while another 6 million were counted twice," arriving at the same 4 million undercount for 1990.

27. Finally upheld by the U.S. Supreme court in March 1996.

28. *The New York Times* May 3, 11 and 27, 1997. In June, the Republican congressional majority attempted to ban sampling as part of the census through an amendment to a disaster relief bill, but withdrew the amendment after a Presidential veto.

29. Examples are the use of surrogate end points as discussed in Day and Duffy (1996), or proposals to move away from crude use of age by decade and identifying menopause as a marker in breast cancer studies (Chalmers, 1993). Improvements such as these require modeling, as formally adopted out by Day and Duffy.

30. Smart et al. (1995) combine results from eight large RPTs and, by excluding the large Canadian study whose randomization has been seriously questioned, manage a $p$-value $< 0.05$ corresponding to a 23% observed mortality reduction pooled over a variety of treatment and control policies.

31. Reporter Gary Taubes gives detailed news reports in *Science* **275** 1056–1059 and **276** 27–28. See also letters in the March 14 and 21, 1997, issues.

32. An ideal experiment might have every member of the treated group being screened on a precise schedule, and none of the control group ever being screened. Neither group follows this protocol, so, as with randomized clinical trials, there is an element of uncontrolled selection present. With RCTs the "intent to treat" analysis is the easy way out, but the science wants to know the effect of treatment, assuming the protocol was followed precisely. In the real world, this requires nontrivial modeling of the selection process. With RPTs, the analogous problem may be harder since a large fraction of the controls may self-screen in various ways.

33. Only Moskowitz (1986), of the papers that I have seen, makes quantitative estimates of lead-time for women 40–49 of 2 to 3 years from then available data.

34. A good illustration of detailed modeling associated with modern Bayesian inference is the study of HIV incidence by De Angelis, Gilks and Day (1998).

35. The key chapter (Santer et al. 1996) of Houghton et al. (1996) on "detection" and "attribution" concludes with the paragraph, "The body of statistical evidence in Chapter 8, when examined in the context of our physical understanding of the climate system, now points towards a discernible human influence on global climate. Our ability to quantify the magnitude of this effect is currently limited by uncertainties in key factors, including the magnitude and patterns of long-term natural variability and the time-evolving patterns of forcing by (and response to) greenhouse gases and aerosols." An earlier draft that was circulated on the Web in April 1995 was more conservative, referring to "large uncertainties," possibly "flawed" noise estimates, and placing a "burden of proof" on the scientists involved. A news report in *Physics Today* of August 1996 describes some of the controversies surrounding the wording of the final report. A news report in *Science* by Richard A. Kerr (Kerr, 1997) quotes a range of scientific opinion ranging from cautious to skeptical, especially regarding whether the science is really in place to justify claims of anthropogenic effects.

36. A recent generalization involves "multifingerprint" (Hegerl et al., 1997) or "multi-pattern fingerprint" (Hasselman, 1997) analysis. These terms mean the comparison of joint space–time patterns not necessarily interpretable as trends over specified intervals such as 30 or 50 years. The linear compounds of space–time variables used for fingerprints and multifingerprints alike are typically derived from principal components analyses whose multivariate covariance inputs are computed by treating successive times as a sample.

## ACKNOWLEDGMENT

Supported in part by NSF Grant DMS-94-04396 to Harvard University and through the Geophysical Statistics Project at the National Center for Atmospheric Research.

## REFERENCES

AITKIN, M. (1997). The calibration of $P$-values, posterior Bayes factors and the AIC from the posterior distribution of likelihood. *Statist. Comput.* **7** 253–272. [Includes discussion by M. Stone of Aitkin (1997) and Dempster (1997), and replies by Dempster and Aitkin.]

BELIN, T. R. and ROLPH, J. E. (1994). Can we reach consensus on census adjustment? *Statist. Sci.* **9** 486–508.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory.* Wiley, New York.

BLOOMFIELD, P. (1992). Trends in global temperature. *Climate Change* **21** 1–16.

BLOOMFIELD, P. and NYCHKA, D. (1992). Climate spectra and detecting climate change. *Climate Change* **21** 275–287.

BOOLE, G. (1854). *An Investigation into the Laws of Thought.* Walton and Maberly, London. [Reprinted (1951) Dover, New York.]

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

BREIMAN, L. (1994). The 1991 census adjustment: undercount or bad data? *Statist. Sci.* **9** 458–475.

BRYANT, B. E. (1993). Guest commentary. Census-taking for a litigious data driven society. *Chance* **6** 44–49.

CHALMERS, T. C. (1993). Screening for breast cancer: What should national policy be? *Journal of the National Cancer Institute* **85** 1619–1621.

COCHRAN, W. G. (1967). Footnote by William G. Cochran. *Science* **156** 1462–1463.

COX, D. R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5** 179–174.

COX, D. R. (1995). The relation between theory and application in statistics (with discussion). *TEST* **4** 207–261.

COX, D. R. and WERMUTH, N. (1996). *Multivariate Dependencies.* Chapman and Hall, London.

CROWLEY, T. J. and NORTH, G. R. (1991). *Paleoclimatology.* Oxford Univ. Press.

DAY, N. E. and DUFFY, S. W. (1996). Trial design based on surrogate end points—application to comparison of different breast screening frequencies. *J. Roy. Statist. Soc. Ser. A* **159** 49–60.

DE ANGELES, D., GILKS, W. R. and DAY, N. E. (1998). Bayesian projection of the acquired immune deficiency syndrome epidemic. *J. Roy. Statist. Soc. Ser. C* **47**. To appear.

DEMENOCAL, P. B., RUDDIMAN, W. F. and POKRAS, E. M. (1993). Influences of high- and low-latitude processes on African terrestrial climate: pleistocene eolian records from equatorial, Atlantic Ocean Drilling Program Site 663. *Paleoceanography* **8** 209–242.

DEMPSTER, A. P. (1964). On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc.* **59** 56–66.

DEMPSTER, A. P. (1971). Model searching and estimation in the logic of inference. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 56–78. Holt, Rinehart and Winston, Toronto.

DEMPSTER, A. P. (1990). Causality and statistics. *J. Statist. Plann. Inference* **25** 261–278.

DEMPSTER, A. P. (1997). The direct use of likelihood for significance testing. *Statist. Comput.* **7** 247–252. [Reprinted from *Proceedings of the Conference on Foundational Issues in Statistical Inference* (O. Barndorff-Nielsen et al., eds.) Aarhus, Denmark, 1974.]

DEMPSTER, A. P. (1998a). Comment on "R. A. Fisher in the 21st century," by Bradley Efron. *Statist. Sci.* **13** 120–121.

DEMPSTER, A. P. (1998b). Logicist statistics II. Inference. In preparation.

DEMPSTER, A. P. and LIU, C. (1995). Trend and drift in climatological time series. In *Conference Proceedings, 6th International Meeting on Statistical Climatology* 21-24. University College, Galway, Ireland.

DRAPER, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 45–97.

EDGEWORTH, F. Y. (1884). The philosophy of chance. *Mind* **9** 223–235.

EFRON, B. (1998). R. A. Fisher in the 21st century (with discussion). *Statist. Sci.* **13** 95–114.

ERICKSEN, E. P., FIENBERG, S. E. and KADANE, J. B. (1994). Comment on "The 1991 Census Adjustment, Undercount or Bad Data" by Leo Breiman, "Heterogeneity and Census Adjustment for the Intercensal Base" by D. Freedman and K. Wachter, and "Can We Reach Consensus on Census Adjustment?" by Thomas E. Belin and John E. Rolph. *Statist. Sci.* **9** 511–515.

FIENBERG, S. E. (1993). The New York City Census adjustment trial: witness for the plaintiffs. *Jurimetrics Journal* **34** 65–83.

FISHER, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh. (Many later editions exist, up to the 14th in 1973.)

FISHER, R. A. (1955). Statistical methods and scientific induction. *J. Roy. Statist. Soc. Ser. B* **17** 69–78.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh. (Slightly revised versions appeared in 1958 and 1960.)

FISHER, R. A. (1958). The nature of probability. *Centennial Review* **2** 261–274.

FISHER, R. A. (1959). Mathematical probability in the natural sciences. *Technometrics* **1** 21–29.

FLEHINGER, B. J. and MELAMED, M. R. (1994). Current status of screening for lung cancer. *Current Perspectives in Thoracic Oncology* **4** 1–15.

FREEDMAN, D. and WACHTER, K. (1994). Heterogeneity and census adjustment for the intercensal base. *Statist. Sci.* **9** 476–485.

GLANTZ, M. H., KATZ, R. W. and NICHOLLS, N. (1991). *Teleconnections Linking Worldwide Climate Anomalies.* Cambridge Univ. Press.

GOOD, I. J. (1950). *Probability and the Weighing of Evidence.* Griffin, London.

GOOD, I. J. (1961–62). A causal calculus. *British J. Philos. Sci.* **11** 305–318, **12** 43–51, **13** 88. [Reprinted (1983) in *Good Thinking* 197–217. Univ. Minnesota Press.]

HASSELMANN, K. (1979). On the signal-to-noise problem in atmospheric response studies. In *Meteorology of Tropical Oceans* (D. B. Shaw, ed.) 251–259. Royal Meteorological Society, London.

HASSELMANN, K. (1993). Optimal fingerprints for the detection of climate change. *J. Climate* **6** 1957–1971.

HASSELMANN, K. (1997). Multi-pattern fingerprint method for detection and attribution of climate change. *Climate Dynamics* **13** 601–611.

HEGERL, G. C., VON STORCH, H., HASSELMANN, K., SANTER, B. D., CUBASCH, U. and JONES, P. D. (1996). Detecting greenhouse-gas-induced climate change with an optimal fingerprint. *J. Climate* **9** 2281–2306.

HEGERL, G. C., HASSELMANN, K., CUBASCH, V., MITCHELL, J. F. B., ROECKNER, E., VOSS, R. and WASKEWITZ, J. (1997). Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. *Climate Dynamics* **13** 613–634.

HEGERL, G. C. and NORTH, G. R. (1997). Statistically optimal approaches to detecting anthropogenic climate change. *J. Climate* **10** 1125–1133.

HOUGHTON, J. T., JENKINS, G. J. and EPHRAUMS, J. J., eds. (1991). *Climate Change*. Cambridge Univ. Press. (The Intergovernmental Panel on Climate Change Scientific Assessment.)

HOUGHTON, J. T., CALLANDER, B. A. and VARNEY, S. K., eds. (1992). *Climate Change 1992.* Cambridge Univ. Press. (Supplementary Report to the Intergovernmental Panel on Climate Change Scientific Assessment. Prepared for IPCC by Working Group I.)

HOUGHTON, J. T., MEIRA FILHO, L. G., CALLANDER, B. A., HARRIS, N., KATTENBERG, A. and MASKELL, K., eds. (1996). *Climate Change 1995. The Science of Climate Change*. Cambridge Univ. Press. (Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change.)

JONES, P. D. (1994). Hemispheric surface air temperature variations: a reanalysis and update to 1993. *J. Climate* **7** 1794–1802.

JONES, P. D., OSBORN, T. J. and BRIFFA, K. R. (1997). Estimating sampling errors in large-scale temperature averages. *J. Climate* **10** 2548–2568.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KERR, R. A. (1997). Greenhouse forecasting still cloudy. *Science* **276** 1040–1042. (16 May 1997.)

LAPLACE, P. S. (1814). *Essai Philosophique sur les Probabilities.* Courcier, Paris. [A 1902 translation by F. W. Truscott and F. L. Emory was reprinted (1951) by Dover, New York.]

LINDZEN, R. S. (1994). Climate Dynamics and global change. *Ann. Rev. Fluid Mech.* **26** 353–378.

LINDZEN, R. S. (1995). The importance and nature of the water vapor budget in nature and models. In *Climate Sensitivity to Radiative Perturbations: Physical Mechanisms and Their Validation* (H. Le Treut, ed.) 51–66. Springer, Berlin.

LINDZEN, R. S. (1997). Can increasing carbon dioxide cause climate change? *Proc. Nat. Acad. Sci. U.S.A.* **94** 8335–8342.

MADDEN, R. A. and JULIAN, P. R. (1972). Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmospheric Sci.* **29** 1109–1123.

MORRISON, A. S. (1992). *Screening in Chronic Disease*, 2nd ed. Oxford Univ. Press.

MOSKOWITZ, M. (1986). Breast cancer: age-specific growth rates and screening strategies. *Radiology* **161** 37–41.

NELDER, J. (1986). Statistics, science, and technology. *J. Roy. Statist. Soc. Ser. A* **149** 109–121.

NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *J. Roy. Statist. Soc. Suppl.* **2** 107–180.

NEYMAN, J. (1955). The problem of inductive inference. *Comm. Pure Appl. Math.* **8** 13–46.

NEYMAN, J. (1957). Inductive behavior as a basic concept of the philosophy of science. *Rev. Internat. Statist. Inst.* **25** 22–35.

NEYMAN, J. (1960). Indeterminism in science and new demands on statisticians. *J. Amer. Statist. Assoc.* **55** 625–639.

NEYMAN, J. (1961). Silver jubilee of my dispute with Fisher. *J. Oper. Res. Soc. Japan* **3** 145–154.

NEYMAN, J. (1967). R. A. Fisher (1890–1962): an appreciation. *Science* **156** 1456–1462.

NEYMAN, J. (1977). Frequentist probability and frequentist statistics. *Synthèse* **36** 97–131.

NORTH, G. R. and STEVENS, M. J. (1998). Detecting climate signals in the surface temperature record. *J. Climate.* **11** 563–577.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparisons (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 99–138.

PAIS, A. (1991). *Niels Bohr's Times, in Physics, Philosophy, and Polity.* Clarendon, Oxford.

PARKER, D. E., FOLLAND, C. K. and JACKSON, M. (1995). Marine surface temperature: observed variations and data requirements. *Climate Change* **31** 559–600.

PEARL, J. (1997). Structural and probabilistic causality. *The Psychology of Learning and Motivation* **34** 393–435.

PEARSON, E. S. (1962). Thoughts on statistical inference. *Ann. Math. Statist.* **33** 394–403.

PEARSON, K. (1920). The fundamental problem of practical statistics. *Biometrika* **13** 1–20, 300–301.

PORTER, T. M. (1986). *The Rise of Statistical Thinking, 1820–1900.* Princeton Univ. Press.

ROLPH, J. (1993). The census adjustment trial: reflections of a witness for the plaintiffs. *Jurimetrics Journal* **34** 85–97.

SANTER, B. D., WIGLEY, T. P., BARNETT, T. P. and ANYAMBA, E. (1996). Detection of climate change and attribution of causes. In *Climate Change 1995. The Science of Climate Change* (J. T. Houghton, et al., eds.) 407–444. Cambridge Univ. Press.

SAVAGE, L. J. (1962). *The Foundations of Statistical Inference.* Methuen, London.

SCHNEIDER, S. H. (1994). Detecting climatic change signals: are there any "fingerprints"? *Science* **263** 341–347.

SMART, C. R., HENDRICK, R. E. and RUTLEDGE, J. H. (1995). Benefit of mammography screening in women ages 40–49 years. *Cancer* **75** 1619–1626.

SMITH, A. F. M. (1995). Discussion of "Fractional Bayes factors for model comparison," by A. O'Hagan. *J. Roy. Statist. Soc. Ser. B* **57** 120–122.

STIGLER, S. (1986). *The History of Statistics.* Harvard Univ. Press.

STIGLER, S. (1989). The role of probability models in statistical inference in 19th century Europe. *Bull. Internat. Statist. Inst.* **53**(Book 3), 157–162.

STRAUSS, G. M. (1997). Measuring effectiveness of lung cancer screening. *Chest* **112** 216S–228S.

STRAUSS, G. M., GLEASON, R. E. and SUGARBAKER, D. J. (1997). Screening for lung cancer. *Chest* **111** 754–768.

STRAUSS, G. M. (1998). Randomized population trials: implications for cancer early detection. Unpublished manuscript.

SUPPES, P. (1970). *A Probabilistic Theory of Causality.* North-Holland, Amsterdam.

TETT, S. F. B., MITCHELL, J. F. B., PARKER, D. E. and ALLEN, M. (1996). Human influence on the atmospheric vertical temperature structure: detection and observations. *Science* **274** 1169–1173.

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67.

TUKEY, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, MA.

U.S. BUREAU OF THE CENSUS (1996). *The Plan for Census 2000.* Economics and Statistics Administration, U.S. Dept. Commerce, Washington, DC. (Revised and reissued February 28, 1996.)

WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.

WHITAKER, A. (1996). *Einstein, Bohr and the Quantum Dilemma.* Cambridge Univ. Press.

ZELEN, M. (1986). Case–control studies and Bayesian inference. *Statistics and Medicine* **5** 261–269.