Last 2 pags give title & copyright
pags of book

Statistical Practice as Argumentation: A Sketch of a Theory of Applied Statistics

James S. Hodges
Division of Biostatistics, University of Minnesota

Abstract

We have theories of statistics and we use statistical methods to solve subject-matter problems. One might expect that the theories would affect how the methods are used, but they do so only superficially, because all statistical theories are quite incomplete as descriptions of and prescriptions for statistical practice. This paper sketches an extension of Bayesian theory that might address this incompleteness. The extension is based on five ideas:

1. The product of a statistical analysis is an argument – not an HPD region, posterior distribution, decision, or other data summary, but the entire argument, including premises and logical steps.
2. Arguments come in several logically distinct types, with an argument's type being defined by the form of its conclusion. The paper catalogs the types of argument and identifies the main burden of each.
3. EDA and model-building activities establish a plausible, tractable baseline argument for a given problem and dataset.
4. Diagnostics and sensitivity analyses vary the premises of the baseline argument and display the resulting variation in the conclusion (as opposed to intermediate quantities).
5. An argument is strong to the extent that:
   - its premises are conclusions of strong arguments, or
   - the region of premises yielding the same conclusion is large.

A simple example demonstrates the mechanics of the extended theory. The example is then generalized to draw implications for statistical foundations, methods, and computing. This paper amounts to a research agenda, so it poses more problems than it solves.

Keywords: applied statistics, Bayesian statistics, diagnostics, exploratory data analysis, foundations of statistics, rhetoric, sensitivity analysis.

1 Introduction.

We have theories of statistics and we use statistical methods to solve subject-matter problems. One might expect that the theories affect how the methods are used and, superficially, they do. For example, Fisherians report P-values and Bayesians do not, at least, not to Bayesian audiences. But the theoretical veneer is thin. When we work in substantive fields, we all do the same things: learn something about the subject matter; on obtaining data, poke around in it and settle on a few models; compute data summaries; do diagnostics; and report conclusions, perhaps after repeated cycles through these steps. The theoretical styles lean toward different data summaries and models, but otherwise, foundational leanings have little affect on statistical activities.

This is not just historical inertia. Rather, it happens because all statistical theories are incomplete as descriptions of and prescriptions for statistical practice. This being a largely Bayesian event, I will focus on Bayesian theory. Section 2 makes a brief argument for the incompleteness of Bayesian theory. Section 3 argues that Bayesian theory can be extended and gives five ideas that frame the extension. A simple example in Section 4 illustrates the mechanics of the extended theory. Section 5 generalizes from the example to draw implications for foundations, methods, and computing. Section 6 discusses direct antecedents and inspirations for the extended theory. This theory is embryonic; the present paper amounts to a research agenda, and I beg the reader's indulgence.

2 Bayesian theory is incomplete.

All of us, Bayesians included, learn from data by means other than Bayes' Theorem, the obvious example being exploratory data analysis (EDA). EDA makes little or no use of the Bayesian formalism and prominent Bayesians have certified this as kosher (Smith [1], Hill [2], Berger [3]). Predictive calibration is also difficult to square with Bayesian ideology, but many Bayesians consider it essential. (Geisser and Zellner are the best-known advocates of this view, but see also West and Harrison [4], Chapter 10.) Bayesian theory brings no special facility to the problem of learning enough subject matter to avoid being a menace. And so on; the examples are easily multiplied.

What might we ask of a theory of applied statistics? Such a theory should name the *activities* of applied statistics and the *products* of those activities. It should provide a *rationale* for doing certain activities in specific situations and not doing other activities. Finally, it should explain *how the activities combine* to yield the products. Although Bayesian theory supplies some necessary language, the next few sections will show by construction that it provides none of these elements of a theory of applied statistics – nor, it should be obvious, does any other statistical theory.

Why is this a problem? For the fastidious, incompleteness is problem enough. For the more pragmatic, it is useful to be able to explain what we

do when we ta
matical world,
about the origi
(or personalisti
some allegiance
an individual's
is to analyze ho
uals. If we swe
in practice.

3 But B

The problem in
connect mathen
proach of axiom
is needed, instea
cal activities and
proposes such a
ideas frame the

1. The produ
2. Arguments
is defined b
3. EDA and m
argument.
4. Diagnostics
gument and
5. An argumen
- its pre
- the reg

These ideas will n

3.1 Idea #1 argumen

The product of a
posterior distribut

do when we take something from a substantive problem, build a little mathematical world, do operations in that world, and then claim to be better informed about the original problem. What is the basis of this latter claim? Subjective (or personalistic) Bayesian theory – to which most of this audience would pledge some allegiance – describes and prescribes how to use data to make changes in an individual's beliefs. With few exceptions, however, a statistician's problem is to analyze how data can and should change the beliefs of many or all individuals. If we sweep this under the rug of informality, we risk a variety of errors in practice.

3 But Bayesian theory can be extended.

The problem in extending Bayesian theory to a theory of applied statistics is to connect mathematical reasoning and verbal reasoning. The time-honored approach of axiomatics will not do, because axioms are within mathematics. What is needed, instead, is a rhetorical structure that rationalizes specific mathematical activities and explicitly connects them to subject-matter issues. This section proposes such a structure as an extended Bayesian theory of statistics. Five ideas frame the extended theory:

1. The product of a statistical analysis is an argument.
2. Arguments come in several logically distinct types. An argument's type is defined by the form of its conclusion.
3. EDA and model-building activities establish a plausible, tractable baseline argument.
4. Diagnostics and sensitivity analyses vary the premises of the baseline argument and display the resulting variation in the conclusion.
5. An argument is strong to the extent that:
    - its premises are conclusions of strong arguments, or
    - the region of premises yielding the same conclusion is large.

These ideas will now be described.

3.1 Idea #1: The product of a statistical analysis is an argument.

The product of a statistical analysis is *not* a highest-posterior-density region, posterior distribution, likelihood, predictive distribution, or decision. Instead,

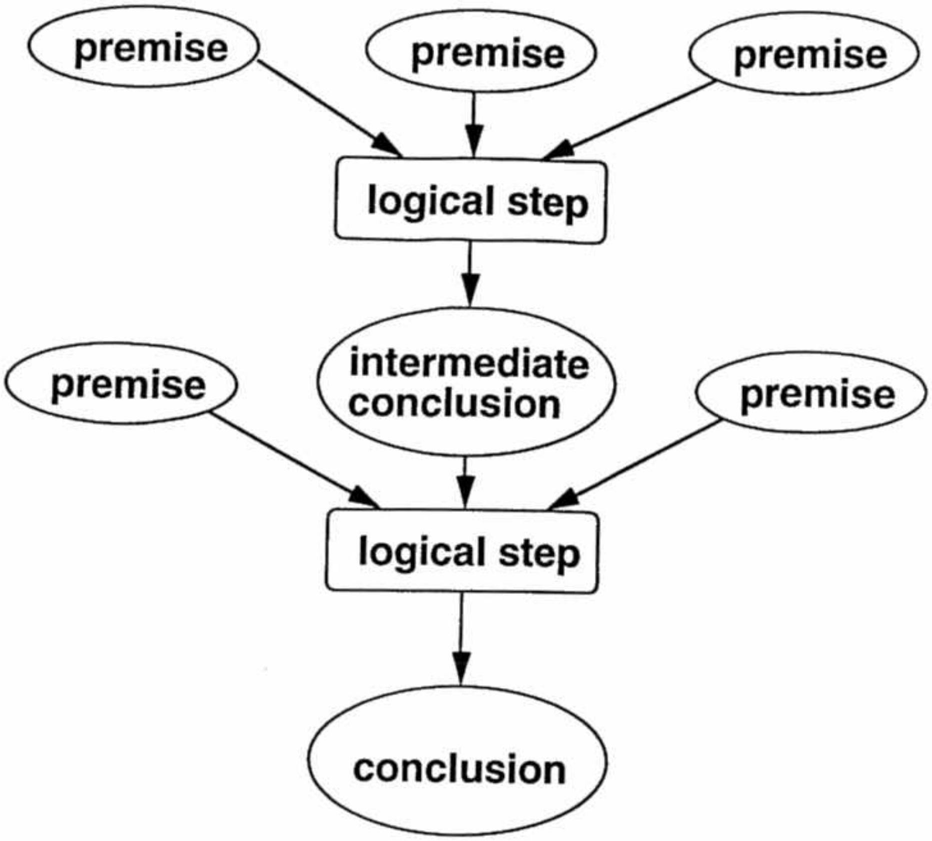

Figure 1: Schematic diagram of an argument

the product is an argument – the entire argument – parts of which may use data summaries like those just mentioned. An argument begins with premises, which are inputs to a logical step, which yields the conclusion. Sometimes premises are the conclusion of an argument that has its own premises and logical step. This is shown in Figure 1.

A premise is "a proposition upon which an argument is based or from which a conclusion is drawn" (*American Heritage College Dictionary*, 3rd. ed.). Statistical models are premises, as are assumptions like exchangeability or missing-at-random. The observed data are also a premise: different data could have occurred but did not, and an argument must use the data that did occur.

Binary logic provides many logical steps combining premises, such as *modus*

ponens: A implie
culus, interprete
Fisher's disjunct
unlikely under a
posulated model
approaches – of
quentist, and pe
premises they us
premises and Ba
clusions are least
the most specific
cause they use c
Further discussio

3.2 Idea
types.
its conc

This subsection c
subtypes. There
of statistics. The
Each conclusion
be replaced by p
types of argumen

- *Causal arg*
  - inferer
  - predic
  - action
    "Takir
- *Non-causal*
  - descrip
  - action:
    come."
- *Description*
  - direct:
  - inferen

ponens: A implies B, A is true, therefore B is true. So does the probability calculus, interpreted as an extension of binary logic. I also include arguments like Fisher's disjunction (Fisher [5], p. 42): if we observe a datum that is extremely unlikely under a postulated model, then *either* a rare event has occurred, *or* the posulated model is not true. In a future paper, I will argue that foundational approaches – of which I count three: randomization-theoretic, model-based frequentist, and personalistic Bayesian – differ in the explicitness and variety of premises they use. Randomization-theoretic arguments generally use the fewest premises and Bayesian arguments the most, while randomization-theoretic conclusions are least specific to the circumstances at hand and Bayesian conclusions the most specific (*cf* Efron [6]). No approach is correct or incorrect; rather, because they use different premises, they tend to produce different arguments. Further discussion of this topic is beyond the scope of this paper.

3.2 Idea #2: Arguments come in several logically distinct types. An argument's type is defined by the form of its conclusion.

This subsection discusses five logically distinct types of arguments, each having subtypes. There may be other types of arguments, but these five cover most uses of statistics. The arguments are defined below by the form of their conclusions. Each conclusion is stated non-probabilistically, but statements labeled B can be replaced by probability distributions in the personalistic fashion. The five types of arguments are:

- *Causal arguments*
  - inferential: "B *was* caused by A."
  - predictive: "A *will* cause B."
  - action: "B *was* caused by A, therefore action C is desirable," or "Taking action C in situation D *will* cause a desirable outcome."
- *Non-causal predictive arguments*
  - descriptive: "B *will* occur."
  - action: "Following rule C in situation D *will* yield a desirable outcome."
- *Description arguments*
  - direct: "We *summarize* the data as B."
  - inferential: "The unobserved constant C *takes the value* B."

  - action: "Constant C *takes the value* B, therefore, action D is desirable."

- *Existence arguments*
  - predictive: "B *has* occurred; therefore, it *can* occur."
  - action: "B *has* occurred, so it *can* occur; therefore, action C is desirable."

- *Hypothesis-generation arguments*
  - data-mining: "We ransacked the data and found B, but we have reason to fear that this finding is spurious."
  - innocuous identification: "We have found B, for which we have no explanation; explanations should be sought."
  - action: "On the basis of these data, we should invest in the study of A to make a more definitive argument."

All five types of arguments have "action" as a subtype. The non-action kinds of arguments will be considered first, followed by the action subtypes.

Before doing so, though, it is reasonable to ask why we should bother with a taxonomy like this. The basis of the taxonomy – and the strongest argument for using it – is that arguments can and should be distinguished according to their burdens of proof. The remainder of this subsection makes the case that the five types of arguments and their subtypes do, in fact, have distinct burdens of proof. Drawing on this material, Section 5 makes the case that the usual taxonomies of statistical problems – for example, inference vs. prediction vs. decision, or hypothesis testing vs. point or interval estimation vs. everything else, and so on – do not identify an argument's burden of proof and thus do not differentiate statistical problems usefully.

3.2.1 Non-action arguments.

Inferential causal arguments refer to the causation of an event that occurred in the past or was determined in the past, while predictive causal arguments refer to events, often counterfactual, that will or could occur in the future. The burden of an inferential causal argument is to rule out all but one causal agent. The additional burden of a predictive causal argument is to show that a postulated causal agent will produce a given effect in specified future situations.

For example, consider BW02, the clinical trial that justified US licensing of AZT to treat HIV infection (Fischl et al [7]). BW02 was a randomized, double-blind, placebo-controlled trial in patients who had had an AIDS-defining illness or who had AIDS-related complex (ARC). Competent use of randomization,

double-blind
caused the d
P an inferent
It is harc
a general po
BW02:

- enrolle
  of *Pne*
  ment;
- enrollec
  no prio
- restrict
  prophyl

It might be p
patients mee
tious predicti
argument th
medications
ple, lengthy
For either
– in this case,
or not at all ir
estimation pr
Non-causa
ture, but the
predictive arg
properties cla
tor autoregres
postulating c
is medical use
shown that in
cyte count is
This is true e
clinical events
[21], Choi et
of proof of a
this case, an
formulation of
Descriptio
scription argu

double-blinding, and placebo controls are the crux of an argument that AZT caused the difference in the number of deaths between the two arms of the trial P an inferential causal argument.

It is harder to make the predictive causal argument that giving AZT to a general population with AIDS will extend their lifespans. This is because BW02:

- enrolled only people with ARC or people who had had exactly one episode of *Pneumocystis carinii* pneumonia (PCP) within six months of enrollment;

- enrolled only people with specific medication histories, in particular, with no prior use of antiretroviral drugs; and

- restricted concomitant medications, in particular, patients could not use prophylaxes for PCP, the most common AIDS-defining condition.

It might be possible to make a predictive causal argument in favor of AZT for patients meeting these conditions, but few such patients exist. A more ambitious predictive causal argument based on BW02 would require a subsidiary argument that AIDS-defining illnesses, medication history, and concomitant medications do not affect the efficacy of AZT. This is most unlikely; for example, lengthy use of AZT degrades its effectiveness (Kahn et al [8]).

For either type of causal argument, the burden of proof is met by a procedure – in this case, the data collection procedure – which is reflected either implicitly or not at all in formulating the statistical problem as a hypothesis test or interval estimation problem. This pattern recurs in the argument types to follow.

Non-causal predictive arguments refer to events that will occur in the future, but they make no assertion of causality. The burden of a non-causal predictive argument is showing that the predictive rule or procedure has the properties claimed for it. One example is economic predictions made with vector autoregressions, which capture relationships in economic time series without postulating causality (Litterman [9]). Another non-causal predictive argument is medical use of prognostic measurements. For example, dozens of studies have shown that in an unselected population of HIV-infected people, CD4+ lymphocyte count is a strong predictor of time to an AIDS-defining disease or to death. This is true even though *changes* in CD4+ are poor predictors of subsequent clinical events in interventional clinical trials (Fleming [20], De Gruttola et al [21], Choi et al [22]). Note that as with the causal arguments, the burden of proof of a non-causal predictive argument is provided by a procedure – in this case, an out-of-sample validation – that is not part of the mathematical formulation of a prediction problem.

Description arguments involve neither causality nor prediction. Direct description arguments also involve no uncertainty or sampling: all of the entities exist and have been measured. For example, certain vendors sell summaries of sales databases gathered by bar-code scanners in consumer-goods stores, mostly grocery stores [23]; in some regions, the sales data are exhaustive. Among other things, grocery stores use these data to answer the question "How am I doing?" compared to nearby grocery stores. The burden of proof in a direct description argument is showing that the summary actually describes the whole for the purpose at hand (Mallows [10], Draper et al [11]). An inferential description is almost a direct description – the entities involved exist and are *potentially* measurable – but some of the entities either were not or could not be measured. The obvious example is a census. Large censuses inevitably involve an inference, as the US census undercount controversy has made clear. Another example is estimating the number of birds killed in an oil spill (Carter, Page, and Ford [12]): some number of birds actually died, but it is impossible to count the number that died at sea and were not washed ashore, and if a census of beached carcasses is not exhaustive, it is impossible to count the number that washed ashore dead or died on shore. Any count of dead birds must be by inference. The burden of proof here is showing that the inferential statement has the properties claimed for it.

Estimating natural constants might seem to be inferential description, but it is not, necessarily. Some constants, such as the gravitational constant in Newton's theory of gravity, do not exist in any meaningful sense. Rather, because the theory is manifestly false, although useful, the gravitational constant is a mere tuning knob, like those on a radio, that is adjusted to make the theory fit the data as well as possible – that is, to facilitate predictions. Arguments estimating such constants must therefore be considered predictive. By contrast, other entities, such as the speed of light or the electron's mass, truly exist and thus admit of inferential description arguments. Estimating a given natural constant might initially be a predictive argument but later become an inferential description argument as the constant's *bona fides* are established.

Existence arguments are have a simple structure: some event *has* occurred; therefore, it *can* occur. The burden of existence arguments is showing that the event actually did happen. In medicine, case series are a popular but much-maligned form of existence argument. For example, Blakeman et al [13] described a series of patients who were on heart-transplant waiting lists because of severe congestive heart failure and coronary artery disease. Such patients have bleak prognoses but because of the risk of death during surgery, they are widely believed to be poor candidates for palliatives like coronary artery bypass grafts. Nonetheless, 17 of the 20 patients in Blakeman et al's series survived bypass grafting and 10 of the 17 were radically more able to perform ordinary activities. Thus, bypass grafts can be done on high-risk patients with good results and acceptable operative risk. The series does not identify *which* patients should be bypassed, but the result is significant nonetheless.

Another e
sometimes ar
AIDS becaus
would have u
is similar in s
keys precisely
has happened
ply that HIV
does undermi
conventional

For an in
be met, consi
HIV-positive
low-dose oral
sero-deconver
investigator h
concluded th
burden of pro
– that are no

The foreg
Suppose, for
unobserved p
a very intere
occurred, the
predictive cau
unobserved p

Data-mini
ticians and o
this: you ran
so on, and fin
searches are
I errors (Tuk
Leamer [15] c
nocuous. Wi
attributed to
mit that data
although the

The tradit
has suggestec
pelling argun
current state
directed searc

Another example indicates the possibilities of existence arguments. It is sometimes argued that the human immunodeficiency virus (HIV) cannot cause AIDS because lentiviruses such as HIV cannot work as the conventional view would have us believe. But the simian immunodeficiency virus (SIV) – which is similar in structure to HIV – has been shown experimentally to do in monkeys precisely what HIV is supposed to do in humans. Such viral behavior *has* happened, in monkeys; thus, it *can* happen. This argument does not imply that HIV behaves in humans in accordance with the conventional view; it does undermine the claim that no lentiviruses *can* work in accordance with the conventional view.

For an instance in which the burden of this argument's proof could not be met, consider Koech et al. [24] and Obel and Koech [25], who claimed that HIV-positive patients in their care became HIV-negative upon administration of low-dose oral α-interferon. At the time, no-one had claimed to induce so-called sero-deconversion, so these two reports were greeted with skepticism, and no investigator has been able to replicate the result. All but a few observers have concluded that sero-deconversion did not, in fact, occur. Note again that the burden of proof for existence arguments lies in procedures – external validation – that are not represented mathematically.

The foregoing notwithstanding, existence arguments are limited in scope. Suppose, for example, that an astronomer uses measurements to infer that an unobserved planet must exist. There is an existence argument here, but not a very interesting one: orbital irregularities (say) for this list of planets have occurred, therefore they can occur. The interesting argument is, instead, a predictive causal argument: the observed orbital irregularities were caused by an unobserved planet, which will cause this list of further observable irregularities.

Data-mining arguments are a loose end which is necessary because statisticians and others have not sorted out the relevant issues. The situation is this: you ransack a dataset, check subgroupings, delete apparent outliers, and so on, and find a nominally significant result. The traditional view is that such searches are extremely prone to spurious findings: they do not control Type I errors (Tukey [14]). Few Bayesians have commented explicitly on this issue. Leamer [15] can be interpreted as meaning that data-mining is essentially innocuous. With somewhat greater possibility of injustice, this view might be attributed to Lindley [16], [17]. On the other hand, Berry [18] and Hill [2] admit that data-mining affects the interpretation of a nominally significant result, although they do not give prescriptions.

The traditional view is a hypothesis-generation argument: this data-mining has suggested an interesting hypothesis; perhaps we can make a more compelling argument using another dataset. It captures the belief that, at the current state of theory at least, a causal argument cannot be based on a data-directed search because the chances of error are unknowable. The contrasting

Bayesian view is that data-directed searches can, in fact, be the basis of a predictive causal argument. An intermediate position is that data- directed searches can be the basis of an action causal argument, if not a predictive causal argument (Berry [18], discussed below).

Innocuous identification arguments present a result, usually from an observational study, refute uninteresting explanations like selection effects, and pose a challenge: "explain this result." For example, Neaton and Wentworth [19] used data on over 330,000 men screened for the MRFIT study between 1973 and 1975, along with the National Death Index and Social Security mortality data, to argue that in men between the ages of 35 and 57, low cholesterol and low blood pressure are risk factors for death from AIDS. Cross-sectional studies have shown that patients with HIV disease who are sicker, by various definitions, tend to have lower cholesterol, but it was unclear whether this was a consequence of the disease or a pre-existing condition. Neaton and Wentworth's result imples that it is at least partly a pre-existing condition and cannot be explained by misclassification of cause of death or by an association of sexual orientation with blood pressure and cholesterol. The result poses a challenge: show how cholesterol and blood pressure are causally related to death from AIDS or show that they are not. Neaton and Wentworth do not assert that raising cholesterol or blood pressure can reduce the risk of infection, so it is not a causal or predictive argument.

For both subtypes of hypothesis-generation argument, the burden is to find the result and to rule out as many uninteresting explanations as possible. Yet again, meeting this burden involves procedures not generally reflected in the explicit mathematical formulation of the statistical problem.

3.2.2 Action arguments.

Each type of argument has an action subtype, because the other subtypes need not imply action and vice versa. For example, one may be able to make a compelling action causal argument without being able to sustain either a predictive or inferential causal argument (*cf* Berry [18]). A living example of this is a colleague who has AIDS and uses an unproven anti-HIV drug, peptide-T. Peptide-T is an analog for the protein in the HIV virus's envelope that binds to the CD4 receptor in human cells; it blocks the binding of HIV to these receptors *in vitro*, and thus has been suggested for treating HIV. Observational studies suggest that peptide-T has a potent effect, but the sole controlled trial of peptide-T is measuring only neurological symptoms. Thus, it is difficult to sustain a predictive causal argument of more general clinical benefit. But peptide-T has no known side effects or interactions with other drugs, and my colleague gets his supply *gratis*. The cost and risk are small and the potential benefits immense, so an argument for action – for using peptide-T – is compelling.

Conversely, it is possible to make a cogent inferential or predictive causal argument without being able to sustain an action argument. For example, no-one questions that aerosolized pentamidine (AP) is an effective prophylaxis for PCP, i.e., it is easy to make a strong predictive causal argument that it reduces the chance of PCP. However, neither does anyone question that trimethoprim/sulfamethoxazole (TMS) is a much more effective prophylaxis for people who can tolerate it (Schneider et al [27], Hardy et al [26]). In this case, the action argument about taking AP ("don't take AP, take TMS if you can tolerate it") requires more information than the predictive causal argument about AP ("AP reduces the incidence of PCP").

Any action argument involves an implicit or explicit tabulation of costs and benefits associated with possible courses of action. In decision-theoretic views of statistics, the tabulation is explicit; in both of the arguments above, the tabulation is mostly implicit, although in each case it could be made explicit. (The value of such an exercise is a separate issue.) Whether the tabulation is implicit or explicit, it adds premises to the action arguments.

For the other argument types, as for causal arguments, a non-action argument need not imply the corresponding action argument, and vice versa. Brief examples follow. Non-causal predictive: Litterman [9] and those following his lead make strong arguments for predictions, not arguments for actions; on the other hand, political and corporate actors must constantly rationalize action by predictive arguments that no well-informed person could find compelling. Descriptive: Counts of birds killed in oil spills have partly rationalized legal judgments against oil tanker companies; but less well-supported population counts have been used to draw inferences about sea-bird population dynamics. (Gross inferences, to be sure.) Existence: When Blakeman et al (1990) showed that patients with severe heart failure could be radically improved with bypass, avoiding a costly and hazardous heart transplant, it became necessary to act by identifying which patients could be so treated. But if someone discovered the remains of Noah's ark, this would confirm that it existed but the standards by which its existence was verified could not sensibly involve a calculus of costs and benefits. Hypothesis-generation: Had Neaton and Wentworth tried to use a specific calculus of gains and losses, it would have had only formal content. But a pharmaceutical company might reasonably use a subset analysis of a clinical trial, producing weaker evidence than Neaton/Wentworth's, to select among possible future trials in promising patient subgroups.

Now that the argument types have been differentiated according to their burdens of proof, a new diagram of arguments, Figure 2, can replace Figure 1. Commonly, researchers approach an investigation with the intention of eventually drawing a particular kind of conclusion, that is, of making a particular type of argument. The desired type of argument determines the mathematical

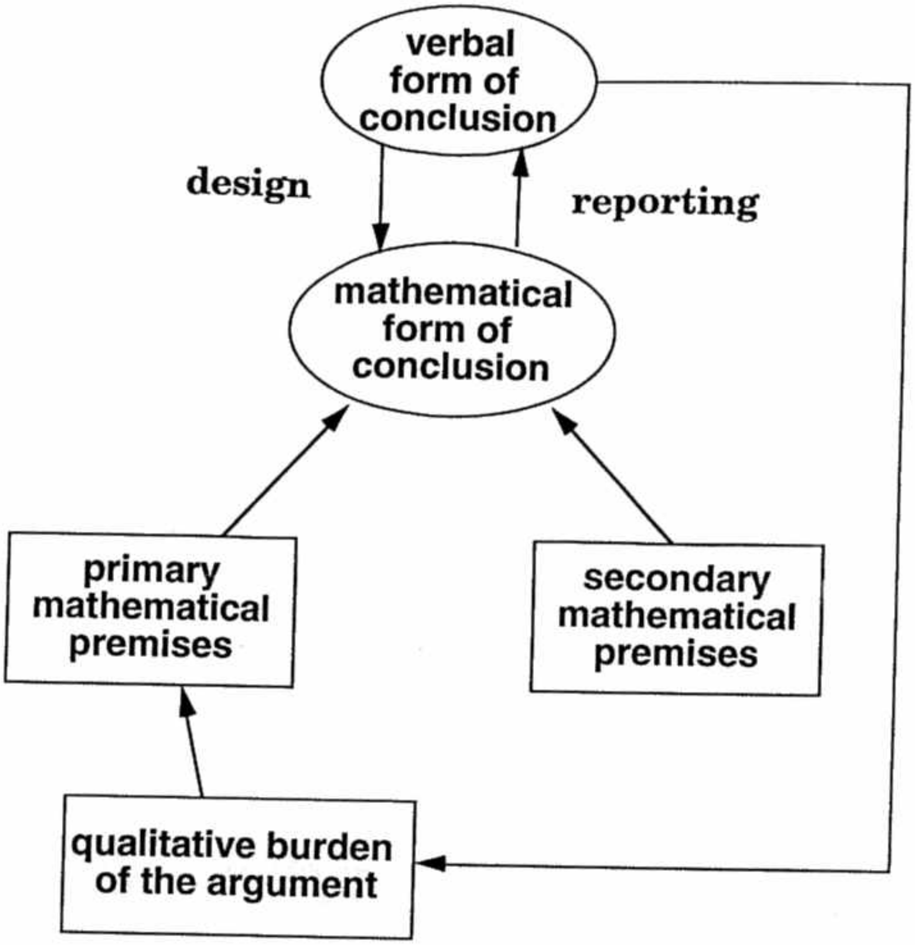
verbal form of conclusion
design
reporting
mathematical form of conclusion
primary mathematical premises
secondary mathematical premises
qualitative burden of the argument

Figure 2: Revised schematic diagram of an argument, showing the relation of the argument type to the premises

3.3 Ide
ble,

In the world
fitting. Firs
data, that i
coding scher
the analyst
for example,
that can be
and some te
not plainly i

But in sc
example – n
are collected
run through
and bridle.

Whether
trial organiz
argument in
it passed th
has achieved
argument to
have little ef
"baseline" c
the extended
report the b
acceptability
not be accep

3.4 Idea
the
resu

This idea ha
choosing the

Measurin
analyses to
induces a lar
p. 115) mea
inverse of its
by the chang

3.3 Idea #3: Model-building activities establish a plausible, tractable baseline argument.

In the world portrayed in data analysis classes, certain activities precede model-fitting. First, the dataset is summarized until the analyst understands it as data, that is, to ensure that data items have not been garbled, to figure out coding schemes, and the like. Then, the dataset is summarized some more until the analyst has a few ideas of what it says about the subject-matter problem; for example, pictures are drawn to reveal trends, groupings, and other features that can be modeled. With luck, the result is a few models or data summaries and some tentative arguments that are tractable, have simple elements, and are not plainly inconsistent with the data.

But in some places where statistics is used – clinical trial organizations, for example – models and types of arguments are often specified before the data are collected. As soon as the data are entered into the computer system and run through basic edits, proportional-hazards models are fit to them like a bit and bridle.

Whether your world looks more like the data-analysis class or the clinical trial organization, what you have at this point is a baseline argument. It is an argument in the sense used above; it is a baseline in two senses. First, because it passed the EDA test (in the world of the data-analysis class) or because it has achieved the status of a convention (in the clinical trial world), it is the argument to which variations will be compared, in the hope that the variations have little effect and the baseline argument can be reported. Second, the term "baseline" commonly implies that attempts will be made to improve on it. In the extended theory of statistics, both senses are important. It is desirable to report the baseline argument because it is often simple and has conventional acceptability. On the other hand, it is important that the baseline argument not be accepted without substantial scrutiny.

3.4 Idea #4: Diagnostics and sensitivity analyses vary the premises of the baseline argument and display the resulting variation in the conclusion.

This idea has consequences for measuring the influence of perturbations and for choosing the class of perturbations to be examined.

Measuring the influence of perturbations. We do diagnostics and sensitivity analyses to see if a small change in the data, model, prior, or loss function induces a large change in . . . what? Cook's distance (Cook and Weisberg [28], p. 115) measures the change in the vector of regression coefficients against the inverse of its covariance matrix. Cook [29] assessed the effect of a perturbation by the change it induces in the parameters of the baseline model, measured by the associated change in the value of the baseline likelihood. Some Bayesian analogs use Kullback-Liebler distance to measure how perturbations change the posterior or predictive distributions (Johnson and Geisser [30], McCulloch [31], Carlin and Polson [32]).

These measures are helpful but we lack the words to describe how they help. What *does* one do if a point is influential? Cook and Weisberg ([28], p. 104) gave little advice, noting that"final judgments [about influential cases] must necessarily depend on context, making global recommendations impossible." They suggested some useful considerations – Is the influential point erroneous? Are more data available near the influential point? – but otherwise, they concentrated on methods for detecting influential cases.

The extended theory has a clear implication here: if a perturbation of the baseline argument has large influence according to some measure but does not change the outcome of the argument or suggest a need to change its structure, then . . . who cares? Influence measures should focus on conclusions of arguments, not intermediate steps.

The method of Kass, Tierney, and Kadane [33] allows this focus: it assesses the influence of perturbations on the posterior expectation of a general class of functions of the baseline model's parameters. The disadvantage of this method is that it measures influence locally, around the posterior mode. However, recent advances in Bayesian computing (e.g., Gelfand, Dey, and Chang [34]) should permit non-local influence measures for general functions.

Choosing the class of perturbations to be examined. Premises of a statistical argument often include exact mathematical specifications. Some specifications are explicit, such as probability density functions, while others are implicit, such as exchangeability assumptions (Draper et al [35]). It is rare to see a demonstration of strong support for the specifics of an argument's premises. More commonly, a case is made that the argument does not depend on the specifics, that is, the baseline argument is used for convenience and then shown to be *only* a convenience.

This is in the spirit of the extended theory, but one generally sees only modest variations on premises. Leave-one-out diagnostics are fairly common (Carlin [36]), as are checks of (say) logistic-regression-in-place-of-probit (Gibbons et al [37]) or if-we-used-this-different-measurement (Caulkins and Padman [38]). However, we see hardly any checks for failures of proportional hazards – and SAS even supplies a test! It is truly extraordinary to see the full range of an argument's premises bent to see where the argument breaks.

Why are we so timid? Why not bend all of an argument's premises? Tractability was an acceptable excuse ten years ago; perhaps it still is, but it will certainly not be in ten years. Why do we not use our computing power – raw CPU power and flexible modeling systems like generalized linear models – to put experimental design methods to work exploring variations of premises?

Of course
ple, we think
variations. B
sions with no
not support s
models that c
[39].) And w
them? Judgir
other fields, b
examined, the

Well, can
of medical jo
problems are
middle groun
an argument
(Crawford et

At bottom
never sustain
need to find o
people sort ou
how much the

3.5 Idea the p the re large.

A premise is st
(This appropri
strongly suppo
large variation
for example, th
but difficult to
normality or li
field of market
and Bound's ca

It is more c
of the premises
the errors are a
same for all va
of the argumen
on the space of

Of course, any applied statistician can supply plausible excuses. For example, we think we know which variations are important, and we examine those variations. But do we know which variations are important? For linear regressions with normal errors, we probably do. For other models, the research does not support such an assertion, especially not for the bigger, less comprehensible models that computing power has made possible. (For example, see Zaslavsky [39].) And when we do know which variations are important, do we examine them? Judging from the medical literature, almost never. It is harder to assess other fields, but based on editing experience and conferences, if variations are examined, they are not discussed.

Well, can we not handle such difficulties qualitatively? For example, editors of medical journals expect papers to report losses to follow-up, and if such problems are bad enough the editors will reject the paper. But there is a middle ground between rejecting papers and dropping standards entirely: if an argument can be shown to hold in spite of problems, then it has value. (Crawford et al [40] make an argument like this.)

At bottom, one senses a worry that if we bend premises too much, we will never sustain any results. But perhaps that is the right answer: maybe we need to find out how far we must bend premises to break a result, and then let people sort out their beliefs about the result according to their beliefs about how much the premises *should* be bent. This leads to the last of the five ideas.

3.5 Idea #5: An argument is strong to the extent that the premises are supported by strong arguments or the region of premises yielding the same conclusion is large.

A premise is strongly supported when it is the conclusion of a strong argument. (This appropriately raises the specter of infinite regress.) When a premise is strongly supported, there is no need to examine variations of it or, at least, not large variations. It is easy to think of strongly supported qualitative premises – for example, that a coefficient should be positive because it can't be negative – but difficult to find instances of strong support for mathematical specifics like normality or linearity. Ehrenberg and Bound [41] give some exceptions in the field of marketing; their exceptional quality suggests that we heed Ehrenberg and Bound's call to seek regularity across many sets of similar data.

It is more common to find arguments that are strong because the specifics of the premises do not matter, for example, the same result is obtained whether the errors are assumed to be normal or t. If the outcome of an argument is the same for all variations of its premises, the argument is strong. If the outcome of the argument varies as the premises vary, then you can argue about metrics on the space of premises, but ultimately the choice of a metric is subjective.

This is another way to reach the conclusion, sometimes attributed to D.F. Andrews, that "objectivity is a hoax": an argument's strength can only be defined with respect to a specific body of knowledge and belief. One cannot foreclose the possibility that ten years into the future, someone will invent a new way to bend premises that *does* change the outcome of the argument, at which point the old, sturdy argument falls or surrenders some of the ground it occupies, as Newtonian mechanics did a century ago.

4 A simple example illustrates the extended theory's mechanics.

This example is from Carlin and Louis [42] (henceforth CL), who were motivated by a clinical trial in which Bayesian monitoring was conducted in parallel to conventional monitoring (Carlin et al [43]). The trial compared pyrimethamine to placebo as prophylaxes against toxoplasmic encephalitis in HIV-infected people. For the time-to-event analysis below, the endpoint was death from any cause. The investigators wanted to make an inferential causal argument about the efficacy of pyrimethamine, which CL formulated as a Bayesian test of a point-null hypothesis against a composite alternative. (Other formulations might have made sense; this is discussed in Section 5.) The primary premises are that the treatment groups did not differ systematically in baseline characteristics, concomitant treatments, or in ascertainment of deaths. These are supported by the procedural features of randomization and blinding. CL focused on the secondary mathematical premises.

The baseline model – specified in the trial's protocol document – was a proportional-hazards regression. CL used two explanatory variables for each patient: a dummy indicating treatment group (1 for pyrimethamine, 0 for placebo) and the CD4+ count at entry into the trial. CL integrated the CD4+ coefficient out of the partial likelihood, leaving the marginal (partial) likelihood for the pyrimethamine coefficient. Call that coefficient θ and call the marginal likelihood $f(x|\theta)$, where x represents the data from the trial. The explicit mathematical form of the baseline argument was:

- Null hypothesis, $H_0 : \theta = 0$; the prior probability of H_0 is $\pi\ = 0.25$;
- Alternative hypothesis, $H_a : \theta \neq 0$, where the prior cumulative distribution function of θ under H_a is $G(\theta)$; and
- Test: if $P(H_0|x) > p$, choose H_0, otherwise choose H_a, where $p = 0.1$.

Applying the extended theory. To use the extended theory, one must first fix an aspect of the problem that is essential to its definition. For CL, that aspect was θ, the log relative hazard of patients receiving pyrimethamine compared to those receiving placebo. Any other aspect of the problem formulation is a premise that can be varied: exchangeability of the treatment groups, the proportional hazards assumption, the use of a partial instead of a full likelihood, $\pi, p, G(\theta)$, etc.

Next, the baseline model is expanded to represent the premise variations under consideration, with the baseline premises corresponding to a specific prior on the expanded space. The term "model expansion" may cause some confusion: in the extended theory, it includes not just familiar (and typically modest) model expansions, but any and all premise variations.

Having expanded the model, the next step is to examine priors on the expanded space to find ones that change the baseline argument's conclusion. If any are found, the last step is to understand why those variations change the conclusion.

CL considered only variations on G and constrained them by specifying quantiles. That is, they defined an indifference zone (an interval such that they are indifferent between θ in this region and $\theta = 0$; see Freedman and Spiegelhalter [44]) $(x_L, x_U) = (-0.288, 0)$ for θ and defined the constraints by $P_G(\theta \leq x_L) = a_L$ and $P_G(\theta > x_U) = a_U$, for specified $a_L, a_U \in [0,1]$ satisfying $a_L + a_U < 1$. The maximum partial likelihood estimate of θ was 0.6, that is, pyrimethamine appeared to decrease survival time. Given the data, π, p, and the indifference zone, for any pair (a_L, a_U) the boundary of interest is linear: if $a_U > 0.145a_L + 0.273$, then a prior G exists that satisfies the constraints and results in the rejection of H_0. Otherwise, no prior permits rejection of H_0.

This is the boundary of variations at which the result changes. Its meaning is simple: if G puts enough probability above the indifference zone, it is possible to reject H_0; otherwise, it is not possible. This much is trivial; the non-trivial part is that CL quantified "enough" precisely, in the linear equation given above. (Sargent and Carlin [45] extended this result three ways, by using interval null hypotheses, including sample size considerations, and allowing π to vary.)

5 A generalization of the simple example suggests implications for foundations, methods, and computing.

The example in Section 4 can be generalized to derive five steps in applying the extended theory. In practice, it will often be necessary to iterate among the steps.

Step 1. Fix the aspect of the problem that is essential to its definition. There may be more than one way to do this. For example, suppose CL had varied the proportional hazards premise, replacing θ, the log relative hazard of pyrimethamine, by $\gamma + g(t)$, where t is time and $g(t)$ is a function of time. Are we interested in γ, or the whole function $\gamma + g(t)$, or some average of it? When the essential aspect of the problem has no single formulation, it may be necessary to vary the formulation in addition to varying baseline premises.

Step 2. Do EDA/model-building to form a baseline argument (Ideas #2 and #3). This is still art, not science, but the extended theory gives it a bottom line: get to an argument, or to the conclusion that no interesting argument can be made.

Step 3. Specify model expansions (Ideas #3 and #4). Each mathematical specific of the baseline premises is a *convenience* parameter: each specific is of no intrinsic interest but conveniently permits arguments to be built. Even the essential aspect of the problem is often just a convenience. This step requires a model expansion corresponding to each baseline premise that is not strongly supported.

Recall that "model expansion" includes *everything*. Whatever the premise variations, they can all be formulated as model expansions even if the various models cannot be nested. Sometimes this is awkward, but it can be done.

Step 4. Do a restricted search for priors on the convenience parameters that change the conclusion (Ideas #3 and #4). The search is restricted because it rarely makes sense to consider absolutely everything. One unhappy feature of robust Bayesian approaches generally and CL's approach in particular is that the G yielding extreme solutions are usually bizarre priors that nobody would advocate. Thus, it is desirable to impose smoothness constraints by means of, say, bounds on derivatives. Such constraints will partly address the concern mentioned in Section 3, that no result will ever stand up under the premise variations advocated here. The trick is to constrain priors in ways that are substantively innocuous but express genuine beliefs, like unimodality and continuity.

Step 5. On finding the boundary of such priors, figure out why the conclusion changes there (Idea #5). The baseline argument should be reported if it stands up under variations, but we also need to report where the baseline argument breaks in the expanded premise space, and why it breaks there. Does it break when we delete a few odd points? Is it so fragile that it breaks with some combination of modest deviations from several premises, or is a catastrophic failure of some premise needed? Such qualifications of the baseline argument are as important as the baseline argument itself.

5.1 Implications for foundations.

The most important implication is that the sensitivity analysis *is* the analysis, and that a sensitivity analysis must focus on the effects of perturbations on the *conclusion* of the baseline argument. Sensitivity checks are not a mere nicety; they are central to the analysis.

Another implication is that the usual taxonomies of statistical problems – for example, inference vs. prediction vs. decision, or hypothesis testing vs. point or interval estimation vs. everything else – do not make useful distinctions among statistical problems. One might say, paraphrasing de Finetti, that inference does not exist: "inference" is not a separate type of argument, but several subtypes of different types of argument. Perhaps the futile quality of the foundational dispute over inference arises in part because there is no such thing as "inference", but, rather, qualitatively different types of inference. Similarly, there is no single kind of prediction or decision, but qualitatively different kinds of each. Finally, although it serves a mechanical purpose to formulate a problem as a hypothesis test or an interval estimate, it is clear that tests, estimates, and intervals play a role in most or all of the types of arguments.

The extended theory also makes it clear – if you are not already convinced – that it is not helpful to reduce all statistical arguments to exercises in decision theory. While non-action arguments can often be cast in decision-theoretic terms, it is sterile to do so. How could any meaningful loss function be constructed for the Neaton/Wentworth argument or for the causal arguments? The main burden of such arguments is not picking a particular estimate or posterior distribution, but sustaining the finding against alternative explanations. But fear not, purists: it is possible, after all, to use Savage's axioms for probability without using the axioms for utility.

A final implication, which cannot be treated here, is that past a certain point it is usually futile to try to express all variability and uncertainty as probability. For most argument types, it would be a waste of time to specify a probability distribution on the premise variations and integrate it out, because it would be just one more aspect of the problem to vary. (But perhaps not always; Hodges [46] discusses possible exceptions.)

5.2 Implications for methods.

Four of the five steps in applying the extended theory have immediate implications for methodological development.

Step 1. Each type of argument has a small group of characteristic problem formulations. For example, CL's inferential causal argument was formulated as a Bayesian hypothesis test. We need to catalog problem formulations for each argument type and develop CL-style setups friendly to perturbations of the generality discussed here.

Step 3. Step 2, the EDA/model-building step, produces a tractable baseline argument. Often this will involve a standard model, such as linear regression, so statisticians will routinely need to examine variations on the premises of these standard models. Thus, we will need standard "baskets" of model expansions for standard models. Of course, some users will need to go beyond the standard basket of model expansions, and one research challenge is to figure out how to allow them to do so without respecifying the standard model expansions.

Step 4. This step is a search through priors on the expanded model space for priors that change the conclusion of the baseline argument. To do this readily, we need friendly classes of priors on the standard baskets of model expansions, and constraints on them corresponding to conditions like continuity and unimodality.

Step 5. We need more ways to assess the effects of perturbations on conclusions of arguments. Kass, Tierney, and Kadane [33] suggests one approach; a computational idea will be suggested below. When the search is finished, we need ways to describe the boundary where the baseline argument's conclusion changes. CL drew a simple picture to summarize their search; more complicated model expansions will require more ingenious pictures.

5.3 Implications for computing.

It is not easy to search for the boundary where the conclusion changes. Explicit solutions are awkward for minor elaborations of CL's problem and for more general problems we cannot avoid computer-intensive searches.

One possible approach is an environment that searches premise variations stochastically, maps the boundary, and suggests diagnostics to elucidate why the conclusion changes there. Ordinarily, if we are using (say) a linear regression model, we have a dozen or so standard diagnostics that we apply, one by one, looking for problems. We may use all the standard diagnostics and find nothing, or a particular diagnostic may indicate a problem which we then pursue with other diagnostics. The computing system described above would invert this process by searching for places in premise space where the conclusion of the baseline argument changes and, by reference to the model expansion that seems to cause the change, displaying diagnostics that indicate what has gone wrong.

One might view the latter as an intelligent agent that does diagnostics for the human user. Other expert systems can help the human implement Step 2, the EDA/model-building step. Particularly in large data sets, it may not be cost-effective to have the human waiting while the computer does arithmetic; rather, it may be more efficient to have an intelligent agent do EDA/model-building and bring back a summary that the human can peruse at her convenience. The difficulty is designing an interface that allows the human to direct the agent so that it does not bring back mostly junk.

6 Antec

The ideas in tl
innumerable les
former can be g

The nearest
analysis (Hodge
ploratory (as op
ly-oriented unpu
and incorporate

Carlin and I
technically mod
a departure am
(Others, such a

Smith [51] s
that Bayes' the
pings induced l
the idea can rea
aspect of a mo
aspects, which
ancestor of the
tion between pı
and Weisberg [
credits to Cox [
for diverse pert
robust decision
case influence o
ness literature (
prior distributi
aspects of the l

Other ante
and Berger [3]
sus that non-B
and Schum [57
legal argument
type of evidenc
prior informati
ization theory
observational s
tween the mecl
to overturn an
outcome. Fina

6 Antecedents.

The ideas in this paper have several immediate antecedents and, of course, innumerable less immediate predecessors. In a paper of this scope, only the former can be given their due; my apologies to the latter.

The nearest antecedent is a catalog of uses of simulation models in policy analysis (Hodges [47], Hodges and Dewar [48]) and Bankes' [49] notion of exploratory (as opposed to predictive) modeling which was followed by statistically-oriented unpublished work by Bankes and JL Adams. These streams merged and incorporated experimental design ideas in Dewar et al [50].

Carlin and Louis [42] is also a direct antecedent. Although their results are technically modest and related to other results in Bayesian robustness, they are a departure among statisticians in their focus on the conclusion of the analysis. (Others, such as decision analysts, have long focused on conclusions.)

Smith [51] suggested the idea of extravagant sensitivity analyses by noting that Bayes' theorem allows us to "report a rich range of the possible belief mappings induced by a data set"; he was referring to individual parameters, but the idea can readily be expanded. Cox and Snell [52] differentiated the primary aspect of a model, specifying the main question of interest, and the secondary aspects, which complete the model and indicate a suitable analysis. This is an ancestor of the notion of the essential aspect of a problem and of the distinction between primary and secondary premises. Cook and Weisberg (e.g., Cook and Weisberg [28], Weisberg [53]) advocated model expansion, which Weisberg credits to Cox [54]. Cook [29] and Ramsay and Novick [55] introduced methods for diverse perturbations, the latter in the form of PLU (prior-likelihood-utility) robust decision theory. Kass, Tierney, and Kadane [33] developed measures of case influence on general functions of parameters. Finally, the Bayesian robustness literature (Wasserman [56] is a recent survey) has focused on robustness to prior distributions, although some authors have considered sensitivity to other aspects of the baseline argument.

Other antecedents come from off the beaten track. Smith [1], Hill [2], and Berger [3] discussed Bayesian notions of data analysis, with the consensus that non-Bayesian methods are acceptable during model-building. Kadane and Schum [57] discussed Bayesian thinking within a scheme of diagramming legal arguments. Lindley and Singpurwalla [58] considered the amount and type of evidence needed to obtain agreement between adversaries with different prior information and loss functions. Rosenbaum (e.g., [59]) pushed randomization theory in an unusual direction with methods for sensitivity analyses in observational studies. These methods evaluate the strength of relationship between the mechanism treatment assignment and the outcome that is necessary to overturn an apparent causal relation between the treatment itself and the outcome. Finally, Leamer [15] incorporated model-searching into formal statistical inference and Leamer [60] catalogued distinct patterns of model-searching. I find it unhelpful to think of uses of statistics as specification searches, but the debt of the present paper to Leamer is clear.

Acknowledgements

The RAND blender mixed my ideas inextricably with those of John Adams, Steve Bankes, and Jim Dewar. Thanks also go to Brad Carlin, David Freedman, David Lane, Tom Louis, and Dan Sargent for helpful comments and discussion.

References

[1] Smith, A.F.M. 'Some Bayesian thoughts on modelling and model choice', *The Statistician*, **35**, 97–102 (1986).

[2] Hill, B.M. 'A theory of Bayesian data analysis', In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, eds. S Geisser, JS Hodges, SJ Press, A Zellner, Amsterdam: North-Holland, 49–74 (1990).

[3] Berger, J.O. 'Contributed discussion', In *Case Studies in Bayesian Statistics*, eds. C Gatsonis, JS Hodges, RE Kass, ND Singpurwalla. New York: Springer-Verlag, 302–303 (1993).

[4] West, M. and Harrison, J. *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag (1989).

[5] Fisher, R.A. *Statistical Methods and Scientific Inference*, 3rd edition, New York: Hafner (1973).

[6] Efron, B. 'Why isn't everyone a Bayesian?', *American Statistician*, **40**, 1–11 (1986).

[7] Fischl, M.A., Richman, D.D., Grieco M.H., et al. 'The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex', *New England Journal of Medicine*, **317**, 185–191, (1987).

[8] Kahn, J.O., Lagakos, S.W., Richman, D.D., et al. 'A controlled trial comparing continued zidovudine [AZT] with didanosine in human immunodeficiency virus infection', *New England Journal of Medicine*, **337**, 581-587, (1992).

[9] Litterman, R.B. 'A statistical approach to economic forecasting', *Journal of Business and Economic Statistics*, **4**, 1–4, (1986).

[10] Mallows, C.L. 'Data description', In *Scientific Inference, Data Analysis, and Robustness*, eds. GEP Box, T Leonard, C-F Wu, Academic Press, 135-152 (1983).

[11] Draper, D.C., Hodges, J.S., Mallows, C.L., and Pregibon, D. 'Exchangeability and data analysis (with discussion)', *Journal of the Royal Statistical Society*, Series A, **156**, 9–37 (1993).

[12] Carter, H.R., Page, G.W., and Ford, R.G. 'The importance of rehabilitation center data in determining the impacts of the 1986 oil spill on marine birds in central California', *Wildlife Journal*, **10**, 9–14 (1987).

[13] Blakeman, B.M., Pifarre, R., Sullivan H., et al. 'High-risk heart surgery in the heart transplant candidate', *Journal of Heart Transplantation*, **5**, 468–472 (1990).

[14] Tukey, J W. 'Some thoughts on clinical trials, especially problems of multiplicity', *Science*, **198**, 679–684 (1977).

[15] Leamer, E.E. 'False models and post-data model construction', *Journal of the American Statistical Association*, **69**, 122-131 (1974).

[16] Lindley, D.V. Discussion of Efron [6]

[17] Lindley, D.V. 'The 1988 Wald Memorial Lectures: The present position in Bayesian statistics (with discussion)', *Statistical Science*, **5**, 44-89 (1990).

[18] Berry, D.A. 'Subgroup analyses', *Biometrics*, **47**, 1227-1230 (1990).

[19] Neaton, J. and Wentworth, D. 'Relationship of serum cholesterol and blood pressure measured prior to HIV-infection with risk of death from AIDS', Unpublished manuscript, Division of Biostatistics, School of Public Health, University of Minnesota.

[20] Fleming, T.R. 'Surrogate markers in AIDS and cancer trials', *Statistics in Medicine*, in press (1995).

[21] De Gruttola, V., Wulfsohn, M., Fischl, M.A., and Tsiatis, A.A. 'Modeling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex', *Journal of Acquired Immune Deficiency Syndromes*, **6**, 359–365 (1993).

[22] Choi, S., Lagakos, S.W., Schooley, R.T., and Volberding, P.A. 'CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine', *Annals of Internal Medicine*, **118**, 674–680 (1993).

[23] Schmitz, J. 'Massive Marketing Datasets', presented on 7 July 1995 at "Statistical Challenges and Possible Approaches in the Analysis of Massive Data Sets," a conference sponsored by the Committee on Applied and Theoretical Statistics; Washington, DC.

[24] Koech, D., et al. 'Low-dose oral alpha-interferon therapy for patients seropositive for human immunodeficiency virus type-1 (HIV-1)', *Molecular Biotherapy*, **2**, 91–95 (1990).

[25] Obel, A.O. and Koech, D. 'Outcome of intervention with or without low-dose oral alpha-interferon in 32 HIV-1 seropositive patients in a referral hospital', *East African Medical Journal*, **67(7)**, 71–76 (1990).

[26] Hardy, W.D., Feinberg, J., Finkelstein, D.M., et al. 'A controlled trial of trimethprim-sulfamethoxazole or aerosolized pentamidine for secondary prophylaxis of *Pneumocystis carinii* pneumonia in patients with the acquired immuneodeficiency syndrome', *New England Journal of Medicine*, **327**, 1842–1848 (1992).

[27] Schneider, M.M.E., Hoepelman, A.I.M., Schattenkerk, J.K.M.E, et al. 'A controlled trial of aerosolized pentamidine or trimethoprim-sulfamethoxazole as primary prophylaxis against *Pneumocystis carinii* pneumonia in patients with human immunodeficiency virus infection', *New England Journal of Medicine*, **327**, 1836–1841 (1992).

[28] Cook, R.D. and Weisberg, S. *Residuals and Influence in Regression*, New York: Chapman and Hall. (1982).

[29] Cook, R.D. 'Assessment of local influence (with discussion)', *Journal of the Royal Statistical Society*, Series B, **48**, 133–169 (1986).

[30] Johnson, W.O. and Geisser, S. 'A predictive view of the detection and characterization of influential observations in regression analysis', *Journal of the American Statistical Assn.*, **78**, 137–144 (1983).

[31] McCulloch, R.E. 'Local model influence', *Journal of the American Statistical Assn.*, **84**, 473–478 (1989).

[32] Carlin, B.P. and Polson, N.G. 'Monte Carlo Bayesian methods for discrete regression models and categorical time series', In *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, Oxford University Press, 577–586, (1992).

[33] Kass, R.E., Tierney, L., and Kadane, J.B. 'Approximate methods for assessing influence and sensitivity in Bayesian analysis', *Biometrika*, **76**, 663–674, (1989).

[34] Gelfand, A
dictive dist
Bayesian S
Smith, Ox

[35] Draper, D.
ity and da
Society, Se

[36] Carlin, B.I
working m
tion', *Jour*

[37] Gibbons, I
effects prol
the Americ

[38] Caulkins,
for illicit d
(1993).

[39] Zaslavsky,
data to est
Assn., **88**,

[40] Crawford, S
based meth
tice Survey
Hodges, R.
(1993).

[41] Ehrenberg,
discussion)
206 (1993).

[42] Carlin, B.I
duce specif
Bayesian A
Arnold Zel
Wiley. (199

[43] Carlin, B.P
monitoring
Bayesian S
N.D. Singp

[34] Gelfand, A.E., Dey, D.K., and Chang, H. 'Model determination using predictive distributions with implementation via sampling-based methods', In *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, Oxford University Press, 147–167, (1992).

[35] Draper, D., Hodges, J.S., Mallows, C.L., and Pregibon, D. 'Exchangeability and data analysis (with discussion)', *Journal of the Royal Statistical Society*, Series A, **156**, 9–37 (1993).

[36] Carlin, B.P., Kass, R.E., Lerch, F.J., and Huguenard, B.R. 'Predicting working memory failure: A subjective Bayesian approach to model selection', *Journal of the American Statistical Assn.*, **87**, 319–327 (1992).

[37] Gibbons, R.D., Hedeker, D., Charles, S.C., and Frisch, P. 'A random-effects probit model for predicting medical malpractice claims', *Journal of the American Statistical Assn.*, **89**, 760–767 (1994).

[38] Caulkins, J.P. and Padman, R. 'Quantity discounts and quality premia for illicit drugs', *Journal of the American Statistical Assn.*, **88**, 748–757 (1993).

[39] Zaslavsky, A.M. 'Combining census, dual-system, and evaluation study data to estimate population shares', *Journal of the American Statistical Assn.*, **88**, 1092-1105 (1993).

[40] Crawford, S.L., Johnson, W.G., and Laird, N.M. 'Bayes analysis of model-based methods for nonignorable nonresponse in the Harvard Medical Practice Survey', In *Case Studies in Bayesian Statistics*, eds. C. Gatsonis, J.S. Hodges, R.E. Kass, N.D. Singpurwalla, New York: Springer-Verlag, 78–117 (1993).

[41] Ehrenberg, A.S.C. and Bound, J.A. 'Predictability and prediction (with discussion)', *Journal of the Royal Statistical Society*, Series A, **156**, 167–206 (1993).

[42] Carlin, B.P. and Louis, T.A. 'Identifying prior distributions that produce specific decisions, with application to monitoring clinical trials.', In *Bayesian Analysis of Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D.A. Berry, K.M. Chaloner, J.K. Geweke, New York: Wiley. (1996).

[43] Carlin, B.P., Chaloner, K.M., Louis, T.A., and Rhame, F.S. 'Elicitation, monitoring, and analysis for an AIDS clinical trial', In *Case Studies in Bayesian Statistics, Volume II*, eds. C. Gatsonis, J.S. Hodges, R.E. Kass, N.D. Singpurwalla, New York: Springer-Verlag (1995).

[44] Freedman, L.S. and Spiegelhalter, D.J. 'Application of Bayesian statistics to decision making during a clinical trial', *Statistics in Medicine*, **11**, 23–35 (1992).

[45] Sargent, D. and Carlin, B.P 'Robust Bayesian design and analysis of clinical trials via prior partitioning (with discussion)', Research Report 94-016, Division of Biostatistics, University of Minnesota, 1994. To appear in the IMS Lecture Note Series.

[46] Hodges, J.S. 'Uncertainty, policy analysis, and statistics (with discussion)', *Statistical Science*, **2**, 259–291 (1987).

[47] Hodges, J.S. 'Six (or so) things you can do with a bad model', *Operations Research*, **39**, 355–365 (1991).

[48] Hodges, J.S. and Dewar, J.A. 'Is it you or your model talking? A framework for model validation', RAND, R-4114-AF/A/OSD, Santa Monica, California (1992).

[49] Bankes, S.C. 'Exploratory modeling and the use of simulation for policy analysis', RAND, N-3093-A, Santa Monica, California (1992).

[50] Dewar, J.A., Bankes, S.C., Hodges, JS., et al. 'Credible uses of the Distributed Interactive Simulation (DIS) System' RAND, MR-607-A, Santa Monica, California (1995).

[51] Smith, A.F.M. 'Present position and potential developments: Some personal views of Bayesian statistics (with discussion)', *Journal of the Royal Statistical Society*, Series A, **147**, 245–259 (1984).

[52] Cox, D.R. and Snell, E.J. *Applied Statistics: Principles and Examples.* London: Chapman and Hall (1981).

[53] Weisberg, S. 'Some principles for regression diagnostics and influence analysis: Comments on "Developments in linear regression methodology: 1959-1982" by RR Hocking', *Technometrics*, **25**, 240–244 (1983).

[54] Cox, D.R. 'Nonlinear models, residuals, and transformations', *Math. Operationsforsch. Statist. Ser. Statistics*, **8**, 3–22 (1977).

[55] Ramsay, J.O. and Novick, M.R. 'PLU robust Bayesian decision theory: Point estimation', *Journal of the American Statistical Association*, **75**, 901–907 (1980).

[56] Wasserman, L. 'Recent methodological advances in robust Bayesian inference', In *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, Oxford University Press, 483–502 (1992).

[57] Kada
case'
Daw

[58] Lind
agre
pling

[59] Rose
in m

[60] Lear

[57] Kadane, J.B. and Schum, D.A. 'Opinions in dispute: The Sacco-Vanzetti case', In *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, Oxford University Press, 267–287 (1992).

[58] Lindley, D.V. and Singpurwalla, N.D. 'On the evidence needed to reach agreed action between adversaries, with application to acceptance sampling', *Journal of the American Statistical Association*, **86**, 933–937 (1991).

[59] Rosenbaum, P.R. 'Sensitivity analysis for certain permutation inferences in matched observational studies', *Biometrika*, **74**, 13–26 (1987).

[60] Leamer, E.E. *Specification Searches*, New York: Wiley (1978).

Jack C. Lee
Wesley O. Johnson
Arnold Zellner
Editors

Modelling and Prediction Honoring Seymour Geisser

With 76 Figures

Subtitle: A Terminally Obscure Book

Springer

Jack C. Lee
National Chiao Tung University
Institute of Statistics
1001 Ta-Hseuh Road
Hsichu, Taiwan R.O.C.

Wesley O. Johnson
University of California, Davis
Department of Statistics
Davis, CA 95616
USA

Arnold Zellner
University of Chicago
Graduate School of Business
1101 East 58th Street
Chicago, IL 60637
USA

Library of Congress Cataloging-in-Publication Data
Modelling and prediction : honoring Seymour Geisser / edited by Jack C.
Lee, Wesley O. Johnson, Arnold Zellner.
p. cm.
"Proceedings of the 'Conference on Forecasting, Prediction, and
Modelling in Statistics and Econometrics' which was held at National
Chiao Tung University, Hsinchu, Taiwan during the period December
12–14, 1994"—Pref.
Includes bibliographical references and index.
ISBN 0-387-94808-2 (hardcover : alk. paper)
1. Bayesian statistical decision theory—Congresses. I. Geisser,
Seymour. II. Lee, Jack C. III. Johnson, Wesley O. IV. Zellner,
Arnold V. Conference on Forecasting, Prediction, and Modelling in
Statistics and Econometrics (1994 : National Chiao Tung University)
QA279.5.M595 1996
519.5′4—dc20
96-19132

Printed on acid-free paper.

Production managed by Allan Abrams; manufacturing supervised by Jeffrey Taub.
Camera-ready copy prepared from the author's files.
Printed and bound by Braun-Brumfield, Inc., Ann Arbor, MI.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-94808-2 Springer-Verlag New York Berlin Heidelberg SPIN 10524242

T
Universit
Predictic
Tung Uni
was orga
Chicago
Educatio
Analysis
T
following
G.C. Tiao
Chou, L
Conferer
President
crucial t
Director
NRC's S
T
extremel
reasonabl
ultimatel
The failu
more use
natural u
Because
him.
P
New Yo
Universi
Professor
for the N
section o
1965. D
at the G
Departme
Geisser h
at the Uni
of unive
Universit
and he w