Bayesian Model Specification: Toward a Theory of Applied Statistics

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
www.ams.ucsc.edu/~draper

ICSA 2010: Guangzhou, China

21 December 2010

(based in part on Draper D, Krnjajić M (2010). Calibration results for Bayesian model specification. Submitted.)
Outline (Inaccurate Homage to David Blackwell)

(1) **Axiomatization** of statistics.

(2) **Foundations** of **probability** secure: (RT Cox, 1946)

Principles → Axioms → Theorem:

Logical consistency in uncertainty quantification → Bayes.

(3) **Foundations** of **inference, prediction** and **decision-making** not yet secure: fixing this would yield a Theory of Applied Statistics, which we do not yet have.

(a) Cox’s Theorem doesn’t require You to pay attention to a basic scientific issue: how often do You get the **right answer**?

(b) Too much ad hockery in model specification: still lacking

Principles → Axioms → Theorems.

(4) A **Calibration Principle** fixes 3 (a) via decision theory.

(5) Log scores help with 3 (b) via a Modeling-As-Decision Principle and a Prediction Principle; some people have claimed that log scores are not asymptotically consistent, but in my view this position reflects an incorrect problem formulation.
An Axiomatization of Statistics

1 (definition) Statistics is the study of uncertainty: how to measure it well, and how to make good choices in the face of it.

2 (definition) Uncertainty is a state of incomplete information about something of interest to You (Good, 1950: a generic person wishing to reason sensibly in the presence of uncertainty).

3 (axiom) (Your uncertainty about) “Something of interest to You” can always be expressed in terms of propositions: true/false statements $A, B, \ldots$.

Examples: You may be uncertain about the truth status of
- $A = \text{(Barack Obama will be re-elected U.S. President in 2012)}$
- $B = \text{(the in-hospital mortality rate for patients at hospital $H$ admitted in calendar 2011 with a principal diagnosis of heart attack will be between 5\% and 25\%)}$

4 (implication) It follows from 1-3 that statistics concerns Your information (NOT your beliefs) about $A, B, \ldots$. 
Axiomatization (continued)

5 (axiom) But Your information cannot be assessed in a vacuum: all such assessments must be made relative to (conditional on) Your background assumptions and judgments about how the world works vis à vis $A, B, \ldots$.

6 (axiom) These assumptions and judgments, which are themselves a form of information, can always be expressed in a set $B$ of propositions (examples below).

7 (definition) Call the “something of interest to You” $\theta$; in applications $\theta$ is often a vector or matrix of real numbers, but in principle it could be almost anything (an image of the surface of Mars, a phylogenetic tree, ...).

8 (axiom) There will typically be an information source (data set) $D$ that You judge to be relevant to decreasing Your uncertainty about $\theta$; in applications $D$ is often again a vector or matrix of real numbers, but in principle it too could be almost anything (a movie, the words in a book, ...).

Examples of $B$: 

**Axiomatization (continued)**

- If $\theta$ is the **mean survival time** for a **specified group of patients** (who are alive now), then $\mathcal{B}$ includes the proposition ($\theta \geq 0$).

- If $D$ is the result of an **experiment** $E$, then $\mathcal{B}$ might include the proposition (Patients were **randomized** into one of two groups, **treatment** (new drug) or **control** (current best drug)).

  $\Box$ (implication) The presence of $D$ creates a **dichotomy**:

  - Your information about $\theta \{\text{internal, external}\}$ to $D$.

  (People often talk about a **different dichotomy**: Your information about $\theta \{\text{before, after}\}$ $D$ arrives (prior, posterior), but temporal considerations are actually irrelevant.)

  $\Box$ (implication) It follows from $\Box$-$\Box$ that statistics concerns itself principally with **five things** (omitted: description, data integrity, ...)

  (1) **Quantifying Your information** about $\theta$ **external** to $D$ (given $\mathcal{B}$), and doing so well;
(2) **Quantifying Your information** about $\theta$ internal to $D$ (given $\mathcal{B}$), and doing so **well**;

(3) **Combining** these two **information sources** (and doing so **well**) to create a **summary** of **Your uncertainty** about $\theta$ (given $\mathcal{B}$) that includes all available **information** You judge to be **relevant** (this is **inference**); and

Using all Your **information** about $\theta$ (given $\mathcal{B}$) to make

(4) **Predictions** about future data values $D^*$, and

(5) **Decisions** about how to **act sensibly**, even though Your **information** about $\theta$ may not be **complete**.

**Foundational question:** How should these tasks be **accomplished**?

This question has **two parts**: **probability** and **statistics**; in my view, the **probability foundations** are **secure**, the **statistics foundations** are **not**.

From the **1650s** (Fermat, Pascal) through the **18th century**
Theory of Probability

(Bayes, Laplace) to the 1860s (Venn, Boole), three different ideas about how to think about uncertainty quantification — classical, Bayesian, and frequentist probability — were put forward in an intuitive way, but no one ever tried to prove a theorem of the form \{given these premises, there’s only one sensible way to quantify uncertainty\} until de Finetti (1937) and a physicist named RT Cox (1946).

Cox’s goal was to identify what basic rules \( pl(A|B) \) — the plausibility of \( A \) given \( B \) — should follow so that \( pl(A|B) \) behaves sensibly, where \( A \) and \( B \) are propositions with \( B \) known by You to be true and the truth status of \( A \) unknown to You.

He did this by identifying a set of principles making operational the word “sensible” (Jaynes, 2003):

- You need to be willing to represent degrees of plausibility by real numbers (i.e., \( pl(A|B) \) is a function from propositions \( A \) and \( B \) to \( \mathbb{R} \));
- You insist that Your reasoning be logically consistent:
— If a plausibility assessment can be arrived at in more than one way, then every possible way must lead to the same value.

— You always take into account all of the evidence You judge to be relevant to the plausibility assessment under consideration.

— You always represent equivalent states of information by equivalent plausibility assignments.

From these principles Cox derived a set of axioms:

• The plausibility of a proposition determines the plausibility of the proposition’s negation; each decreases as the other increases.

• The plausibility of the conjunction $AB = (A$ and $B)$ of two propositions $A, B$ depends only on the plausibility of $B$ and that of $\{A$ given that $B$ is true$\}$ (or equivalently the plausibility of $A$ and that of $\{B$ given that $A$ is true$\}$).

• Suppose $AB$ is equivalent to $CD$; then if You acquire new information $A$ and later acquire further new information $B$, and update all plausibilities each time, the updated plausibilities will be the same as if You
had first acquired new information $C$ and then acquired further new information $D$.

From these axioms Cox proved a theorem showing that uncertainty quantification must behave in one and only one way:

**Theorem:** If You accept Cox’s axioms, plus the convention that You always want plausibilities to be finite real numbers, then to be logically consistent You must quantify uncertainty as follows:

- Your *plausibility operator* $p_l(A|B)$ — for propositions $A$ and $B$ — can be referred to as Your *probability* $p(A|B)$ that $A$ is true, *given* that You regard $B$ as true, and $0 \leq p(A|B) \leq 1$, with certain truth of $A$ (given $B$) represented by 1 and certain falsehood by 0.

  - $p(A|B) + p(\bar{A}|B) = 1$, where $\bar{A} = \text{not } A$.

  - $p(AB|C) = p(A|C) \cdot p(B|AC) = p(B|C) \cdot p(A|BC)$.

The proof (see, e.g., Jaynes (2003)) involves deriving two functional equations $F[F(x, y), z] = F[x, F(y, z)]$ and $x \cdot S \left[ \frac{S(y)}{x} \right] = y \cdot S \left[ \frac{S(x)}{y} \right]$. 


Optimal Reasoning

that \( p(A|B) \) must satisfy and then **solving** those equations.

A number of **important corollaries** arise from Cox’s Theorem:

- For **real-valued** \( \theta \), You can use **notation** such as \( p(\theta \leq t|B) \) to stand for \( p(A|B) \) with \( A = (\theta \leq t) \), which opens up the world of **cumulative distribution functions** (CDFs) and **density functions** to You; indeed, You **MUST** be prepared to **quantify uncertainty** about a **real-valued** \( \theta \) given \( B \) via a **CDF** of the form \( F_{\theta}(t|B) = p(\theta \leq t|B) \), with associated **density function** \( p(\theta|B) \) if it exists — i.e., to be **logically consistent** You **MUST** reason probabilistically about **ALL sources of uncertainty** (including real-valued population parameters), and **this makes you a Bayesian**.

- Strictly speaking, Cox’s Theorem only tells You how to **reason sensibly** about a **finite** collection \( \mathcal{C} \) of propositions, but the **accuracy** with which You measure a **real-valued quantity** \( \theta \) is limited in practice to a **finite** number of **significant figures**, meaning that in principle You only need to think about situations with **finite** \( |\mathcal{C}| \); I agree with Jaynes (2003) that You can only claim to have **earned** the **convenience** of **continuous models** for **real-valued** \( \theta \) by first (a) **defining the discourse** with finite \( |\mathcal{C}| \) and only then
(b) explicitly identifying the **unique way** in which You will let $|\mathcal{C}| \to \infty$
(similar remarks apply if $\theta$ is a **function**: this can be defined on a **finite grid**, with **smooth interpolation** an assumption in $\mathcal{B}$ to fill in the **gaps**).

- Given that set $\mathcal{B}$, of **propositions** summarizing Your **background assumptions and judgments** about how the world works as far as $\theta$, $D$ and future data $D^*$ are concerned,

  (a) You must be prepared to specify

  **two conditional probability distributions:**

  — $p(\theta|\mathcal{B})$, to quantify **all information** about $\theta$ **external** to $D$ that You judge relevant; and

  — $p(D|\theta \mathcal{B})$, to quantify Your **predictive uncertainty**, given $\theta$, about the data set $D$ before it’s arrived.

  (b) Given the distributions in (a), the distribution $p(\theta|D \mathcal{B})$ quantifies **all relevant information** about $\theta$, both **internal and external** to $D$, and **must** be computed via Bayes’s Theorem:
Optimal Inference, Prediction and Decision

\[ p(\theta|D B) = c p(\theta|B) p(D|\theta B), \]  
\text{(inference)} \tag{1} \]

where \( c > 0 \) is a \textbf{normalizing constant} chosen so that the left-hand side of (1) integrates (or sums) over \( \Theta \) to 1 (here \( \Theta \) is the set of \textbf{possible} \( \theta \) values);

(c) Your \textbf{predictive distribution} \( p(D^*|D B) \) for future data \( D^* \) given the \textbf{observed data set} \( D \) must be \textbf{expressible} as follows:

\[ p(D^*|D B) = \int_{\Theta} p(D^*|\theta D B) p(\theta|D B) d\theta; \]

typically there’s \textbf{no information} about \( D^* \) contained in \( D \) if \( \theta \) is known, in which case this expression \textbf{simplifies} to

\[ p(D^*|D B) = \int_{\Theta} p(D^*|\theta B) p(\theta|D B) d\theta; \]  
\text{(prediction)} \tag{2} \]

(d) to make a sensible \textbf{decision} about which \textbf{action} \( a \) to take in the face of \textbf{uncertainty} about \( \theta \), You \textbf{must be prepared to specify}

(i) the set \( A \) of \textbf{feasible actions} among which You’re choosing, and
(ii) a utility function $U(a, \theta)$, taking values on $\mathbb{R}$ and quantifying Your judgments about the rewards (monetary or otherwise) that would ensue if You chose action $a$ and the unknown actually took the value $\theta$;

then the optimal decision is to choose the action $a^*$ that maximizes the expectation of $U(a, \theta)$ over $p(\theta|D \mathcal{B})$: 

$$a^* = \arg\max_{a \in A} E_{\theta|D \mathcal{B}} U(a, \theta) = \arg\max_{a \in A} \int_{\Theta} U(a, \theta) p(\theta|D \mathcal{B}) d\theta. \quad (3)$$

These corollaries to Cox’s theorem solve problems (3–5) above — they leave no ambiguity about how to draw inferences, and make predictions and decisions, in the presence of uncertainty — but problems (1) and (2) are still unaddressed: to implement this logically-consistent approach in a given application, You have to specify

- $p(\theta|\mathcal{B})$, often referred to as Your prior information about $\theta$ (given $\mathcal{B}$; this is better understood as a summary of all relevant information about $\theta$ external to $D$, rather than by appeal to any temporal (before-after) considerations);
Specification Burden (continued)

- \( p(D|\theta \mathcal{B}) \), often referred to as Your **sampling distribution** for \( D \) given \( \theta \)
  (and \( \mathcal{B} \); this is better understood as Your **conditional predictive distribution** for \( D \) given \( \theta \), before \( D \) has been observed, rather than by appeal to other data sets that might have been observed); and

- the **action space** \( \mathcal{A} \) and the **utility function** \( U(a, \theta) \) for decision-making purposes.

The results of **implementing** this approach are

- \( p(\theta|D \mathcal{B}) \), often referred to as Your **posterior** distribution for \( \theta \) given \( D \) (and \( \mathcal{B} \); as above, this is better understood as the **totality of Your current information** about \( \theta \), again without appeal to temporal considerations);

- Your **posterior predictive distribution** \( p(D^*|D \mathcal{B}) \) for future data \( D^* \) given the **observed data set** \( D \); and

- the **optimal decision** \( a^* \) given all available information (and \( \mathcal{B} \)).

**To summarize:** Inference and prediction require You to **specify** \( p(\theta|\mathcal{B}) \) and \( p(D|\theta \mathcal{B}) \); decision-making requires You to **specify** the same two
ingredients plus $A$ and $U(a, \theta)$; how should this be done in a **sensible** way?

Cox’s Theorem and its corollaries provide **no constraints on the specification process**, apart from the requirement that **all probability distributions** be **proper** (integrate or sum to 1).

In my view, in seeking **answers** to these **specification questions**, as a **profession** we’re approximately where the **discipline of statistics** was in arriving at an **optimal theory of probability** before Cox’s work: many people have made **ad-hoc suggestions**, but little **formal progress** has been made.

Developing (1) **principles**, (2) **axioms** and (3) **theorems** about **optimal specification** could be regarded as creating a **Theory of Applied Statistics**.

$p(\theta|\mathcal{B})$, $p(D|\theta \mathcal{B})$ and $\{A, U(a, \theta)\}$ are all important; for lack of time I’ll focus here on the **problem of specifying** $\{p(\theta|\mathcal{B}), p(D|\theta \mathcal{B})\}$ — call such a **specification** a **model** $M$ for Your uncertainty about $\theta$. 
Calibration Principle

How should \( M \) be specified? Where is the progression

Principles \( \rightarrow \) Axioms \( \rightarrow \) Theorems

to guide You, the way Cox’s Theorem settled the foundational questions
for probability?

In my view this is the central unsolved foundational problem in
statistical inference and prediction.

As a contribution to closing the gap between ad-hoc practice and lack of
theory, here’s a principle worth considering:

Calibration Principle: In model specification, You should pay attention
to how often You get the right answer, by creating situations in which
You know what the right answer is and seeing how often Your
methods recover known truth.

The reasoning behind the Calibration Principle is as follows:

(axiom) You want to help positively advance the course of science.
(remark) However, there’s nothing in the Bayesian paradigm to prevent You from making one or both of the following mistakes — (a) choosing $p(D|\theta \mathcal{B})$ unwisely; (b) inserting \{strong information\} about $\theta$ external to $D$ into the modeling process that turns out after the fact to have been (badly) out of step with reality — and, if You repeatedly do this,

(i) it would seem likely that Your colleagues will stop inviting You into their projects as a statistical collaborator and

(ii) this runs counter to Your axiomatic desire to aid in the scientific enterprise.

(remark) Calibration can be given an entirely Bayesian justification via decision theory:

Taking a broader perspective over Your career, not just within any single attempt to solve an inferential/predictive problem in collaboration with other investigators, Your desire to avoid the loss of collaborative opportunities, arising from getting the wrong answer too often, and to take part positively in the progress of science can be quantified in a
utility function that incorporates a bonus for being well-calibrated, and in this context (Draper and Von Brzeski, 2010) calibration-monitoring emerges as a natural and inevitable Bayesian activity.

This seems to be a new idea: logical consistency yields Bayesian uncertainty assessment but does not provide guidance on model specification; if You accept the Calibration Principle, some of this guidance is provided, via Bayesian decision theory, through a desire on Your part to pay attention to how often You get the right answer, which is a central scientific question.

But the Calibration Principle is not enough: in problems of realistic complexity You’ll generally notice that (a) You’re uncertain about $\theta$ but (b) You’re also uncertain about how to quantify Your uncertainty about $\theta$, i.e., You have model uncertainty.

This acknowledgment of Your model uncertainty implies a willingness by You to consider two or more models in an ensemble $\mathcal{M} = \{M_1, M_2, \ldots \}$,
The Modeling-As-Decision Principle

which gives rise immediately to two questions:

\[
\begin{align*}
Q_1 & : \text{Is } M_1 \text{ better than } M_2? \\
Q_2 & : \text{Is } M_1 \text{ good enough?}
\end{align*}
\]

These questions sound fundamental but are not: better for what purpose? Good enough for what purpose? This implies (see, e.g., Bernardo and Smith, 1995; Draper, 1996; Key et al., 1999) a

\textbf{Modeling-As-Decision Principle:} Making clear the purpose to which the modeling will be put transforms model specification into a decision problem, which should be solved by maximizing expected utility with a utility function tailored to the specific problem under study.

Some examples of this may be found (e.g., Draper and Fouskakis, 2008: variable selection in generalized linear models under cost constraints), but this is hard work; there’s a powerful desire for generic model-comparison tools whose utility structure may provide a decent approximation to problem-specific utility elicitation.
Two such tools are

- **Bayes factors**, based on a utility structure in which you have to **pretend** that one of the **models** in $\mathcal{M}$ is the **actual data-generating mechanism** $M_{DG}$ and you **reward** yourself with $c > 0$ utiles if your **chosen** $M_{j^*}$ is $M_{DG}$ and 0 otherwise;

- **Log scores**, in which the utility function is based on predictive **accuracy**; an example, with the **simple data structure** $D = y = (y_1, \ldots, y_n)$, is the **full-sample log score**

$$LS_{FS}(M_j | y \mathcal{B}) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | y M_j \mathcal{B}).$$

I prefer log scores, for two reasons:

- The utility function underlying Bayes factors is a bit **far-fetched** in many applications, whereas the utility motivation for log scores is based on the (to me) reasonable
**Prediction Principle:** Good models make good predictions, and bad models make bad predictions; that’s one scientifically important way you know a model is good or bad.

- When the available information about $\theta$ external to $D$ in a model $M_j$ is weak relative to the information about $\theta$ internal to $D$ under $M_j$ (the so-called diffuse-prior situation), Bayes factors suffer from a serious problem of instability in how the diffuse prior information is specified, whereas log scores have no such instability.

Some people (e.g., Mukhopadhyay, Ghosh and Berger, 2005) have claimed that log scores are asymptotically inconsistent, but in my view this arises from a scientifically inappropriate problem formulation; when this is remedied (Draper and Krnjajić, 2010), log scores have no problem with large-sample model discrimination, and perform well calibratively in small samples too.
Conclusions

- I’ve offered an axiomatization of inferential, predictive and decision-theoretic statistics based on information, not belief, and RT Cox’s (1946) notion of probability as a measure of the weight of evidence in favor of the truth of a true-false proposition whose truth status is uncertain for You.

- Cox’s Theorem lays out a logical progression

  Principles → Axioms → Theorem

  to prove that

  If logical consistency then Bayesian reasoning;

  this secures the foundations of probability.

- But Cox’s Theorem does not go far enough for statistical work in science, in two ways related to model specification:

  — Nothing in its consequences requires You to pay attention to how often You get the right answer, which is a basic scientific concern, and
Conclusions (continued)

— it doesn’t offer any advice on how to specify the required ingredients: With $\theta$ as the unknown of principal interest, $\mathcal{B}$ as Your relevant background assumptions and judgments, and an information source (data set) $D$ relevant to decreasing Your uncertainty about $\theta$,

* $\{p(\theta|\mathcal{B}), p(D|\theta \mathcal{B})\}$ for inference and prediction, and

* in addition $\{\mathcal{A}, U(a, \theta)\}$ for decision, where $\mathcal{A}$ is Your set of available actions and $U(a, \theta)$ is Your utility function (mapping from actions $a$ and $\theta$ to real-valued consequences).

• To secure the foundations of statistics, work is needed laying out the logical progression

Principles $\rightarrow$ Axioms $\rightarrow$ Theorems

for model specification; progress in this area is part of the Theory of Applied Statistics.

• A Calibration Principle helps address the first of the two concerns above:
Conclusions (continued)

**Calibration Principle:** In model specification, You should pay attention to how often You get the right answer, by creating situations in which You know what the right answer is and seeing how often Your methods recover known truth.

- A Modeling-As-Decision Principle and a Prediction Principle help to address the second of the two concerns:

**Modeling-As-Decision Principle:** Making clear the purpose to which the modeling will be put transforms model specification into a decision problem, which should be solved by maximizing expected utility with a utility function tailored to the specific problem under study.

**Prediction Principle:** Good models make good predictions, and bad models make bad predictions; that’s one scientifically important way You know a model is good or bad.

- In problems of realistic complexity You’ll generally notice that (a) You’re uncertain about θ but (b) You’re also uncertain about how to quantify Your uncertainty about θ, i.e., You have model uncertainty.
• This acknowledgment of Your model uncertainty implies a willingness by You to consider two or more models in an ensemble $\mathcal{M} = \{M_1, M_2, \ldots \}$, which gives rise immediately to two questions:

\[
\begin{align*}
Q_1 & : \text{Is } M_1 \text{ better than } M_2? \\
Q_2 & : \text{Is } M_1 \text{ good enough?}
\end{align*}
\]

• The full-sample log score

\[
LS_{FS}(M_j | y \mathcal{B}) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | y M_j \mathcal{B}),
\]

which arises from a combination of the Modeling-As-Decision Principle and the Prediction Principle, is a good method for addressing $Q_1$; it’s completely stable with respect to diffuse-prior specification, and its calibration properties for model discrimination in both small- and large-sample settings (contrary to what some people have said) are good.