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

Center For Statistical Research & Methodology
U.S. Census Bureau

Workshop: 19–21 Sep 2011
(1) An axiomatization of statistics (Draper 2011).

(2) Foundations of probability seem (to me) to be secure: (RT Cox, 1946) Principles $\rightarrow$ Axioms $\rightarrow$ Theorem:

Logical consistency in uncertainty quantification $\rightarrow$ justification of Bayesian reasoning.

(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; two remaining challenges:

(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 $\rightarrow$ Axioms $\rightarrow$ Theorems.

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

(5) The Modeling-As-Decision Principle, the Prediction Principle and the Decision-Versus-Inference Principle help with 3 (b).
Example (Krnjajić, Kottas, Draper [KKD] 2008): In-home geriatric assessment (IHGA). In an experiment conducted in the 1980s (Hendriksen et al. 1984), 572 elderly people, representative of \( P = \{ \text{all non-institutionalized elderly people in Denmark} \} \), were randomized, 287 to a control (C) group (who received standard health care) and 285 to a treatment (T) group (who received standard care plus IHGA: a kind of preventive medicine in which each person’s medical and social needs were assessed and acted upon individually).

One important outcome was the number of hospitalizations during the two-year life of the study:

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Hospitalizations</th>
<th>( n )</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>( n_{C0} ) ( n_{C1} ) \ldots ( n_{Ck} )</td>
<td>( n_C = 287 )</td>
<td>( \bar{y}_C )</td>
<td>( s_C )</td>
</tr>
<tr>
<td>Treatment</td>
<td>( n_{T0} ) ( n_{T1} ) \ldots ( n_{Tk} )</td>
<td>( n_T = 285 )</td>
<td>( \bar{y}_T )</td>
<td>( s_T )</td>
</tr>
</tbody>
</table>

Let \( \mu_C \) and \( \mu_T \) be the mean hospitalization rates (per two years) in \( P \) under the \( C \) and \( T \) conditions, respectively.

Here are four statistical questions that arose from this study:
The Four Principal Statistical Activities

**Q₁:** Did IHGA reduce the **mean number of hospitalizations per two years** by an **amount** that was **large** in **practical** terms?

\[(\bar{y}_T - \bar{y}_C)\]

**Q₂:** Did IHGA reduce the **mean number of hospitalizations per two years** by an **amount** that was **large** in **statistical** terms?

\[(\mu_T - \mu_C)\]

**Q₃:** On the **basis** of this study, how **accurately** can You **predict** the **total decrease in hospitalizations** over a period of \(N\) years if IHGA were **implemented throughout Denmark**? **[prediction]**

**Q₄:** On the **basis** of this study, is the **decision** to **implement IHGA throughout Denmark** **optimal** from a **cost-benefit** point of view? **[decision-making]**

These questions **encompass** almost all of the **discipline of statistics**: describing a data set \(D\), generalizing outward inferentially from \(D\), predicting new data \(D^*\), and helping people make decisions in the presence of uncertainty (I include sampling/experimental design under decision-making; omitted: data quality assurance (QA), ...).
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

- $A = \text{(Barack Obama will be re-elected U.S. President in 2012), or}$

- $B = \text{(the in-hospital mortality rate for patients at hospital H admitted in calendar 2010 with a principal diagnosis of heart attack was between 5% and 25%).}$

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, or array) of real numbers, but in principle it could be almost anything (a function, an image of the surface of Mars, a phylogenetic tree, ...).

IHGA example: $\theta = \text{mean relative decrease } \left( \frac{\mu_T - \mu_C}{\mu_C} \right)$ in hospitalization rate in $\mathcal{P}$.

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, or array) of real numbers, but in principle it too could be almost anything (a movie, the words in a book, ...).
Examples of $\mathcal{B}$:

- In the **IHGA study**, based on the **experimental design**, $\mathcal{B}$ would include the **propositions**
  
  (Subjects were like a random sample from $\mathcal{P}$) and
  
  (Subjects were randomized into one of two groups, **treatment** (standard care + IHGA) or **control** (standard care)).

9 (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**.)

10 (implication) It follows from $[1-9]$ that **statistics** concerns itself principally with **five things** (omitted: description, data QA, ...):

(1) **Quantifying Your information** about $\theta$ **internal** to $D$ (given $\mathcal{B}$), and doing so **well** (this term is **not yet defined**);
(2) Quantifying Your information about $\theta$ external to $D$ (given $B$), and doing so well;

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

and using all Your information about $\theta$ (given $B$) to make

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

(5) Decisions about how to act sensibly, even though Your information about $\theta$ may be incomplete.

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

This question has two parts: probability and statistics; in my view, the probability foundations are secure, but the statistics foundations still need attending to.

Let’s look first at the probability foundations.
From the 1650s (Fermat, Pascal) through the 18th century (Bayes, Laplace) to the period 1860–1930 (Venn, Boole, von Mises), three different approaches for 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 Kolmogorov, de Finetti, and RT Cox.

— Kolmogorov (1933): following (and rigorizing) Venn, Boole and von Mises, probability is a function on (possibly some of) the subsets of a sample space $\Omega$ of uncertain possibilities, constrained to obey some reasonable axioms; this is excellent, as far as it goes, but many types of uncertainty cannot (uniquely, comfortably) be fit into this framework (examples follow).

Kolmogorov was trying to make precise the intuitive notion of repeatedly choosing a point at random in a Venn diagram and asking how frequently the point falls inside a specified set, i.e., his concept of probability had a repeated-sampling, frequentist character:
Frequentist Probability: Kolmogorov

“The basis for the applicability of the results of the mathematical theory of probability to real ‘random phenomena’ must depend on some form of the frequency concept of probability, the unavoidable nature of which has been established by von Mises in a spirited manner.”

*Example:* You’re about to roll a pair of dice and You regard this dice-rolling as fair, by which You mean that (in Your judgment) all \(6^2 = 36\) elemental outcomes in \(\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}\) are equally probable; then the Kolomogorov probability of snake eyes \((1, 1)\) exists and is unique (from Your fairness judgment), namely \(\frac{1}{36}\); but

*Example:* You’re a doctor; a new patient presents saying that he may be HIV positive; what’s the Kolmogorov probability that he is?

What’s \(\Omega\)? This patient is not the result of a uniquely-specifiable repeatable “random” process, he’s just a guy who walked into Your doctor’s office, and — throughout the repetitions of whatever repeatable phenomenon anyone might imagine — his HIV status is not fluctuating “randomly”: he’s either HIV positive or he’s not.
The closest You can come to making Kolmogorov’s approach work here is to imagine the set $\Omega$ of all people \{similar to this patient in all relevant ways\} and ask how often You’d get an HIV-positive person if You repeatedly chose one person at random from $\Omega$, but to make this operational You have to specify what You mean by “similar to, in all relevant ways,” and if You try to do this You’ll notice that it’s not possible to do so uniquely (in such a way that all other reasonable people would unanimously agree with You).

— de Finetti (1937): rigorizing Bayes, probability is a quantification of betting odds about the truth of a proposition, constrained to obey axioms guaranteeing coherence (absence of internal contradictions); this is more general than Kolmogorov — in fact, it’s as general as You can get: any statement about sets can be expressed in terms of propositions — but betting odds are not fundamental to science.

de Finetti made many important contributions — in particular, his concept of exchangeability (more on this later) is crucial in Bayesian modeling — but science is about information, not betting.
- **RT Cox** (1946): following Laplace, probability is a quantification of information about the truth of one or more propositions, constrained to obey axioms guaranteeing internal logical consistency; this is both fundamental to science and as general as You can get.

Cox’s goal was to identify what basic rules $p(A|B)$ — the plausibility (weight of evidence in favor) of (the truth of) $A$ given $B$ — should follow so that $p(A|B)$ behaves sensibly, where $A$ and $B$ are propositions with $B$ assumed 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):

- Suppose You’re willing to represent degrees of plausibility by real numbers (i.e., $p(A|B)$ 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.
Cox’s Principles and Axioms

— You always take into account all of the evidence you judge to be relevant to the plausibility assessment under consideration (this is the Bayesian version of objectivity).

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

- 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 $D$. 
Cox’s Theorem

From these axioms Cox proved a theorem showing that uncertainty quantification about propositions behaves in one and only one way:

**Theorem:** If you accept Cox’s axioms, then to be logically consistent you must quantify uncertainty as follows:

- Your plausibility operator \( pl(A|B) \) — 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.

- **(normalization)** \( P(A|B) + P(\bar{A}|B) = 1 \), where \( \bar{A} = \) (not \( A \)).

- **(the product rule):**

\[
P(AB|C) = P(A|C) \cdot P(B|AC) = P(B|C) \cdot P(A|BC).
\]

The proof (see, e.g., Jaynes (2003)) involves deriving two functional equations \( F[F(x, y), z] = F[x, F(y, z)] \) and \( x S \left[ \frac{S(y)}{x} \right] = y S \left[ \frac{S(x)}{y} \right] \) that \( pl(A|B) \) must satisfy and then solving those equations.

A number of important corollaries arise from Cox’s Theorem:
(the sum rule):
\[ P(A \text{ or } B | C) \equiv P(A + B | C) = P(A | C) + P(B | C) - P(A B | C). \]

Extensions of the product and sum rules to an arbitrary finite number of propositions are easy, e.g.,

\[ P(A B C | D) = P(A | D) \cdot P(B | A D) \cdot P(C | A B D) \text{ and} \]
\[ P(A + B + C | D) = P(A | D) + P(B | D) + P(C | D) - P(A B | D) - P(A C | D) - P(B C | D) + P(A B C | D). \]

This framework (obviously) covers optimal reasoning about uncertain quantities \( \theta \) taking on a finite number of possible values; less obviously, it also handles (equally well) situations in which the set \( \Theta \) of possible values of \( \theta \) has infinitely many elements.

---

**Example:** You’re studying quality of care at the 17 Kaiser Permanente (KP) northern California hospitals in 2003–7, before the era of electronic medical records; during that time there was a population \( \mathcal{P} \) of \( N = 8,561 \) patients at these facilities with a primary admission diagnosis of heart attack.
You take a simple random sample of \( n = 112 \) of these admissions and record whether or not each patient had an unplanned transfer to the intensive care unit (ICU), observing \( s = 4 \) who did; \( \theta \) is the proportion of such unplanned transfers in all of \( P \); here \( \Theta = \{ \frac{0}{N}, \frac{1}{N}, \ldots, \frac{N}{N} \} \), which can be conveniently approximated by \( \Theta' = [0, 1] \).

Prior to 2003, the proportion of such unplanned transfers for heart attack patients at KP in the northern California region was about \( q = 0.07 \), so interest focuses on \( P(A|DB) \), where \( A \) is the proposition \( (\theta \leq q) \), \( D \) is the proposition \( (s = 4) \), and \( B \) includes (among other things) details about the sampling experiment (e.g., \( (n = 112) \)).

In this setup \( \theta \) is usually called a (population) parameter, and is not itself the result of any sampling experiment (random or otherwise); for this reason, it’s not possible to (directly) quantify uncertainty about \( \theta \) from the Kolmogorov (set-theoretic) point of view, but it makes perfect sense to do so from the RT Cox (propositional) point of view.
You could now more generally define a function
\[ F_{(\theta|DB)}(q) = P(\theta \leq q|DB) \]
and call it the **cumulative distribution function (CDF)** for (not of) \( \theta|DB \), which is shorthand for the CDF for Your uncertainty about \( \theta \) given \( D \) and \( B \).

If \( F_{(\theta|DB)}(q) \) turns out to be **continuous** and **differentiable** in \( q \) (I haven’t said yet how to calculate \( F \)), it will be convenient to write
\[
F_{(\theta|DB)}(b) - F_{(\theta|DB)}(a) = P(a < \theta \leq b|DB) = \int_a^b p_{(\theta|DB)}(q) \, dq, \tag{1}
\]
where the **(partial) derivative** \( p_{(\theta|DB)}(q) \) of \( F_{(\theta|DB)} \) with respect to \( q \) can be called the **density** for (not of) (Your uncertainty about) \( \theta \) given \( D \) and \( B \).

In a small abuse of notation it’s common to write \( F(\theta|DB) \) and \( p(\theta|DB) \) instead of \( F_{(\theta|DB)}(q) \) and \( p_{(\theta|DB)}(q) \) (respectively), letting the argument \( \theta \) of \( F(\cdot|DB) \) and \( p(\cdot|DB) \) serve as a **reminder** of the uncertain quantity in question.
In the Kolmogorov approach a random variable $X$ is a function from $\Omega$ to some outcome space $O$, and if $O = \mathbb{R}$ you’ll often find it useful to summarize $X$’s behavior through the CDF of $X$: $F_X(x) = P(\text{the set of } \omega \in \Omega \text{ such that } X(\omega) \leq x)$, usually written in propositional-style shorthand as $F_X(x) = P(X \leq x)$.

In the RT Cox approach, there are no random variables; there are uncertain things $\theta$ whose uncertainty (when $\Theta = \mathbb{R}^k$, for integer $1 \leq k < \infty$) can usefully be summarized with CDFs and densities.

Jaynes (2003) makes a worthwhile distinction: the statements

| There is noise in the room. | The room is noisy. |

seem quite similar but are in fact quite different: the former is ontological (asserting the physical existence of something), whereas the latter is epistemological (expressing the personal perception of the individual making the statement).

Talking about “the density of $\theta$” would be to confuse ontology and epistemology;
The Mind-Projection Fallacy

Jaynes calls this confusion of \{the world\} (ontology) with \{Your uncertainty about the world\} (epistemology) the mind-projection fallacy, and it’s clearly a mistake worth avoiding.

Returning to the corollaries of Cox’s Theorem,

- Given the 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) It’s natural (and indeed You must be prepared in this approach) 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 \( p(\theta|\mathcal{D} \mathcal{B}) \) 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

\[ p(\theta | D B) = c \ p(\theta | B) \ p(D | \theta B), \quad \text{(inference)} \]  \hspace{1cm} (2)

where \( c > 0 \) is a normalizing constant chosen so that the left-hand side of (2) integrates (or sums) over \( \Theta \) to 1;

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

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

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

\[ p(D^* | D B) = \int_{\Theta} p(D^* | \theta B) \ p(\theta | D B) d\theta ; \quad \text{(prediction)} \]  \hspace{1cm} (3)

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

(i) the set \( A \) of 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 $\theta$ — without loss of generality You can take large values of $U(a, \theta)$ to be better than small values;

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

$$a^* = \arg\max_{a \in A} E_{(\theta|DB)} U(a, \theta) = \arg\max_{a \in A} \int_{\Theta} U(a, \theta) p(\theta|DB) d\theta. \quad (4)$$

The equation solving the inference problem is traditionally attributed to Bayes (1764), although it’s just an application of the product rule (page 14), which was already in use by (James) Bernoulli and de Moivre around 1715, and Laplace made much better use of this equation from 1774 to 1827 than Bayes did in 1764; nevertheless the Laplace/Cox propositional approach is typically referred to as Bayesian reasoning.
Cox’s Theorem is equivalent to the assertion

If You wish to quantify Your uncertainty about an unknown $\theta$ (and make predictions and decisions in the presence of that uncertainty) in a logically internally consistent manner (as specified through Cox’s axioms), on the basis of data $D$ and background assumptions/judgments $B$, then You can achieve this goal with Bayesian reasoning, by specifying $p(\theta|B)$, $p(D|\theta B)$, and $\{A, U(a, \theta)\}$ and using equations (2–4).

This assertion has not rendered Bayesian analyses ubiquitous, although the value of Bayesian reasoning has become increasingly clear to an increasingly large number of people in the last 20 years, now that advances in computing have made the routine use of equations (2–4) feasible.

Advantages include a unified probabilistic framework: e.g., in my earlier ICU example, Kolmogorov’s non-Bayesian approach does not permit direct probability statements about a population parameter, but Cox’s Theorem permits You to make such statements (summarizing all relevant available information) in a natural way.
It’s **worth noting**, however, that **there really is a theorem here**, of the form \( A \rightarrow B \), from which \( \overline{B} \rightarrow \overline{A} \); this **comes close to the assertion**

If You employ **non-Bayesian reasoning** then You’re **open to the possibility of logical inconsistency**, and indeed there have been some **embarrassing moments** in **non-Bayesian inference** over the past 100 years (e.g., **negative estimates** for quantities that are **constrained** to be **non-negative**).

**Challenges:** These **corollaries to Cox’s theorem** solve problems (3–5) above (page 8) — 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**

- \( p(\theta|B) \), **usually called** Your **prior information** about \( \theta \) (given \( B \); 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);
The Specification Burden (continued)

- \( p(D|\theta B) \), often referred to as Your **sampling distribution** for \( D \) given \( \theta \) (and \( B \); 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

- \( p(\theta|DB) \), often referred to as Your **posterior** distribution for \( \theta \) given \( D \) (and \( B \); 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** \( p(D^*|DB) \) for future data \( D^* \) given the **observed data set** \( D \); and

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

**To summarize:** Inference and prediction require You to **specify** \( p(\theta|B) \) and \( p(D|\theta B) \); decision-making requires You to **specify** the same
two ingredients plus $\mathcal{A}$ 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 (some of them good), 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, which we need but do not yet have.

$p(\theta|B)$, $p(D|\theta B)$ and $\{A, U(a, \theta)\}$ are all important; I’ll focus here on the problem of specifying $\{p(\theta|B), p(D|\theta B)\}$ — call such a specification a model $M$ for Your uncertainty about $\theta$ (I’ll have one brief comment about decision theory at the end).
The 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, I’ll focus in the rest of this Workshop on four principles worth considering, the first of which is the

**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, and repeatedly getting the wrong answer runs counter to this desire.
Reasoning Behind the Calibration Principle

(remark) There’s nothing in the Bayesian paradigm to prevent You from making one or both of the following mistakes — (a) choosing $p(D|\theta B)$ badly; (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 repeatedly doing this violates the axiom above.

(remark) Paying attention to calibration is a natural activity from the frequentist point of view, but a desire to be well-calibrated 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 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, 2011) calibration-monitoring emerges as a natural and inevitable Bayesian activity.

This seems to be a new idea: logical consistency justifies 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 activity.

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.

Cox’s Theorem says that You can draw logically-consistent inferences about an unknown \( \theta \), given data \( D \) and background information \( B \), by specifying \( M = \{ p(\theta|M|B), p(D|\theta|M|B) \} \), but item (b) in the previous paragraph implies that there will typically be more than one such plausible \( M \); what should You do about this?

It would be nice to be able to solve the inference problem by using Bayes’s Theorem to compute \( p(\theta|D,M_{\text{all}}|B) \), where \( M_{\text{all}} \) is the set of all possible models, but this is not feasible: just as Kolmogorov had to resort to \( \sigma \)-fields because the set of all subsets of an \( \Omega \) with uncountably many elements is too big to meaningfully assign probabilities to all of the subsets, with a finite data set \( D \),
\( \mathcal{M}_{\text{all}} \) is too big for \( D \) to permit meaningful plausibility assessment of all the models in \( \mathcal{M}_{\text{all}} \).

Having adopted the **Calibration Principle**, it makes sense to talk about an **underlying data-generating model** \( M_{DG} \), which is unknown to You (more on this below).

Not being able to compute \( p(\theta|D \mathcal{M}_{\text{all}} \mathcal{B}) \), in practice the best You can do is to compute \( p(\theta|D \mathcal{M} \mathcal{B}) \), where \( \mathcal{M} \) is an **ensemble of models** (finite or countably or uncountably infinite) chosen “well” by You, where “well” can and should be brought into focus by the **Calibration Principle** (and some of the other Principles to be introduced later): evidently what You want, among other things, is for \( \mathcal{M} \) to contain one or more models that are identical (or at least close) to \( M_{DG} \) (in a sense I’ll make precise below).

Suppose initially, for the sake of discussion, that You’ve identified such an ensemble (I’ll present some ideas for how to do this later) and that it turns out to be finite: \( \mathcal{M} = (M_1, \ldots, M_k) \) for \( 2 \leq k < \infty \); what next?
Are You **supposed** to try to **choose** one of these **models** (the **model selection problem**) and **discard** the rest, or **combine** them in some way (if so, **how**?), or **what**?

To move toward an **answer** to this **question**, suppose (continuing the Kaiser example on page 15) that You also **observe** (for each of the \( n = 112 \) randomly-sampled patients from the population \( \mathcal{P} \) of \( N = 8,561 \) heart-attack patients) a real-valued conceptually-continuous non-negative quality-of-care score \( y_i \), and **inferential interest** focuses on the **mean** \( \theta \) of these **scores** in \( \mathcal{P} \); here the **data set** \( D \) is just \( y = (y_1 \ldots y_n) \).

One possible **Bayesian parametric model** for this setting is

\[
M_1: \left\{ \begin{array}{c}
(\theta \sigma^2 | M_1 B) \sim p(\theta \sigma^2 | M_1 B) \\
(y_i | \theta \sigma^2 M_1 B) \overset{\text{IID}}{\sim} \text{Gaussian}(\theta, \sigma^2)
\end{array} \right\},
\]

for some **scientifically appropriate prior distribution** \( p(\theta \sigma^2 | M_1 B) \); another possible **parametric model** is

\[
M_2: \left\{ \begin{array}{c}
(\theta \tau^2 | M_2 B) \sim p(\theta \tau^2 | M_2 B) \\
(y_i | \theta \tau^2 M_2 B) \overset{\text{IID}}{\sim} \text{Lognormal}(\theta, \tau^2)
\end{array} \right\},
\]
with the **Lognormal distribution parameterized** so that $\theta$ and $\tau^2$ are the **mean** and **variance** on the $y$ (rather than $\log y$) scale.

I’ll use the **notation** $\gamma_j = (\theta, \eta_j)$ for the **parameter vector** (of **length** $k_j$) for model $M_j$, where each model has its own **vector** of so-called **nuisance parameters** $\eta_j$: here $\eta_1 = (\sigma^2)$ and $\eta_2 = (\tau^2)$.

By the **Product Rule**, $p(\theta \eta_j | M_j B) = p(\theta | M_j B) p(\eta_j | \theta M_j B)$, and the priors $p(\theta | M_j B)$ are the **same** for all $j$ (and can therefore just be **referred to** as $p(\theta | B)$); thus, in this **setting**, in which **two or more parametric models** may be **plausible**, **model uncertainty** has **three parts**: the **prior** $p(\theta | B)$ on $\theta$, the **conditional prior** on the **nuisance parameters** $p(\eta_j | \theta M_j B)$, and the **sampling distribution** (in this case, **Gaussian** ($j = 1$) or **Lognormal** ($j = 2$)).

As noted above, under the **Calibration Principle** it makes sense to talk about an **underlying data-generating model** $M_{DG}$, which is **unknown to You**; an example here might be

$$M_{DG}: y_i \overset{\text{IID}}{\sim} \text{Gaussian}(\theta_{DG}, \sigma^2_{DG}), \quad (7)$$

with (e.g.) $(\theta_{DG}, \sigma^2_{DG}) = (50, 10^2)$; I’ll use the **notation**
γ_{DG} = (θ_{DG}, η_{DG}) for the parameter vector of M_{DG}.

The fact that M_{DG} is unknown to You presents a challenge in both Bayesian and non-Bayesian paradigms; the form this challenge takes in the Bayesian approach can be seen by examining the following argument:

- All Bayesian reasoning under uncertainty is based on
  \[ P(A|B) = \frac{P(A \cap B)}{P(B)} \]
  for true/false propositions A and B, and this is undefined if B is false; this gives rise to

**Rule 1:** You should try hard not to condition on propositions (a) that You know to be false and (b) that MAY be false.

- Choosing a specific model M_j amounts to conditioning on it; in other words, in practice You may want to compute \( p(θ|D B) \), but by choosing M_j You’re really computing \( p(θ|D M_j B) \).

- Having chosen a particular model M_j (say), this makes me wonder what happens when M_j \( \neq \) M_{DG}, because in that case choosing M_j sounds like conditioning on a false proposition.
• However, it’s not quite meaningful to write something like $M_j \neq M_{DG}$, because the sampling-distribution part of $M_j$ actually contains many models from an $M_{DG}$ perspective; in the Gaussian-Lognormal example above, for instance, $M_{DG}$ specifies the single model $N(50, 10^2)$ but $p(y_i|\theta \sigma^2 M_1 B)$ specifies $N(\theta, \sigma^2)$ for all $(\theta, \sigma^2)$ in the support of the prior $p(\theta, \sigma^2|M_1 B)$ (i.e., all $(\theta, \sigma^2)$ such that $p(\theta, \sigma^2|M_1 B) \neq 0$).

• Theorem (Doob, 1948): In repeated sampling under $M_{DG}$, as $n$ increases, the posterior distribution $p(\theta|D M_j B)$ becomes more and more concentrated around \{point mass at $\theta_{DG}$\}, as long as $\theta_{DG}$ is in the support of $p(\theta|M_j B)$ (this theorem demonstrates what’s known as asymptotic consistency of Bayesian inference).

• This theorem motivates the following

**Definition** (Draper 2011): $M_j$ is consistent with $M_{DG}$ ($M_j \bowtie M_{DG}$) if
(a) the support of $p(\gamma_j|M_j B)$ includes $\gamma_{DG}$ and (b) $p(D|\gamma_{DG} M_j B) = p(D|M_{DG})$. 


Intuitively $M_j \overset{c}{=} M_{DG}$ means that (a) Your prior on the parameters includes the data-generating parameter values as valid possibilities and (b) You got the sampling distribution right.

So now the correct wording of the question is: what happens if I choose $M_j$ but (unknown to me) $M_j \not\overset{c}{=} M_{DG}$?

Good news — what happens is not like conditioning on a false proposition (i.e., not like dividing by 0); (possibly) bad news —

**Theorem** (Berk, 1964): if $M_j \not\overset{c}{=} M_{DG}$, then as $n$ increases, the posterior distribution $p(\theta|D M_j B)$ becomes more and more concentrated around \{point mass at $\theta^*$\}, where $\gamma_j^* = (\theta^*, \eta_j^*)$ and $\theta^*$ is such that $p(D|\gamma_j^* M_j B)$ is as close as possible to $p(D|\gamma_{DG} M_{DG})$ (here closeness is measured by Kullback-Leibler (KL) divergence: for densities $p$ and $q$, $D_{KL}(p||q) = \int p \log \frac{p}{q}$).

In the Gaussian-Lognormal example, if $M_{DG}$ is Lognormal($\theta_{DG}, \tau_{DG}^2$) but You choose as Your model Gaussian($\theta, \sigma^2$), with more data it will look increasingly to You as though $M_{DG}$ is Gaussian($\theta^*, \sigma_*^2$), where ($\theta^*, \sigma_*^2$) is such that Gaussian($\theta^*, \sigma_*^2$) minimizes the KL divergence.
Model Mis-Specification (continued)

from Lognormal($\theta_{DG}, \tau_{DG}^2$).

It’s nice that $p(D|\gamma_j^* M_j B)$ is as close as possible to $p(D|\gamma_{DG} M_{DG})$, but this provides no guarantee that they are in fact close; the point is that model mis-specification can have serious inferential consequences in both Bayesian and non-Bayesian paradigms.

Having introduced this idea of a model $M_j$ being consistent (or not) with an underlying data-generating mechanism $M_{DG}$, it would be nice — from a calibration point of view — to be able to compute $p(\theta|D M_c B)$, where $M_c$ includes all models $M_j$ such that $M_j \equiv M_{DG}$;

Q: Are there any Bayesian approaches that can achieve this goal?

A: Bayesian nonparametric methods can come close, in large samples (more on this below).

Solving the model uncertainty problem. People used to “solve” the problem of what to do about model uncertainty by ignoring it: it was common, at least through the mid-1990s, to

(a) use the data $D$ to conduct a search among possible models,
settling on a single (apparently) “best” model $M^*$ arising from the search, and then

(b) draw inferences about $\theta$ pretending that $M^* \subseteq M_{DG}$.

This of course can lead to quite bad calibration, almost always in the direction of pretending You know more than You actually do, so that, e.g., Your nominal 90% posterior predictive intervals for data values not used in the modeling process would typically include substantially fewer than 90% of the actual observations.

The $M^*$ approach “solves” the problem of how to specify $\mathcal{M}$ by setting $\mathcal{M} = \{M^*\}$; I’ll continue to postpone for the moment how You might do a better job of arriving at $\mathcal{M}$.

Having chosen $\mathcal{M}$ in some way, how can You assess Your uncertainty across the models in $\mathcal{M}$, and appropriately propagate this through to Your uncertainty about $\theta$, in a well-calibrated way?

I’m aware of three approaches to improved assessment and propagation of model uncertainty: BMA, BNP, CCV.
**Bayesian model averaging (BMA):** If interest focuses on something that has the same meaning across all the models in $\mathcal{M}$ — for example, a set of future data values $D^*$ to be predicted — calculation reveals (e.g., Draper 1995, on the workshop webpage) that

$$p(D^* | D \mathcal{M} B) = \int_{\mathcal{M}} p(D^* | D M B) p(M | D \mathcal{M} B) \, dM ,$$

which is *eminently reasonable*: equation (8) tells you to form a weighted average of your conditional predictive distributions $p(D^* | D M B)$, given particular models $M \in \mathcal{M}$, weighted by those models’ posterior probabilities $p(M | D \mathcal{M} B)$.

This approach typically provides (substantially) better calibration than that obtained by the $M^*$ method.

**Bayesian nonparametric (BNP) modeling:** The BMA integral in (8) can be thought of as an approximation to the (unattainable?) ideal of averaging over all worthwhile models; a better approximation to this ideal can often be achieved with Bayesian nonparametric modeling, which dates back to de Finetti (1937).
Following the discussion on page 30, continuing the Kaiser example on page 15, suppose you also observe (for each of the $n = 112$ randomly-sampled patients from the population $\mathcal{P}$ of $N = 8,561$ heart-attack patients) a real-valued conceptually-continuous quality-of-care score $y_i$, and (following de Finetti) you’re thinking about your predictive distribution $p(y_1 \ldots y_n | \mathcal{B})$ for these scores before any data have arrived.

De Finetti pointed out that, if you have no covariate information about the patients, your predictive distribution $p(y_1 \ldots y_n | \mathcal{B})$ should remain the same under arbitrary permutation of the order in which the patients are listed, and he coined the term exchangeability to describe this state of uncertainty.

He (and later Diaconis/Freedman) went on to prove that, if your judgment of exchangeability extends from $(y_1 \ldots y_n)$ to $(y_1 \ldots y_N)$ (as it certainly should here, given the random sampling) and $N \gg n$ (as is true here), then all logically-internally-consistent predictive distributions can approximately be expressed hierarchically as follows:
letting \( F \) stand for the \textbf{empirical CDF} of the \textbf{population values} \((y_1 \ldots y_N)\), the \textbf{hierarchical model} is (for \( i = 1, \ldots, n \))

\[
\begin{cases}
(F | \mathcal{B}) \sim p(F | \mathcal{B}) \\
(y_i | F, \mathcal{B}) \overset{\text{iid}}{\sim} F
\end{cases}
\]

This requires placing a \textbf{scientifically-appropriate prior distribution} \( p(F | \mathcal{B}) \) on the set \( \mathcal{F} \) of all CDFs on \( \mathbb{R} \), which de Finetti didn’t know how to do in 1937; thanks to work by Freedman, Ferguson, Lavine, Escobar/West, and others, \textbf{two methods} for doing this \textbf{sensibly} — Pólya trees and \textbf{Dirichlet-process (DP) priors} — are now in \textbf{routine use}: this — placing \textbf{distributions} on function spaces — is \textbf{Bayesian nonparametric} (BNP) modeling.

\textbf{IHGA Example, Revisited:} Visualizing the IHGA data set before it \textbf{arrives}, it would look like the \textbf{table shell} presented back on page 3:

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Hospitalizations</th>
<th>( n )</th>
<th>Mean ( \bar{y} )</th>
<th>SD ( s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>( n_{C0} ) \ldots ( n_{Ck} )</td>
<td>( n_C = 287 )</td>
<td>( \bar{y}_C )</td>
<td>( s_C )</td>
</tr>
<tr>
<td>Treatment</td>
<td>( n_{T0} ) \ldots ( n_{Tk} )</td>
<td>( n_T = 285 )</td>
<td>( \bar{y}_T )</td>
<td>( s_T )</td>
</tr>
</tbody>
</table>
Letting (as before) $\mu_C$ and $\mu_T$ be the mean hospitalization rates (per two years) in the population $\mathcal{P}$ (of all elderly non-institutionalized people in Denmark in the early 1980s) under the $C$ and $T$ conditions, respectively, the inferential quantity of main interest is still

$$\theta = \frac{\mu_T - \mu_C}{\mu_C}$$

(or this could be redefined without loss as $\theta = \frac{\mu_T}{\mu_C}$); how can You draw valid and accurate inferences about $\theta$ while coping with Your uncertainty about the population $C$ and $T$ CDFs — call them $F_C$ and $F_T$, respectively — of numbers of hospitalizations per person (per two years)?

**One approach:** Bayesian “distribution-free” inference (Draper 2005; see the workshop web page).

**Another approach:** Bayesian nonparametric modeling — it turns out that DP priors put all their mass on discrete distributions, so one BNP model for this data set would involve placing parallel DPs priors on $F_C$ and $F_T$; see my encyclopedia article and KKD (2008) on the workshop web page for details on the results.
To serve as the **basis** of the $M^*$ (**cheating**) approach (in which You look at the data for **inspiration** on which models to fit), here’s a **table** of the **actual data values**:

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Hospitalizations</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>138 77 46 12 8 4 0 2</td>
<td>287</td>
<td>0.944</td>
<td>1.24</td>
</tr>
<tr>
<td>Treatment</td>
<td>147 83 37 13 3 1 1 0</td>
<td>285</td>
<td>0.768</td>
<td>1.01</td>
</tr>
</tbody>
</table>

Evidently **(description)** IHGA lowered the **mean hospitalization rate** (for **these elderly Danish people**, at least) by $(0.944 - 0.768) = 0.176$, which is a $\{100 \left( \frac{0.768 - 0.944}{0.944} \right) \div \} 19\%$ reduction from the **control level**, a difference that’s **large in clinical terms**, but **(inference)** how **strong** is the **evidence** for a **positive effect** in $\mathcal{P} = \{\text{all people similar to those in the experiment}\}$?

It’s **natural** to think **initially** of parallel $\text{Poisson}(\lambda_C)$ and $\text{Poisson}(\lambda_T)$ modeling ($M_1$), but there’s **substantial over-dispersion**: the $C$ and $T$ variance-to-mean ratios are $\frac{1.24^2}{0.944} \doteq 1.63$ and $\frac{1.01^2}{0.768} \doteq 1.33$. 
Unfortunately we have no covariates to help explain the extra-Poisson variability, and there’s little information external to the data set about the treatment effect; this latter state of knowledge is expressed in prior distributions on parameters by making them diffuse (i.e., ensuring they have large variability to express substantial uncertainty).

In this situation You could fit parallel Negative Binomial models ($M_2$), but a parametric choice that more readily generalizes is obtained by letting $(x_i, y_i) = (\text{C/T status, outcome})$ — so that $x_i = 1$ if Treatment, 0 if Control and $y_i =$ the number of hospitalizations — for person $i = 1, \ldots, n$ and considering the random-effects Poisson regression model ($M_3$):

$$
(y_i | \lambda_i, M_3 B) \overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i)
$$

$$
\log(\lambda_i) = \gamma_0 + \gamma_1 x_i + \epsilon_i
$$

$$
(\epsilon_i | \sigma_\epsilon^2, M_3 B) \overset{\text{IID}}{\sim} N(0, \sigma_\epsilon^2)
$$

$$
(\gamma_0, \gamma_1, \sigma_\epsilon^2 | M_3 B) \sim \text{diffuse}.
$$

In this model the unknown of main policy interest is
\[ \theta = \frac{\text{population } \tilde{T}}{\text{population } \tilde{C}} = e^{\gamma_1}; \text{ the other parameters can be collected in a} \]

\[ \text{vector } \eta = (\gamma_0, \sigma^2_\epsilon); \text{ and the random effects } \epsilon_i \text{ can be thought of as} \]

\[ \text{proxying for the combined main effect } \sum_{j=2}^{J} \gamma_j (x_{ij} - \bar{x}_j) \text{ of all the} \]

\[ \text{unobserved relevant covariates (age, baseline health status, ...).} \]

The first line of (9) makes good scientific sense (the \( y_i \) are counts of relatively rare events), but the Gaussian assumption for the random effects is conventional and not driven by the science; a potentially better model \((M_4)\) is obtained by putting a prior distribution on the CDF of the \( \epsilon_i \) that’s centered at the \( N(0, \sigma^2_\epsilon) \) distribution but that expresses substantial prior uncertainty about the

Gaussian assumption:

\[
\begin{align*}
(y_i | \lambda_i, M_4, B) & \sim \text{Poisson}(\lambda_i) \\
\log(\lambda_i) & = \gamma_0 + \gamma_1 x_i + \epsilon_i \\
(\epsilon_i | F, M_4, B) & \sim F \\
(F | \alpha, \sigma^2_\epsilon, M_4, B) & \sim DP(\alpha, F_0), \ F_0 = N(0, \sigma^2_\epsilon) \\
(\gamma_0, \gamma_1, \sigma^2_\epsilon | M_4, B) & \sim \text{diffuse}; (\alpha | M_4) \sim \text{small positive}.
\end{align*}
\]
Many Bayesian prior distributions $p(\theta|M_j B)$ have two user-friendly inputs: a quantity $\theta_0$ that acts like a prior estimate of the unknown $\theta$, and a number $n_0$ that behaves like a prior sample size (i.e., a measure of how tightly the prior is concentrated around $\theta_0$); DP priors are no exception to this pattern.

In equation (10), $DP(\alpha, F_0)$ is a Dirichlet-process prior on $F$ with prior estimate $F_0 = N(0, \sigma^2\epsilon)$ and a quantity ($\alpha$) that behaves something like a prior sample size; this is referred to as Dirichlet-process mixture modeling, because (10) is a mixture model — each person in the study has her/his own $\lambda$, drawn from $F_C$ (control) or $F_T$ (treatment) — in which uncertainty about $F_C$ and $F_T$ is quantified via a DP.

NB Bayesian model averaging (BMA) with a finite set of models can be regarded as a crude approximation to what Bayesian nonparametric (BNP) modeling is trying to do, namely average over Your uncertainty in model space to provide an honest representation of Your overall uncertainty that doesn’t condition on things You don’t know are true.
• Calibration cross-validation (CCV): The way the IHGA example unfolded looks a lot like the $M^*$ approach I condemned previously: I used the entire data set to suggest which models to consider.

This has the (strong) potential to underestimate uncertainty; Bayesians (like everybody else) need to be able to look at the data to suggest alternative models, but all of us need to do so in a way that’s well-calibrated.

Cross-validation — partitioning the data (e.g., exchangeably) into subsets used for different tasks (modeling, validation, ...) can help.

— The $M^*$ approach is an example of what might be called 1CV (one-fold cross-validation): You use the entire data set $D$ both to model and to see how good the model is (this is clearly inadequate).

— 2CV (two-fold cross-validation) is frequently used: You (a) partition the data into modeling (M) and validation (V) subsets, (b) use M to explore a variety of models until You’ve found a “good” one $M^*$, and (c) see how well $M^*$ validates in V (a useful Bayesian way to do this is to use the data in M
Calibration Cross-Validation (CCV) to construct **posterior predictive distributions for all of the data values** in V and see how the latter compare with the former).

2CV is a lot better than 1CV, but what do You do (as frequently happens) if $M^*$ doesn’t validate well in V?

— **CCV (calibration cross-validation):** going out one more term in the Taylor series (so to speak),

(a) **partition** the data into **modeling** (M), **validation** (V) and **calibration** (C) subsets,

(b) use M to explore a **variety of models** until You’ve found one or more plausible candidates $\mathcal{M} = \{M_1, \ldots, M_m\}$,

(c) see **how well** the models in $\mathcal{M}$ **validate** in V,

(d) if **none of** them do, **iterate (b) and (c)** until You do get **good validation**, and

(e) **fit** the **best model** in $\mathcal{M}$ (or, better, use BMA) on the data in $M + V$, and report both (i) **inferential conclusions** based on this fit and (ii) the **quality of predictive calibration of Your model/ensemble**) in C.
The goal with this method is both

(1) a good answer, to the main scientific question, that has paid a reasonable price for model uncertainty (the inferential answer is based only on $M + V$, making Your uncertainty bands wider) and

(2) an indication of how well calibrated {the iterative fitting process yielding the answer in (1)} is in $C$ (a good proxy for future data).

You can use decision theory (Draper, 2011) to decide how much data to put in each of $M$, $V$ and $C$: the more important calibration is to You, the more data You want to put in $C$, but only up to a point, because getting a good answer to the scientific question is also important to You.

This is related to the machine-learning practice (e.g., Hastie, Tibshirani, Friedman [HTF] 2009) of Train/Validation/Test partitioning, with one improvement (decision theory provides an optimal way to choose the data subset sizes); I don’t agree with HTF that this can only be done with large data sets: it’s even more important to do it with small and medium-size data sets (You just need to work with multiple ($M$, $V$, $C$) partitions and average).
CCV provides a way to pay the right price for hunting around in the data for good models, motivating the following modeling algorithm:

(a) Start at a model $M_0$ (how choose?); set the current model $M_{\text{current}} \leftarrow M_0$ and the current model ensemble $M_{\text{current}} \leftarrow \{M_0\}$.

(b) If $M_{\text{current}}$ is good enough to stop (how decide?), return $M_{\text{current}}$; else

(c) Generate a new candidate model $M_{\text{new}}$ (how choose?) and set $M_{\text{current}} \leftarrow M_{\text{current}} \cup M_{\text{new}}$.

(d) If $M_{\text{new}}$ is better than $M_{\text{current}}$ (how decide?), set $M_{\text{current}} \leftarrow M_{\text{new}}$.

(e) Go to (b).

For human analysts the choice in (a) is not hard, although it might not be easy to automate in full generality; for humans the choice in (c) demands creativity, and as a profession, at present, we have no principled way to automate it; here I want to focus on the questions in (b) and (d):

$Q_1$: Is $M_1$ better than $M_2$?  
$Q_2$: Is $M_1$ good enough?
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

**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 methods whose utility structure may provide a decent approximation to problem-specific utility elicitation.

Two such methods are Bayes factors and log scores.

- **Bayes factors.** It looks natural to compare models on the basis of their posterior probabilities; from Bayes’s Theorem in odds form,
Bayes Factors

\[
\frac{p(M_2 | DB)}{p(M_1 | DB)} = \left[ \frac{p(M_2 | B)}{p(M_1 | B)} \right] \cdot \left[ \frac{p(D | M_2 B)}{p(D | M_1 B)} \right]; \tag{11}
\]

the **first term** on the right is just the **prior odds** in favor of \( M_2 \) over \( M_1 \), and the **second term** on the right is called the **Bayes factor**, so in words equation (11) says

\[
\begin{pmatrix}
\text{posterior odds} \\
\text{for } M_2 \\
\text{over } M_1
\end{pmatrix}
= 
\begin{pmatrix}
\text{prior odds} \\
\text{for } M_2 \\
\text{over } M_1
\end{pmatrix}
\cdot 
\begin{pmatrix}
\text{Bayes factor} \\
\text{for } M_2 \\
\text{over } M_1
\end{pmatrix}. \tag{12}
\]

(Bayes factors seem to have **first** been **considered** by Turing and Good (\( \sim 1941 \)), as part of the effort to **break** the German Enigma codes.)

**Odds** \( o \) are related to **probabilities** \( p \) via \( o = \frac{p}{1-p} \) and \( p = \frac{o}{1+o} \); these are **monotone increasing transformations**, so the decision rules

\{choose \( M_2 \) over \( M_1 \) if the **posterior odds** for \( M_2 \) are greater\} and

\{choose \( M_2 \) over \( M_1 \) if \( p(M_2 | DB) > p(M_1 | DB) \)\} are equivalent.
This approach does have a decision-theoretic basis, but it’s rather odd: if you pretend that the only possible data-generating mechanisms are $\mathcal{M} = \{M_1, \ldots, M_m\}$ for finite $m$, and you pretend that one of the models in $\mathcal{M}$ must be the true data-generating mechanism $M_{DG}$, and you pretend that the utility function

$$U(M, M_{DG}) = \begin{cases} 1 & \text{if } M = M_{DG} \\ 0 & \text{otherwise} \end{cases}$$

(13)

reflects your real-world values, then it’s decision-theoretically optimal to choose the model in $\mathcal{M}$ with the highest posterior probability (i.e., that choice maximizes expected utility).

If it’s scientifically appropriate to take the prior model probabilities $p(M_j|B)$ to be equal, this rule reduces to choosing the model with the highest Bayes factor in favor of it; this can be found by (a) computing the Bayes factor in favor of $M_2$ over $M_1$,

$$BF(M_2 \text{ over } M_1|D B) = \frac{p(D|M_2 B)}{p(D|M_1 B)},$$

(14)
favoring $M_2$ if $BF(M_2 \text{ over } M_1 | DB) > 1$, i.e., if $p(D|M_2 B) > p(D|M_1 B)$, and calling the better model $M^*$; (b) computing the Bayes factor in favor of $M^*$ over $M_3$, calling the better model $M^*$; and so on up through $M_m$.

Notice that there’s something else a bit funny about this: $p(D|M_j B)$ is the prior (not posterior) predictive distribution for the data set $D$ under model $M_j$, so the Bayes factor rule tells You to choose the model that does the best job of predicting the data before any data arrives.

Let’s look at the general problem of parametric model comparison, in which model $M_j$ has its own parameter vector $\gamma_j$ (of length $k_j$), where $\gamma_j = (\theta, \eta_j)$, and is specified by

$$M_j: \left\{ \begin{array}{l} (\gamma_j|M_j B) \sim p(\gamma_j|M_j B) \\ (D|\gamma_j M_j B) \sim p(D|\gamma_j M_j B) \end{array} \right\}. \tag{15}$$

Here the quantity $p(D|M_j B)$ that defines the Bayes factor is
\[ p(D|M_j B) = \int p(D|\gamma_j M_j B) p(\gamma_j|M_j B) \, d\gamma_j ; \]  

(16)

this is called an \textit{integrated likelihood} (or \textit{marginal likelihood}) because it tells You to take a \textbf{weighted average} of the \textbf{sampling distribution/likelihood} \( p(D|\gamma_j M_j B) \), but \textbf{NB} weighted by the \textbf{prior} for \( \gamma_j \) in model \( M_j \); as noted above, this may seem \textbf{surprising}, but it's \textbf{correct}, and it can lead to \textbf{trouble}, as follows.

The first trouble is \textbf{technical}: the \textbf{integral} in (16) can be \textbf{difficult to compute}, and may not even be easy to \textbf{approximate}.

The second thing to \textbf{notice} is that (16) can be \textbf{rewritten} as

\[ p(D|M_j B) = E(\gamma_j|M_j B) \, p(D|\gamma_j M_j B) . \]  

(17)

In other words the \textbf{integrated likelihood} is the \textbf{expectation} of the \textbf{sampling distribution} over the \textbf{prior} for \( \gamma_j \) in model \( M_j \) (evaluated at the \textbf{observed data set} \( D \)).

A few \textbf{additional words} about \textbf{prior distributions} on \textbf{parameters}:
A distribution (density) for a real-valued parameter $\theta$ that summarizes the information

$$\{\theta \text{ is highly likely to be near } \theta_0\}$$

will have most of its mass concentrated near $\theta_0$, whereas the information

$$\{\text{not much is known about } \theta\}$$

would correspond to a density that’s rather flat (or diffuse) across a broad range of $\theta$ values; thus when the scientific context offers little information about $\gamma_j$ external to the data set $D$, this translates into a diffuse prior on $\gamma_j$, and this spells trouble for Bayes factors:

$$p(D|M_j \mathcal{B}) = E(\gamma_j|M_j \mathcal{B}) \cdot p(D|\gamma_j M_j \mathcal{B}).$$

You can see that if the available information implies that $p(\gamma_j|M_j \mathcal{B})$ should be diffuse, the expectation defining the integrated likelihood can be highly unstable with respect to small details in how the diffuseness is specified.

**Example:** Integer-valued data set $D = (y_1 \ldots y_n); \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$;
Instability of Bayes Factors (continued)

\[ M_1 = \text{Geometric}(\theta_1) \] likelihood with a \( \text{Beta}(\alpha_1, \beta_1) \) prior on \( \theta_1 \);

\[ M_2 = \text{Poisson}(\theta_2) \] likelihood with a \( \text{Gamma}(\alpha_2, \beta_2) \) prior on \( \theta_2 \).

The Bayes factor in favor of \( M_1 \) over \( M_2 \) turns out to be

\[
\frac{\Gamma(\alpha_1 + \beta_1) \Gamma(n + \alpha_1) \Gamma(n\bar{y} + \beta_1) \Gamma(\alpha_2) (n + \beta_2)^{n\bar{y} + \alpha_2} (\prod_{i=1}^{n} y_i!)}{\Gamma(\alpha_1) \Gamma(\beta_1) \Gamma(n + n\bar{y} + \alpha_1 + \beta_1) \Gamma(n\bar{y} + \alpha_2) \beta_2^{\alpha_2}}.
\] (18)

With standard diffuse priors — take \( (\alpha_1, \beta_1) = (1, 1) \) and \( (\alpha_2, \beta_2) = (\epsilon, \epsilon) \) for some \( \epsilon > 0 \) — the Bayes factor reduces to

\[
\frac{\Gamma(n + 1) \Gamma(n\bar{y} + 1) \Gamma(\epsilon) (n + \epsilon)^{n\bar{y} + \epsilon} (\prod_{i=1}^{n} y_i!)}{\Gamma(n + n\bar{y} + 2) \Gamma(n\bar{y} + \epsilon) \epsilon^{\epsilon}}.
\] (19)

This goes to \(+\infty\) as \( \epsilon \downarrow 0 \), i.e., You can make the evidence in favor of the Geometric model over the Poisson as large as You want, no matter what the data says, as a function of a quantity near 0 that scientifically You have no basis to specify.

If instead You fix and bound \( (\alpha_2, \beta_2) \) away from 0 and let \( (\alpha_1, \beta_1) \downarrow 0 \), You can completely reverse this and make the evidence in favor of the Poisson model over the Geometric as large as You want (for any \( y \)).
Approximating Integrated Likelihoods

The **bottom line** is that, when **scientific context** suggests **diffuse priors** on the **parameter vectors** in the **models** being **compared**, the **integrated likelihood values** that are at the **heart of Bayes factors** can be **hideously sensitive** to **small arbitrary details** in how the **diffuseness** is specified.

This has been **well-known** for quite awhile now, and it's given rise to an **amazing amount of fumbling around**, as people who like **Bayes factors** have tried to find a way to **fix** the problem: at this point the **list of attempts** includes \{**partial, intrinsic, fractional**\} Bayes factors, well-calibrated priors, conventional priors, intrinsic priors, expected posterior priors, ... (e.g., Pericchi 2004), and all of them **exhibit** a level of **ad-hockery** that's **otherwise absent** from the **Bayesian paradigm**.

**Approximating integrated likelihoods.** The goal is

\[
p(D|M_j B) = \int p(D|\gamma_j M_j B) p(\gamma_j|M_j B) \, d\gamma_j ; \quad (20)
\]

maybe there's an **analytic approximation** to this that will suggest how to **avoid trouble**.
Laplace (1785) already faced this problem 225 years ago, and he offered a solution that’s often useful, which people now call a Laplace approximation in his honor (it’s an example of what’s also known in the applied mathematics literature as a saddle-point approximation).

Noticing that the integrand $P^*(\gamma_j) \equiv p(D|\gamma_j M_j B) p(\gamma_j|M_j B)$ in $p(D|M_j B)$ is an un-normalized version of the posterior distribution $p(\gamma_j|D M_j B)$, and appealing to a Bayesian version of the Central Limit Theorem — which says that with a lot of data, such a posterior distribution should be close to Gaussian, centered at the posterior mode $\hat{\gamma}_j$ — You can see that (with a large sample size $n$) $\log P^*(\gamma_j)$ should be close to quadratic around that mode; the Laplace idea is to take a Taylor expansion of $\log P^*(\gamma_j)$ around $\hat{\gamma}_j$ and retain only the terms out to second order; the result is

$$\log p(D|M_j B) = \log p(D|\hat{\gamma}_j M_j B) + \log p(\hat{\gamma}_j|M_j B) + \frac{k_j}{2} \log 2\pi - \frac{1}{2} \log |\hat{I}_j| + O \left( \frac{1}{n} \right) ; \quad (21)$$

here $\hat{\gamma}_j$ is the maximum likelihood estimate of the parameter vector $\gamma_j$ under model $M_j$ and $\hat{I}_j$ is the observed information matrix under $M_j$. 
Notice that the prior on $\gamma_j$ in model $M_j$ enters into this approximation through $\log p(\hat{\gamma}_j | M_j \mathcal{B})$, and this is a term that won’t go away with more data: as $n$ increases this term is $O(1)$.

Using a less precise Taylor expansion, Schwarz (1978) obtained a different approximation that’s the basis of what has come to be known as the Bayesian information criterion (BIC):

$$\log p(y | M_j \mathcal{B}) = \log p(y | \hat{\gamma}_j M_j \mathcal{B}) - \frac{k_j}{2} \log n + O(1).$$

People often work with a multiple of this for model comparison:

$$BIC(M_j | D \mathcal{B}) = -2 \log p(D | \hat{\gamma}_j M_j \mathcal{B}) + k_j \log n$$

(the $-2$ multiplier comes from deviance considerations); multiplying by $-2$ induces a search (with this approach) for models with small BIC.

This model-comparison method makes an explicit trade-off between model complexity (which goes up with $k_j$ at a log $n$ rate) — and model lack of fit (through the $-2 \log p(D | \hat{\gamma}_j M_j \mathcal{B})$ term).
BIC is called an information criterion because it resembles AIC (Akaike, 1974). which was derived using information-theoretic reasoning:

\[
AIC(M_j | D B) = -2 \log p(D|\hat{\gamma}_j M_j B) + 2 k_j .
\]  

(24)

AIC penalizes model complexity at a linear rate in \( k_j \) and so can have different behavior than BIC, especially with moderate to large \( n \) (BIC tends to choose simpler models; more on this later).

It’s possible to work out what implied prior BIC is using, from the point of view of the Laplace approximation; the result is

\[
(\gamma_j | M_j B) \sim N_{k_j}(\hat{\gamma}_j, n\hat{I}_j^{-1}).
\]

(25)

In the literature this is called a unit-information prior, because in large samples it corresponds to the prior being equivalent to 1 new observation yielding the same sufficient statistics as the observed data.

This prior is data-determined, but this effect is close to negligible even with only moderate \( n \).
The BIC approximation to Bayes factors has the extremely desirable property that it’s free of the hideous instability of integrated likelihoods with respect to tiny details, in how diffuse priors are specified, that do not arise directly from the science of the problem; in my view, if You’re going to use Bayes factors to choose among models, You’re well advised to use a method like BIC that protects You from Yourself in mis-specifying those tiny details.

I said back on page 49 that there are two generic utility-based model-comparison methods: Bayes factors and log scores.

- Log scores are based on what might be termed the **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.

This suggests developing a generic utility structure based on predictive accuracy: consider first a setting in which $D = y = (y_1 \ldots y_n)$ for real-valued $y_i$ and the models to be compared are (as before)
Log Scores

\[ M_j: \left\{ \begin{array}{l}
\gamma_j | M_j B \sim p(\gamma_j | M_j B) \\
(y | \gamma_j M_j B) \sim p(y | \gamma_j M_j B)
\end{array} \right\}. \]

When comparing a (future) data value \( y^* \) with the predictive distribution \( p(\cdot | y M_j B) \) for it under \( M_j \), it’s been shown that (under reasonable optimality criteria) all optimal scores measuring the discrepancy between \( y^* \) and \( p(\cdot | y M_j B) \) are linear functions of \( \log p(y^* | y M_j B) \) (the log of the height of the predictive distribution at the observed value \( y^* \)).

Using this fact, perhaps the most natural-looking form for a composite measure of predictive accuracy of \( M_j \) is a cross-validated version of the resulting log score,

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

in which \( y_{-i} \) is the \( y \) vector with observation \( i \) omitted.

Somewhat surprisingly, Draper and Krnjajić (2010) have shown that a full-sample log score that omits the leave-one-out idea,
Full-Sample Log Score

\[ LS_{FS}(M_j|y B) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|y M_j B), \] (28)

made operational with the rule \{favor \( M_2 \) over \( M_1 \) if \( LS_{FS}(M_2|y B) > LS_{FS}(M_1|y B) \)}, can have better small-sample model discrimination ability than \( LS_{CV} \) (in addition to being faster to approximate in a stable way).

If, in the spirit of calibration, You’re prepared to think about an underlying data-generating model \( M_{DG} \), \( LS_{FS} \) also has a nice interpretation as an approximation to the Kullback-Leibler divergence between \( M_{DG} \) and \( p(\cdot|y M_j B) \), in which \( M_{DG} \) is approximated by the empirical CDF:

\[
KL[M_{DG}||p(\cdot|y M_j B)] = E_{M_{DG}} \log M_{DG} - E_{M_{DG}} \log p(\cdot|y M_j B) \\
= E_{M_{DG}} \log M_{DG} - LS_{FS}(M_j|y B); \quad (29)
\]

the first term on the right side of (29) is constant in \( p(\cdot|y M_j B) \), so minimizing \( KL[M_{DG}||p(\cdot|y M_j B)] \) is approximately the same as maximizing \( LS_{FS} \).
What follows is a sketch of recent results (Draper, 2011) based on simulation experiments with realistic sample sizes; in my view standard asymptotic calculations — choosing between the models in $\mathcal{M} = \{M_1, M_2\}$ as $n \to \infty$ with $\mathcal{M}$ remaining fixed — are essentially irrelevant in calibration studies, for two reasons:

(1) With increasing $n$, you’ll want $\mathcal{M}$ to grow to satisfy your desire to do a better job of capturing real-world complexities, and

(2) Data usually accumulate over time, and with increasing $n$ it becomes more likely that the real-world process you’re modeling is not stationary.

- Versions of Bayes factors that behave sensibly with diffuse priors on the model parameters tend to have model discrimination performance similar to that of BIC in calibration (repeated-sampling with known $M_{DG}$) environments.

Let’s look first at simple settings in which $M_1$ and $M_2$ have (or appear to have) equal complexity (the same number of parameters).
Clinical Trial to Quantify Improvement

Example: Consider assessing the performance of a drug, for lowering systolic blood pressure (SBP) in hypertensive patients, in a phase–II clinical trial, and suppose that a Gaussian sampling distribution for the outcome variable is reasonable (possibly after transformation).

Two frequent designs in settings of this type have as their goals quantifying improvement and establishing bio-equivalence.

• (quantifying improvement) Here You want to estimate the mean decline in blood pressure under this drug, and it would be natural to choose a repeated-measures (pre-post) experiment, in which SBP values are obtained for each patient, both before and after taking the drug for a sufficiently long period of time for its effect to become apparent.

Let $\theta$ stand for the mean difference $(SBP_{\text{before}} - SBP_{\text{after}})$ in the population of patients to which it’s appropriate to generalize from the patients in Your trial, and let $D = y = (y_1 \ldots y_n)$, where $y_i$ is the observed difference $(SBP_{\text{before}} - SBP_{\text{after}})$ for patient $i$ ($i = 1, \ldots, n$).
The real-world purpose of this experiment is to decide whether to take the drug forward to phase III; under the weight of 20th-century inertia (in which decision-making was strongly — and incorrectly — subordinated to inference), Your first impulse might be to treat this as an inferential problem about \( \theta \), but it’s not; it’s a decision problem that involves \( \theta \).

The action space here is \( \mathcal{A} = (a_1, a_2) = (\text{don’t take the drug forward to phase III, do take it forward}) \), and a sensible utility function \( U(a_j, \theta) \) should be continuous and monotonically increasing in \( \theta \) over a broad range of positive \( \theta \) values (the bigger the SBP decline for hypertensive patients who start at (say) 160 mmHg, the better, up to a drop of about 40 mmHg, beyond which the drug starts inducing fainting spells).

However, to facilitate a comparison between BIC and log scores with frequentist hypothesis-testing, here I’ll compare two models \( M_1 \) and \( M_2 \) that dichotomize the \( \theta \) range, but not at 0: despite a century of textbook claims to the contrary, there’s nothing special about \( \theta = 0 \) in this setting, and in fact You know scientifically that \( \theta \) is
not exactly 0 (because the outcome variable in this experiment is conceptually continuous).

What matters here is whether $\theta > \Delta$, where $\Delta$ is a practical significance improvement threshold below which the drug is not worth advancing into phase III (for example, any drug that did not lower SBP for severely hypertensive patients — those whose pre-drug values average 160 mmHg or more — by at least 15 mmHg would not deserve further attention).

With little information about $\theta$ external to this experimental data set, what counts in this situation is the comparison of the following two models:

$$M_1: \left\{ \begin{array}{c} (\theta|B) \sim \text{diffuse for } \theta \leq \Delta \\ (y_i|\theta B) \sim \text{IID } N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \quad \left(30\right)$$

$$M_2: \left\{ \begin{array}{c} (\theta|B) \sim \text{diffuse for } \theta > \Delta \\ (y_i|\theta B) \sim \text{IID } N(\theta, \sigma^2) \end{array} \right\}, \quad \left(31\right)$$

in which for simplicity I’ll take $\sigma^2$ to be known (the results are similar with $\sigma^2$ learned from the data).
The analogue in the frequentist story is of course to assume the sampling model \( y_i \overset{\text{iid}}{\sim} N(\theta, \sigma^2) \) and test \( H_0: \theta \leq \Delta \) against \( H_A: \theta > \Delta \) at level \( \alpha \); since \( \sigma^2 \) is known, there’s a uniformly-most-powerful (UMP) level-\( \alpha \) test of the form \{favor \( H_A \) (choose \( M_2 \)) if \( \bar{y} > \Delta + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha) \}\}, where \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \).

This gives rise to four model-selection methods that can be compared calibratively:

- **Full-sample log scores**: choose \( M_2 \) if \( LS_{FS}(M_2 | y B) > LS_{FS}(M_1 | y B) \).

- **BIC**: choose \( M_2 \) if \( BIC(M_2 | y B) < BIC(M_1 | y B) \).

- **Posterior probability**: let \( M^* = \{(\theta | B) \sim \text{diffuse on } \mathbb{R}, (y_i | \theta B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)\} \) and choose \( M_2 \) if \( p(\theta > \Delta | y M^* B) > 0.5 \).

- **Hypothesis-testing** with \( \alpha \) set at a traditional level (e.g., 0.05).

**Simulation experiment details**, based on the **SBP drug trial**: \( \Delta = 15 \); \( \sigma = 10 \); \( n = 10, 20, \ldots, 100 \); data-generating \( \theta_{DG} = 11, 12, \ldots, 19 \); \( \alpha = 0.05 \); 1,000 simulation replications.
A note about log-score computation:

\[ \text{LS}_{FS}(M_j | y \ B) = - \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | y M_j B). \]

When the predictive ordinate \( p(y^* | y M_j B) \) has no closed-form expression, there’s an easy Monte-Carlo approach to approximating it: when parametric model \( M_j \) with parameter vector \( \gamma_j \) yields a posterior distribution \( p(\gamma_j | y M_j B) \) that’s straightforward to sample from, let \( (\gamma_j)^*_m \) \((m = 1, \ldots, M)\) be identically-distributed draws from that posterior; then

\[
\begin{align*}
  p(y^* | y M_j B) &= \int p(y^* | \gamma_j M_j B) p(\gamma_j | y M_j B) d\gamma_i \\
  &= E_{(\gamma_j | y M_j B)} [p(y^* | \gamma_j M_j B)] \tag{32} \\
  &= \frac{1}{M} \sum_{m=1}^{M} p(y^* | (\gamma_j)^*_m M_j B), \quad \text{and} \\
  \text{LS}_{FS}(M_j | y B) &= - \frac{1}{n} \sum_{i=1}^{n} \log \left[ \frac{1}{M} \sum_{m=1}^{M} p(y_i | (\gamma_j)^*_m M_j B) \right].
\end{align*}
\]
I used $M = 10,000$ draws in the $LS_{FS}$ approximations; the tables below give Monte-Carlo estimates of the probability that $M_2$ is chosen.

<table>
<thead>
<tr>
<th>$n$</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.095</td>
<td>0.163</td>
<td>0.240</td>
<td>0.376</td>
<td>0.476</td>
<td>0.646</td>
<td>0.725</td>
<td>0.822</td>
<td>0.892</td>
</tr>
<tr>
<td>20</td>
<td>0.035</td>
<td>0.105</td>
<td>0.177</td>
<td>0.316</td>
<td>0.505</td>
<td>0.663</td>
<td>0.824</td>
<td>0.913</td>
<td>0.970</td>
</tr>
<tr>
<td>30</td>
<td>0.011</td>
<td>0.061</td>
<td>0.154</td>
<td>0.294</td>
<td>0.505</td>
<td>0.719</td>
<td>0.857</td>
<td>0.944</td>
<td>0.987</td>
</tr>
<tr>
<td>40</td>
<td>0.006</td>
<td>0.029</td>
<td>0.115</td>
<td>0.266</td>
<td>0.527</td>
<td>0.744</td>
<td>0.903</td>
<td>0.966</td>
<td>0.996</td>
</tr>
<tr>
<td>50</td>
<td>0.003</td>
<td>0.022</td>
<td>0.069</td>
<td>0.219</td>
<td>0.516</td>
<td>0.731</td>
<td>0.916</td>
<td>0.982</td>
<td>1.000</td>
</tr>
<tr>
<td>60</td>
<td>0.003</td>
<td>0.008</td>
<td>0.054</td>
<td>0.216</td>
<td>0.503</td>
<td>0.772</td>
<td>0.939</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>70</td>
<td>0.000</td>
<td>0.006</td>
<td>0.051</td>
<td>0.213</td>
<td>0.493</td>
<td>0.776</td>
<td>0.955</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>80</td>
<td>0.000</td>
<td>0.004</td>
<td>0.037</td>
<td>0.189</td>
<td>0.541</td>
<td>0.807</td>
<td>0.961</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td>90</td>
<td>0.000</td>
<td>0.002</td>
<td>0.027</td>
<td>0.169</td>
<td>0.497</td>
<td>0.834</td>
<td>0.975</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>0.000</td>
<td>0.000</td>
<td>0.024</td>
<td>0.170</td>
<td>0.535</td>
<td>0.854</td>
<td>0.979</td>
<td>0.999</td>
<td>1.000</td>
</tr>
</tbody>
</table>

This exhibits all the monotonicities that it should, and correctly yields $0.5$ for all $n$ with $\theta_{DG} = 15$. 
<table>
<thead>
<tr>
<th>n</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.093</td>
<td>0.162</td>
<td>0.238</td>
<td>0.376</td>
<td>0.476</td>
<td>0.646</td>
<td>0.727</td>
<td>0.821</td>
<td>0.892</td>
</tr>
<tr>
<td>20</td>
<td>0.035</td>
<td>0.104</td>
<td>0.174</td>
<td>0.315</td>
<td>0.508</td>
<td>0.661</td>
<td>0.824</td>
<td>0.913</td>
<td>0.970</td>
</tr>
<tr>
<td>30</td>
<td>0.012</td>
<td>0.061</td>
<td>0.154</td>
<td>0.293</td>
<td>0.506</td>
<td>0.720</td>
<td>0.858</td>
<td>0.944</td>
<td>0.988</td>
</tr>
<tr>
<td>40</td>
<td>0.006</td>
<td>0.029</td>
<td>0.114</td>
<td>0.268</td>
<td>0.526</td>
<td>0.742</td>
<td>0.902</td>
<td>0.966</td>
<td>0.996</td>
</tr>
<tr>
<td>50</td>
<td>0.003</td>
<td>0.022</td>
<td>0.070</td>
<td>0.219</td>
<td>0.515</td>
<td>0.731</td>
<td>0.915</td>
<td>0.981</td>
<td>1.000</td>
</tr>
<tr>
<td>60</td>
<td>0.003</td>
<td>0.008</td>
<td>0.054</td>
<td>0.217</td>
<td>0.503</td>
<td>0.772</td>
<td>0.939</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>70</td>
<td>0.000</td>
<td>0.006</td>
<td>0.050</td>
<td>0.214</td>
<td>0.492</td>
<td>0.776</td>
<td>0.956</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>80</td>
<td>0.000</td>
<td>0.004</td>
<td>0.037</td>
<td>0.189</td>
<td>0.541</td>
<td>0.806</td>
<td>0.959</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td>90</td>
<td>0.000</td>
<td>0.002</td>
<td>0.026</td>
<td>0.169</td>
<td>0.497</td>
<td>0.835</td>
<td>0.975</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>0.000</td>
<td>0.000</td>
<td>0.025</td>
<td>0.169</td>
<td>0.534</td>
<td>0.853</td>
<td>0.979</td>
<td>0.999</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Even though the $LS_{FS}$ and posterior-probability methods are quite different, their information-processing in discriminating between $M_1$ and $M_2$ is identical.
<table>
<thead>
<tr>
<th>n</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.093</td>
<td>0.162</td>
<td>0.238</td>
<td>0.376</td>
<td>0.476</td>
<td>0.646</td>
<td>0.727</td>
<td>0.821</td>
<td>0.892</td>
</tr>
<tr>
<td>20</td>
<td>0.035</td>
<td>0.104</td>
<td>0.174</td>
<td>0.315</td>
<td>0.508</td>
<td>0.661</td>
<td>0.824</td>
<td>0.913</td>
<td>0.970</td>
</tr>
<tr>
<td>30</td>
<td>0.012</td>
<td>0.061</td>
<td>0.154</td>
<td>0.293</td>
<td>0.506</td>
<td>0.720</td>
<td>0.858</td>
<td>0.944</td>
<td>0.988</td>
</tr>
<tr>
<td>40</td>
<td>0.006</td>
<td>0.029</td>
<td>0.114</td>
<td>0.268</td>
<td>0.526</td>
<td>0.742</td>
<td>0.902</td>
<td>0.966</td>
<td>0.996</td>
</tr>
<tr>
<td>50</td>
<td>0.003</td>
<td>0.022</td>
<td>0.070</td>
<td>0.219</td>
<td>0.515</td>
<td>0.731</td>
<td>0.915</td>
<td>0.981</td>
<td>1.000</td>
</tr>
<tr>
<td>60</td>
<td>0.003</td>
<td>0.008</td>
<td>0.054</td>
<td>0.217</td>
<td>0.503</td>
<td>0.772</td>
<td>0.939</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>70</td>
<td>0.000</td>
<td>0.006</td>
<td>0.050</td>
<td>0.214</td>
<td>0.492</td>
<td>0.776</td>
<td>0.956</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>80</td>
<td>0.000</td>
<td>0.004</td>
<td>0.037</td>
<td>0.189</td>
<td>0.541</td>
<td>0.806</td>
<td>0.959</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td>90</td>
<td>0.000</td>
<td>0.002</td>
<td>0.026</td>
<td>0.169</td>
<td>0.497</td>
<td>0.835</td>
<td>0.975</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>0.000</td>
<td>0.000</td>
<td>0.025</td>
<td>0.169</td>
<td>0.534</td>
<td>0.853</td>
<td>0.979</td>
<td>0.999</td>
<td>1.000</td>
</tr>
</tbody>
</table>

In this problem **BIC** and the **posterior-probability approach** are **algebraically identical**, making the **model-discrimination performance** of **BIC** and \(L_{FS}\) **the same**.
### Model-Comparison Results (continued)

<table>
<thead>
<tr>
<th>theta.DG</th>
<th>-------- M.1 correct --------</th>
<th>----- M.2 correct -----</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>20</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>30</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>40</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>50</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>60</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>70</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>80</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>90</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>100</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The **hypothesis-testing approach** with $\alpha = 0.05$ yields quite different **results**, because it’s **constrained to produce 0.05** in the $\theta_{DG} = 15$ **column** for all $n$; as a **result**, it (naturally) does somewhat better on the **left side** of the table, and **substantially worse** on the **right**, than the **Bayesian methods**.
As is well known, with fixed \( n \) a choice such as \( \alpha = 0.05 \) makes the hypothesis test terrified (in the language of this problem) of choosing \( M_2 \) when \( \theta \leq \Delta \) (call this a false positive); the hypothesis test has no built-in counteracting fear of choosing \( M_1 \) when \( \theta > \Delta \) (a false negative).

This makes 0.05–level hypothesis testing not directly comparable with BIC, \( LS_{FS} \) and posterior probabilities, which have a different implicit position on trading off false positives and negatives.

One way to make them directly comparable is to match their false-positive behavior, which can be accomplished by setting \( \alpha = 0.5 \) in the hypothesis-testing approach.

The results (on the next page) are identical to those from the Bayesian approaches.

Thus there’s both a unity between Bayesian and frequentist inferential approaches to this decision problem and a difference: they’re working with the same data information but different implicit choices on how to balance false positives and negatives.
Hypothesis-Testing (alpha = 0.5)

<table>
<thead>
<tr>
<th>theta.DG</th>
<th>n</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>n 10</td>
<td>0.093</td>
<td>0.162</td>
<td>0.238</td>
<td>0.376</td>
<td>0.476</td>
<td>0.646</td>
<td>0.727</td>
<td>0.821</td>
<td>0.892</td>
<td></td>
</tr>
<tr>
<td>n 20</td>
<td>0.035</td>
<td>0.104</td>
<td>0.174</td>
<td>0.315</td>
<td>0.508</td>
<td>0.661</td>
<td>0.824</td>
<td>0.913</td>
<td>0.970</td>
<td></td>
</tr>
<tr>
<td>n 30</td>
<td>0.012</td>
<td>0.061</td>
<td>0.154</td>
<td>0.293</td>
<td>0.506</td>
<td>0.720</td>
<td>0.858</td>
<td>0.944</td>
<td>0.988</td>
<td></td>
</tr>
<tr>
<td>n 40</td>
<td>0.006</td>
<td>0.029</td>
<td>0.114</td>
<td>0.268</td>
<td>0.526</td>
<td>0.742</td>
<td>0.902</td>
<td>0.966</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td>n 50</td>
<td>0.003</td>
<td>0.022</td>
<td>0.070</td>
<td>0.219</td>
<td>0.515</td>
<td>0.731</td>
<td>0.915</td>
<td>0.981</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>n 60</td>
<td>0.003</td>
<td>0.008</td>
<td>0.054</td>
<td>0.217</td>
<td>0.503</td>
<td>0.772</td>
<td>0.939</td>
<td>0.993</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>n 70</td>
<td>0.000</td>
<td>0.006</td>
<td>0.050</td>
<td>0.214</td>
<td>0.492</td>
<td>0.776</td>
<td>0.956</td>
<td>0.993</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>n 80</td>
<td>0.000</td>
<td>0.004</td>
<td>0.037</td>
<td>0.189</td>
<td>0.541</td>
<td>0.806</td>
<td>0.959</td>
<td>0.996</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>n 90</td>
<td>0.000</td>
<td>0.002</td>
<td>0.026</td>
<td>0.169</td>
<td>0.497</td>
<td>0.835</td>
<td>0.975</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>n 100</td>
<td>0.000</td>
<td>0.000</td>
<td>0.025</td>
<td>0.169</td>
<td>0.534</td>
<td>0.853</td>
<td>0.979</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

However, to **elaborate** something said **earlier**: bearing in mind the **real-world purpose** of this **experiment** — to **decide whether or not to take the drug forward to phase III** — neither the **knee-jerk use** of $\alpha = 0.05$ in the **(inferential) hypothesis-testing approach** nor the
off-the-shelf (inferential) use of BIC, log scores or posterior probabilities is optimal here for decision-making.

Even in this setting in which $\Theta$ has been artificially partitioned into $(-\infty, \Delta]$ and $(\Delta, \infty)$, to solve the problem properly, you would have to quantify the real-world consequences of each of the cells in this table specifying $U(a, \theta)$ (here $u_{ij} \geq 0$):

<table>
<thead>
<tr>
<th>Action</th>
<th>Truth</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta \leq \Delta$</td>
<td>$\theta &gt; \Delta$</td>
</tr>
<tr>
<td>$a_1$ (stop)</td>
<td>$u_{11}$</td>
<td>$-u_{12}$</td>
</tr>
<tr>
<td>$a_2$ (phase III)</td>
<td>$-u_{21}$</td>
<td>$u_{22}$</td>
</tr>
</tbody>
</table>

- $u_{11}$ is the gain from correctly not taking the drug forward to phase III (this is clearly 0);
- $u_{12}$ is the loss from incorrectly failing to take the drug forward to phase III;
- $u_{21}$ is the loss from incorrectly taking the drug forward to phase III;
- $u_{22}$ is the gain from correctly taking the drug forward to phase III.
The **optimal Bayesian decision** turns out to be:

choose \( a_2 \) (go forward to phase III) iff

\[
P(\theta > \Delta | y \ B) \geq \frac{u_{21}}{u_{12} + u_{21} + u_{22}} = u^* .
\]

(33)

The frequentist (hypothesis-testing) inferential approach is equivalent to this only if

\[
\alpha = 1 - u^* = \frac{u_{12} + u_{22}}{u_{12} + u_{21} + u_{22}} .
\]

(34)

The implicit trade-off between false positives and negatives in BIC and \( LS_{FS} \) — and the built-in trade-off in level–\( \alpha \) hypothesis-testing for any given \( \alpha \) — may be close to optimal or not, according to the real-world values of \( \{u_{12}, u_{21}, u_{22}\} \).

In phase-II clinical trials or micro-array experiments, when You’re screening many drugs or genes for those that may lead to an effective treatment and — from the drug company’s point of view — a false-negative error (of failing to move forward with a drug or gene that’s actually worth further investigation) can be much more costly than a false-positive mistake, this corresponds to \( u_{12} \gg u_{21} \).
Establishing Bio-Equivalence

and leads in the hypothesis-testing approach in phase-II trials to a willingness to use (much) larger $\alpha$ values than the conventional 0.01 or 0.05, something that good frequentist biostatisticians have long known intuitively.

(In work I’ve done with a Swiss pharmaceutical company, this approach led to $\alpha$ values on the order of 0.45, which is close to the implicit trade-off in BIC and $LS_{FS}$.)

• (establishing bio-equivalence) In this case there’s a previous hypertension drug B (call the new drug A) and You’re wondering if the mean effects of the two drugs are close enough to regard them as bio-equivalent.

A good design here would again have a repeated-measures character, in which each patient’s SBP is measured four times: before and after taking drug A, and before and after taking drug B (allowing enough time to elapse between taking the two drugs for the effects of the first drug to disappear).

Let $\theta$ stand for the mean difference
\[
((SBP_{\text{before},A} - SBP_{\text{after},A}) - (SBP_{\text{before},B} - SBP_{\text{after},B}))
\]

in the population of patients to which it’s appropriate to generalize from the patients in Your trial, and let \( y_i \) be the corresponding difference for patient \( i \) \((i = 1, \ldots, n)\).

Again in this setting there’s nothing special about \( \theta = 0 \), and as before You know scientifically that \( \theta \) is not exactly 0; what matters here is whether \(|\theta| \leq \epsilon\), where \( \epsilon > 0 \) is a practical significance bio-equivalence threshold (e.g., 5 mmHg).

Assuming as before a Gaussian sampling story and little information about \( \theta \) external to this experimental data set, what counts here is a comparison of

\[
M_3: \left\{ \begin{array}{c}
(\theta | \mathcal{B}) \sim \text{diffuse for } |\theta| \leq \epsilon \\
(y_i | \theta \mathcal{B}) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}
\]

and

\[
M_4: \left\{ \begin{array}{c}
(\theta | \mathcal{B}) \sim \text{diffuse for } |\theta| > \epsilon \\
(y_i | \theta \mathcal{B}) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{array} \right\},
\]

in which \( \sigma^2 \) is again taken for simplicity to be known.
A natural alternative to BIC and $LS_{FS}$ here is again based on posterior probabilities: as before, let

$$M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \mathbb{R}, (y_i|\theta, \mathcal{B}) \sim N(\theta, \sigma^2)\},$$

but this time favor $M_4$ over $M_3$ if $p(|\theta| > \epsilon|y M^* \mathcal{B}) > 0.5$.

The analogue in the frequentist story is again to assume the sampling model $y_i \sim N(\theta, \sigma^2)$ and this time test $H_0: |\theta| \leq \epsilon$ against $H_A: |\theta| > \epsilon$ at level $\alpha$; the UMP unbiased test takes the form

$$\{\text{favor } H_A \text{ (choose } M_4 \text{)} \text{ if } |\bar{y}| > \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \frac{\alpha}{2})\}.$$

As before, a careful real-world choice between $M_3$ and $M_4$ in this case would be based on a utility function that quantified the costs and benefits of

$$\{\text{claiming the two drugs were bio-equivalent when they were},$$

$$\text{concluding that they were bio-equivalent when they were not},$$

$$\text{deciding that they were not bio-equivalent when they were},$$

$$\text{judging that they were not bio-equivalent when they were not}\},$$

but here I’ll again simply compare the calibrative performance of $LS_{FS}$, posterior probabilities, BIC and the level-0.05 hypothesis test.
Bio-Equivalence Results

Simulation experiment details, based on the SBP drug trial: $\epsilon = 5$; $\sigma = 10$; $n = 10, 20, \ldots, 100$; data-generating
$
\theta_{DG} = \{-9, -7, -5, -3, -1, 0, 1, 3, 5, 7, 9\}$; $\alpha = 0.05$; 1,000 simulation replications, $M = 10,000$ Monte-Carlo draws for $LS_{FS}$.

My full-grid simulation ran out of time; here’s a summary of the qualitative findings:

- $LS_{FS}$ and the posterior-probability approach have identical model-discrimination performance.

- When $M_{DG} = M_3$ (i.e., when the drugs are bio-equivalent), BIC chooses $M_3$ (gets the right answer) more often than $LS_{FS}$; when $M_{DG} = M_4$ (i.e., when the drugs are not bio-equivalent), $LS_{FS}$ chooses $M_4$ (gets the right answer) more often than BIC.

In frequentist language, in this setting, You would conclude that (a) $LS_{FS}$ is more powerful than BIC at correctly establishing non-bio-equivalence but (b) $LS_{FS}$ achieves this extra power by committing more type-I errors (saying the drugs are not bio-equivalent when they are) than BIC.
Whether this is desirable or undesirable behavior on the part of $LS_{FS}$ depends on the real-world consequences of false-positive and false-negative errors.

- When $\alpha$ is matched to the $LS_{FS}$ behavior at $\theta_{DG} = 0$ (this involves letting $\alpha$ increase as $n$ increases), $LS_{FS}$ and the hypothesis-testing approach have identical model-discrimination performance.

- When $\alpha$ is matched to the BIC behavior at $\theta_{DG} = 0$ (this involves letting $\alpha$ decrease as $n$ increases), BIC and the hypothesis-testing approach have identical model-discrimination performance.

An extreme example of the false-positive/false-negative differences between $LS_{FS}$ and BIC in this setting may be obtained, albeit unwisely, by letting $\epsilon \downarrow 0$.

This is unwise here (and is often unwise) because it amounts, in frequentist language, to testing the sharp-null hypothesis $H_0: \theta = 0$ against the alternative $H_A: \theta \neq 0$.

Sharp-null testing is frequently unwise because
(a) You already know from scientific context, when the outcome variable is continuous, that $H_0$ is false, and (relatedly)

(b) it’s silly from a measurement point of view: with a (conditionally) IID $N(\theta, \sigma^2)$ sample of size $n$, your measuring instrument $\bar{y}$ is only accurate to resolution $\frac{\sigma}{\sqrt{n}} > 0$; claiming to be able to discriminate between $\theta = 0$ and $\theta \neq 0$ is like someone with a scale that’s only accurate to the nearest ounce telling You that Your wedding ring has 1 gram (0.035 ounce) less gold in it than the jeweler claims it does.

Nevertheless, for people who like to test sharp-null hypotheses, here are some results: here I’m comparing the models $(i = 1, \ldots, n)$

\[
M_5: \left\{ \begin{array}{c}
(\sigma^2|B) \sim \text{diffuse on } (0, \text{large}) \\
(y_i|\sigma^2 B) \overset{\text{iid}}{\sim} N(0, \sigma^2)
\end{array} \right\} \quad \text{and} \quad (38)
\]

\[
M_6: \left\{ \begin{array}{c}
(\theta \sigma^2|B) \sim \text{diffuse on } (-\text{large}, \text{large}) \times (0, \text{large}) \\
(y_i|\theta \sigma^2 B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}, \quad (39)
\]
In this case a natural Bayesian competitor to BIC and $LS_{FS}$ would be to construct the central $100(1 - \alpha)\%$ posterior interval for $\theta$ under $M_6$ and choose $M_6$ if this interval doesn't contain 0; with the diffuse priors used here this is equivalent to the usual $100(1 - \alpha)\%$ confidence interval in the frequentist model $y_i \overset{iid}{\sim} N(\theta, \sigma^2)$.

Simulation experiment details: data-generating $\sigma_{DG} = 1$; $n = 10, 20, \ldots, 100$; data-generating $\theta_{DG} = \{0, 0.1, \ldots, 0.5\}$; 1,000 simulation replications, $M = 100,000$ Monte-Carlo draws for $LS_{FS}$; the tables below give Monte-Carlo estimates of the probability that $M_6$ is chosen.

The results on the next page are for $LS_{FS}$; in the limit as $\epsilon \downarrow 0$ this approach makes hardly any false-negative errors but quite a lot of false-positive mistakes.

For $\epsilon = 0$ the calibration probability that $LS_{FS}$ chooses $M_6$ goes to 1 as $n \to \infty$, meaning that it has the “wrong asymptotic behavior” when $\theta_{DG} = 0$, but this is misleading: for any $\epsilon > 0$, $LS_{FS}$ has the right asymptotic behavior both when $M_{DG} = M_5$ and when $M_{DG} = M_6$. 
LS.FS

theta.DG

M.5

correct -------------- M.6 correct ------------

<table>
<thead>
<tr>
<th>n</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.72895</td>
<td>0.74097</td>
<td>0.77593</td>
<td>0.82336</td>
<td>0.87112</td>
<td>0.91514</td>
</tr>
<tr>
<td>20</td>
<td>0.77180</td>
<td>0.79289</td>
<td>0.84355</td>
<td>0.90506</td>
<td>0.95049</td>
<td>0.97902</td>
</tr>
<tr>
<td>30</td>
<td>0.80310</td>
<td>0.82802</td>
<td>0.88982</td>
<td>0.94741</td>
<td>0.98073</td>
<td>0.99470</td>
</tr>
<tr>
<td>40</td>
<td>0.82232</td>
<td>0.85493</td>
<td>0.91903</td>
<td>0.97000</td>
<td>0.99233</td>
<td>0.99867</td>
</tr>
<tr>
<td>50</td>
<td>0.83769</td>
<td>0.87417</td>
<td>0.93899</td>
<td>0.98315</td>
<td>0.99676</td>
<td>0.99967</td>
</tr>
<tr>
<td>60</td>
<td>0.84998</td>
<td>0.88789</td>
<td>0.95317</td>
<td>0.98916</td>
<td>0.99900</td>
<td>0.99990</td>
</tr>
<tr>
<td>70</td>
<td>0.85900</td>
<td>0.90081</td>
<td>0.96485</td>
<td>0.99410</td>
<td>0.99964</td>
<td>0.99999</td>
</tr>
<tr>
<td>80</td>
<td>0.86818</td>
<td>0.91038</td>
<td>0.97262</td>
<td>0.99622</td>
<td>0.99976</td>
<td>1.00000</td>
</tr>
<tr>
<td>90</td>
<td>0.87222</td>
<td>0.91984</td>
<td>0.97899</td>
<td>0.99772</td>
<td>0.99988</td>
<td>1.00000</td>
</tr>
<tr>
<td>100</td>
<td>0.88008</td>
<td>0.92647</td>
<td>0.98358</td>
<td>0.99861</td>
<td>0.99995</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

As the **results** on the **next page** indicate, the **behavior** of the **interval approach** is **quite different**: it makes **many false-negative errors** because its **rate of false-positive mistakes** is fixed at **0.05**.
Testing Sharp-Null Hypotheses (continued)

Interval (alpha = 0.05)

<table>
<thead>
<tr>
<th>theta.DG</th>
<th>M.5 correct</th>
<th>M.6 correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>10</td>
<td>0.05014</td>
<td>0.05913</td>
</tr>
<tr>
<td>20</td>
<td>0.05060</td>
<td>0.07029</td>
</tr>
<tr>
<td>30</td>
<td>0.04965</td>
<td>0.08173</td>
</tr>
<tr>
<td>40</td>
<td>0.04968</td>
<td>0.09494</td>
</tr>
<tr>
<td>50</td>
<td>0.04991</td>
<td>0.10744</td>
</tr>
<tr>
<td>60</td>
<td>0.05039</td>
<td>0.11929</td>
</tr>
<tr>
<td>70</td>
<td>0.05051</td>
<td>0.13161</td>
</tr>
<tr>
<td>80</td>
<td>0.04967</td>
<td>0.14397</td>
</tr>
<tr>
<td>90</td>
<td>0.04990</td>
<td>0.15472</td>
</tr>
<tr>
<td>100</td>
<td>0.05026</td>
<td>0.16734</td>
</tr>
</tbody>
</table>

The next page shows that when the interval method is modified so that $\alpha$ matches the $LS_{FS}$ behavior at $\theta_{DG} = 0$ (letting $\alpha$ vary with $n$), the two approaches have identical model-discrimination ability.
Testing Sharp-Null Hypotheses (continued)

Interval (alpha matched to LS.FS)

\[
\theta_{DG}
\]

M.5

<table>
<thead>
<tr>
<th>n</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.72876</td>
<td>0.74143</td>
<td>0.77665</td>
<td>0.82496</td>
<td>0.87446</td>
<td>0.91642</td>
</tr>
<tr>
<td>20</td>
<td>0.77194</td>
<td>0.79285</td>
<td>0.84486</td>
<td>0.90459</td>
<td>0.95189</td>
<td>0.97948</td>
</tr>
<tr>
<td>30</td>
<td>0.80471</td>
<td>0.82881</td>
<td>0.88986</td>
<td>0.94818</td>
<td>0.98097</td>
<td>0.99489</td>
</tr>
<tr>
<td>40</td>
<td>0.82248</td>
<td>0.85597</td>
<td>0.91904</td>
<td>0.97017</td>
<td>0.99231</td>
<td>0.99872</td>
</tr>
<tr>
<td>50</td>
<td>0.83850</td>
<td>0.87419</td>
<td>0.93918</td>
<td>0.98336</td>
<td>0.99694</td>
<td>0.99967</td>
</tr>
<tr>
<td>60</td>
<td>0.85081</td>
<td>0.8891</td>
<td>0.95347</td>
<td>0.98927</td>
<td>0.99904</td>
<td>0.99990</td>
</tr>
<tr>
<td>70</td>
<td>0.85913</td>
<td>0.90058</td>
<td>0.96483</td>
<td>0.99425</td>
<td>0.99962</td>
<td>0.99998</td>
</tr>
<tr>
<td>80</td>
<td>0.86903</td>
<td>0.91145</td>
<td>0.97292</td>
<td>0.99642</td>
<td>0.99977</td>
<td>1.00000</td>
</tr>
<tr>
<td>90</td>
<td>0.87058</td>
<td>0.91892</td>
<td>0.97874</td>
<td>0.99774</td>
<td>0.99988</td>
<td>1.00000</td>
</tr>
<tr>
<td>100</td>
<td>0.88008</td>
<td>0.92647</td>
<td>0.98358</td>
<td>0.99861</td>
<td>0.99995</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

As the results on the next page indicate, BIC’s behavior is quite different from that of \(LS_{FS}\) and fixed-\(\alpha\) intervals: its false-positive rate goes to 0 as \(n\) grows, but it suffers a high false-negative rate to achieve this goal.
The next page shows that when the interval method is modified so that $\alpha$ matches the BIC behavior at $\theta_{DG} = 0$ (letting $\alpha$ vary with $n$), the two approaches also have identical model-discrimination ability.
Testing Sharp-Null Hypotheses (continued)

Interval (alpha matched to BIC)

<table>
<thead>
<tr>
<th>n</th>
<th>M.5 correct</th>
<th>M.6 correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.15819</td>
<td>0.18017</td>
</tr>
<tr>
<td>20</td>
<td>0.09689</td>
<td>0.12655</td>
</tr>
<tr>
<td>30</td>
<td>0.07172</td>
<td>0.11121</td>
</tr>
<tr>
<td>40</td>
<td>0.05925</td>
<td>0.10871</td>
</tr>
<tr>
<td>50</td>
<td>0.05166</td>
<td>0.10992</td>
</tr>
<tr>
<td>60</td>
<td>0.04632</td>
<td>0.11189</td>
</tr>
<tr>
<td>70</td>
<td>0.04224</td>
<td>0.11569</td>
</tr>
<tr>
<td>80</td>
<td>0.03744</td>
<td>0.11802</td>
</tr>
<tr>
<td>90</td>
<td>0.03598</td>
<td>0.12405</td>
</tr>
<tr>
<td>100</td>
<td>0.03320</td>
<td>0.12853</td>
</tr>
</tbody>
</table>

To emphasize again: none of these model-discrimination behaviors is uniformly optimal; it depends on the real-world consequences of false-positive and false-negative errors; Bayesian decision theory, on a problem-specific basis, is the only sure way to sort this out.
What about $Q_2$: Is $M_1$ good enough?

As discussed previously, by the Modeling-As-Decision Principle a full judgment of adequacy requires real-world input ("To what purpose will the model be put?")}, so it’s not possible to propose generic methodology to answer $Q_2$ (apart from maximizing expected utility, with a utility function that’s appropriately tailored to the problem at hand), but the somewhat related question

$Q_2'$: Could the data have arisen from model $M_j$?

can be answered in a general way by simulating from $M_j$ many times, developing a distribution of (e.g.) $LS_{FS}$ values, and seeing how unusual the actual data set’s log score is in this distribution.

This is related to the posterior predictive model-checking method of Gelman et al. (1996), which produces a $P$-value.

However, this sort of thing needs to be done carefully (Draper 1996), or the result will be poor calibration; indeed, Bayarri and Berger (2000) and Robins et al. (2000) have demonstrated that the
Gelman et al. procedure may be (sharply) conservative: You may get $P = 0.4$ from Gelman et al. (indicating that Your model is fine) when a well-calibrated version of their idea would have $P = 0.04$ (indicating that it’s not fine).

Using a modification of an idea suggested by Robins et al., Draper and Krnjačić (2010) have developed a simulation-based method for accurately calibrating the log-score scale (I’d be happy to send You the paper).

How should You judge how unusual the actual data set’s log score is in the simulation distribution?

In all of Bayesian inference, prediction and decision-making, except for calibration concerns, there’s no need for $P$-values, but — since this is a calibrative question — it’s no surprise that tail areas (or something else equally ad-hoc, such as the ratio of the attained height to the maximum height of the simulation distribution) arise.

I don’t see how to avoid this ad-hockery except by directly answering $Q_2$ with decision theory (instead of answering $Q_2'$ with a tail area).
Summary

- 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 progression from

  Principles $\rightarrow$ Axioms $\rightarrow$ Theorem

  to prove that Bayesian reasoning is justified under natural logical consistency assumptions; for me this secures the foundations of applied 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
— it doesn’t offer any advice on how to specify the required ingredients: with $\theta$ as the unknown of principal interest, $B$ as Your relevant background assumptions and judgments, and an information source (data set) $D$ relevant to decreasing Your uncertainty about $\theta$, the ingredients are

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

* in addition $\{A, U(a, \theta)\}$ for decision, where $A$ is Your set of available actions and $U(a, \theta)$ is Your utility function (mapping from actions $a$ and unknown $\theta$ 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 deficiencies above:
Summary (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.

Interest in calibration can be seen to be natural in Bayesian work by thinking decision-theoretically, with a utility function that rewards both quality of scientific conclusions and good calibration of the modeling process yielding those conclusions.

- 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 $M = \{M_1, M_2, \ldots\}$, which gives rise immediately to two questions:

  - $Q_1$: Is $M_1$ better than $M_2$?
  - $Q_2$: Is $M_1$ good enough?
• These questions sound fundamental but are not: better for what purpose? Good enough for what purpose? To address the second of the two deficiencies above (lack of guidance from Cox’s Theorem on model specification), this implies a

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

This solves the model-specification problem but is hard work; there’s a powerful desire for generic model-comparison methods whose utility structure may provide a decent approximation to problem-specific utility elicitation.

Two such methods are Bayes factors (whose utility justification is less than compelling) and log scores, which are based on the

**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.
• I’m aware of three approaches to improved assessment and propagation of model uncertainty: Bayesian model averaging (BMA), Bayesian nonparametric (BNP) modeling, and calibration (3-fold) cross-validation (CCV).

• CCV provides a way to pay the right price for hunting around in the data for good models, motivating the following modeling algorithm:

(a) Start at a model $M_0$ (how choose?); set the current model $M_{\text{current}} \leftarrow M_0$ and the current model ensemble $M_{\text{current}} \leftarrow \{M_0\}$.

(b) If $M_{\text{current}}$ is good enough to stop (how decide?), return $M_{\text{current}}$; else

(c) Generate a new candidate model $M_{\text{new}}$ (how choose?) and set $M_{\text{current}} \leftarrow M_{\text{current}} \cup M_{\text{new}}$.

(d) If $M_{\text{new}}$ is better than $M_{\text{current}}$ (how decide?), set $M_{\text{current}} \leftarrow M_{\text{new}}$.

(e) Go to (b).

• For the choice in (a), there’s usually a default off-the-shelf initial model based on the structure of the data set $D$ and the scientific context.
• In manual model search the choice in (c) is typically based on the results of a variety of diagnostics, with the new model suggested by deficiencies revealed in this way; at present, we have no better way to automate this choice in many cases than choosing $M_{\text{new}}$ at random (I offer no new ideas on this topic today).

• In comparing $M_1$ with $M_2$ (the choice in (d)), consider a calibrative scenario in which the data-generating model $M_{\text{DG}}$ is one or the other of $\mathcal{M} = \{M_1, M_2\}$ (apart from parameter estimation), and call {choosing $M_2$ when $M_{\text{DG}} = M_1$} a false positive and {choosing $M_1$ when $M_{\text{DG}} = M_2$} a false negative; then

— The right way to do this, following the Modeling-As-Decision Principle, is to build a utility function by quantifying the real-world consequences of

\[
\{\text{choosing } M_1 \text{ when } M_{\text{DG}} = M_1, \text{ choosing } M_1 \text{ when } M_{\text{DG}} = M_2, \\
\text{choosing } M_2 \text{ when } M_{\text{DG}} = M_1, \text{ choosing } M_2 \text{ when } M_{\text{DG}} = M_2 \}
\]

and maximize expected utility.
— If instead You contemplate using Bayes factors/BIC or log scores, it is not the case that one of these two methods uniformly dominates the other in calibrative performance; in some settings they behave the same, in others (for Your sample size) they will have a different balance of false positives and false negatives; it’s a good idea to investigate this before settling on one method or the other.

- See Draper and Krnjajić (2010) for a method for answering the question \( Q_2' \): Could the data have arisen from model \( M_j \)? in a well-calibrated way.

- CCV provides an approach to finding a good ensemble \( M \) of models, and gives You a decent opportunity both to arrive at good answers to Your main scientific questions and to evaluate the calibration of the iterative modeling process that led You to Your answers.

- **Decision-Versus-Inference Principle:** We should all get out of the habit of using inferential methods to make decisions: their implicit utility structure is often far from optimal.
Another Unsolved Foundational Problem

- One more **unsolved foundational problem**: how can **good decisions** be arrived at when “You” is a **collective of individuals**, all with their **own utility functions** that imply **partial cooperation** and **partial competition**?

**Example:** Allocation of finite resources by **two or more people** who have **agreed to band together** in some sense (i.e., **politics**, at the level of **family** or **nation** or ...).

**An instance of this:** Defining and funding **good quality of health care** — the **actors** in the drama include

\{patient, doctor, hospital, state and local regulatory bodies, federal regulatory system\};

all are in **partial agreement** and **partial disagreement** on how (and how many) **resources** should be **allocated** to the **problem** of addressing this patient’s immediate health needs.

(But that’s for **another day**, as is the topic of **Bayesian computing** with **large data sets**.)