False-Positive/False-Negative Trade-Offs in Bayesian Model Comparison

David Draper

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

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

SBIES, SANTA CRUZ

27 Apr 2012

(See Draper (2012: Bayesian model specification: heuristics and examples. In Bayesian Theory and Applications (Damien P, Dellaportas P, Polson N, and Stephens D, editors), forthcoming) for an example of calibration cross-validation (CCV, page 11) in action.)
(1) The big remaining challenge in the Bayesian paradigm is **optimal model specification**, where \( \text{model} = \{\text{prior, sampling distribution/likelihood}\} \) for inference/prediction and \( \text{model} = \{\text{prior, sampling distribution/likelihood, action space, utility function}\} \) for decision-making; in this talk model = \( \{\text{prior, sampling distribution/likelihood}\} \).

(2) In practice You’ll almost always have **uncertainty** about how to specify one or more of \( \{\text{prior, sampling distribution/likelihood}\} \); this means You’ll need tools for **model comparison**: is \( M_1 \) better than \( M_2 \)?

(3) Model comparison is really a decision problem in disguise, which should be solved by formulating a utility function specific to the **given situation** and maximizing expected utility, but this is **hard work**; there’s a strong **desire** for **model-comparison tools** based on generic utility functions.

(4) Two such tools are **Bayes factors** and **log scores**.
(5) All scientists — Bayesian or non-Bayesian — need to pay attention to calibration (how often they get the right answer); this is a basic scientific imperative, and calibration results are part of the definition of “optimal” in (1).

(6) It’s often not OK to do model comparison on the same data set on which You’ll derive your inferential/predictive answers, based on how the model comparison comes out — this can lead to poor calibration; a method called calibration cross-validation (CCV) solves this problem, and allows Bayesians to compare models in a well-calibrated way.

(7) One consequence of (3) is that all statisticians — Bayesian or non-Bayesian — need to make (or at least pay attention to) calibration calculations in which (a) You (temporarily) assume an underlying data-generating mechanism \( M_{DG} \) (truth) and (b) You keep track of how often Your method recovers known truth (e.g., how often does model-comparison method \( A \) correctly identify \( M_{DG} \)?).
When comparing $M_1$ and $M_2$, let’s agree to say that \{\text{choosing } M_2 \text{ when } M_{DG} \text{ is a special case of } M_1\} \text{ is a false-positive mistake}, and \{\text{choosing } M_1 \text{ when } M_{DG} \text{ is a special case of } M_2\} \text{ is a false-negative mistake.}

In evaluating the calibration performance of model-comparison methods, standard asymptotic calculations are often irrelevant; it’s usually far more useful to make calculations or run simulations over a realistic range of finite sample sizes, comparing false-positive and false-negative error rates.

Popular (Bayesian and non-Bayesian) model-comparison methods include \{AIC, Bayes factors, BIC, DIC, log scores\}; it turns out that none of these methods dominates the others simultaneously on false-positive and false-negative performance.

But you can draw the following broad conclusions in the situation where $M_2$ is more complicated than $M_1$ (e.g., when the number of parameters in $M_2$ is greater):
(a) \{AIC, DIC, log scores\} behave similarly;

(b) \{Bayes factors, BIC\} behave similarly;

(c) \{AIC, DIC, log scores\} behave differently from \{Bayes factors, BIC\};

(d) \{AIC, DIC, log scores\} tend to make more false-positive mistakes than \{Bayes factors, BIC\} and \{Bayes factors, BIC\} tend to make more false-negative mistakes than \{AIC, DIC, log scores\}.

(12) To choose a model-comparison method well in Your problem, You need to think about the real-world consequences of false-positive and false-negative mistakes; in other words, choosing a model-comparison method is itself a decision problem!

(13) One popular setting in which $M_2$ is more complicated than $M_1$ is equivalent to sharp-null hypothesis-testing — e.g., the data are IID $N(0, \sigma^2)$ under $M_1$ ($H_0$) and IID $N(\mu, \sigma^2)$ under $M_2$ ($H_A$) — but in practice there are actually very few real-world situations in which this comparison is relevant.
Often Your uncertainty about $\mu$ is continuous, in which case the right comparison is between $M_1$: IID $N(\mu, \sigma^2)$, $\mu \in (-a, +b)$ versus $M_1$: IID $N(\mu, \sigma^2)$, $\mu \notin (-a, +b)$; in this setting comparisons between \{AIC, DIC, log scores\} and \{Bayes factors, BIC\} have very different results from those when comparing $M_1$: IID $N(0, \sigma^2)$ with $M_1$: IID $N(\mu, \sigma^2)$.

Old theorem: $M_1$: IID $N(0, \sigma^2)$ versus $M_1$: IID $N(\mu, \sigma^2)$, sample size $n$, $LS = \text{log scores}$;

(a) $P_{RS}[\text{BIC chooses } M_2 | M_{DG} = M_2(\mu)] \to 1$ as $n \to \infty$ for all $\mu \neq 0$;

(b) $P_{RS}(\text{BIC chooses } M_1 | M_{DG} = M_1) \to 1$ as $n \to \infty$;

(c) $P_{RS}[\text{LS chooses } M_2 | M_{DG} = M_2(\mu)] \to 1$ as $n \to \infty$ for all $\mu \neq 0$;

(d) $P_{RS}(\text{LS chooses } M_1 | M_{DG} = M_1) \to 0$ as $n \to \infty$.

In other words, asymptotic consistency of BIC both under $M_1$ and $M_2$, and asymptotic consistency of LS under $M_2$ but not under $M_1$, when comparing the sharp-null $M_1$ with the composite $M_2$.

However,
New theorem (Draper, 2012): $M_1$: IID $N(\mu, \sigma^2)$, $\mu \in (-a, +b)$ versus $M_1$: IID $N(\mu, \sigma^2)$, $\mu \notin (-a, +b)$, sample size $n$, $LS = \log$ scores;

(a) $P_{RS}[BIC \text{ chooses } M_2| M_{DG} = M_2(\mu)] \to 1$ as $n \to \infty$ for all $\mu \neq 0$;

(b) $P_{RS}(BIC \text{ chooses } M_1| M_{DG} = M_1) \to 1$ as $n \to \infty$;

(c) $P_{RS}[LS \text{ chooses } M_2| M_{DG} = M_2(\mu)] \to 1$ as $n \to \infty$ for all $\mu \neq 0$;

(d) $P_{RS}(LS \text{ chooses } M_1| M_{DG} = M_1) \to 1$ as $n \to \infty$;

In other words, asymptotic consistency of both BIC and LS, both under $M_1$ and $M_2$, when comparing the — often much more realistic — composite $M_1$ with the composite $M_2$.

Thus the comparison between BIC and LS is not as cut-and-dried as the Bayes-factor people would have you believe.
The basic ingredients in a problem involving statistical inference, prediction and/or decision-making are as follows:

- $\theta$, something unknown to You (Good, 1950: a generic person wishing to reason sensibly in the presence of uncertainty).

$\theta$ could be almost anything, but (for concreteness) think of a vector in $\mathbb{R}^k$ for integer $1 < k < \infty$ (all finite-dimensional unknowns can be expressed in this way);

- $D$, an information source (data set) that You judge to be relevant to decreasing Your uncertainty about $\theta$.

$D$ could again be almost anything, but think of a vector in $\mathbb{R}^n$ for integer $1 \leq n < \infty$ (all data sets can be expressed in this way);

- $B$, a (true/false) proposition of the form $(B_1 \text{ and } B_2 \text{ and } \ldots \text{ and } B_b) = (B_1 \ B_2 \ldots \ B_b)$ for integer $1 \leq b < \infty$, where the $B_i$ are propositions, all regarded by You as true, that specify Your background information, assumptions and judgments about the context of the problem and the data-gathering process.
The presence of $D$ creates a dichotomy:

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

**Q:** How should this information be combined for optimal information-processing, to solve the inference, prediction and/or decision-making problem?

**A:** One (not necessarily the only) logically-internally-consistent approach is provided by a theorem of Richard Cox (1946), who regarded probability as an expression of Your rational expectations (de Finetti (1937) has a similar theorem, regarding probability as a quantification of Your betting odds).

The primitive operator in Cox’s framework is $P(A|B)$, where $A$ and $B$ are propositions, with the truth status of $A$ unknown to You and $B$ regarded by You as true; from this You can easily get to CDFs (for real-valued $\theta$) of the form $F_\theta(q|DB) = P(\theta \leq q|DB)$ and densities of the form $p_\theta(q|DB) = \frac{\partial}{\partial q} F_\theta(q|DB)$, which I’ll abbreviate $p(\theta|DB)$ in what follows.
Cox’s Theorem says (informally) that, to be logically internally consistent and not lose any information in Your information-processing, You must be prepared to specify the following two ingredients for inference and prediction:

- $p(\theta|B)$, usually called Your prior distribution for $\theta$ (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);

- $p(D|\theta B)$, 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 following additional two ingredients for decision-making:

- the set $A$ of feasible actions among which You’re choosing, and

- a utility function $U(a, \theta)$, taking values on $\mathbb{R}$ and quantifying Your judgments about the costs and benefits (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**.

The **theorem** further says that, having **specified** these **four ingredients**, You must **combine them** in the following **ways** to solve Your inference, prediction and/or decision-making problem:

- The **distribution** \( p(\theta|D B) \) quantifies all relevant information about \( \theta \), both **internal and external** to \( D \), and must be computed via **Bayes’s Theorem**:

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

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

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

\[
p(D^*|D B) = \int_{\Theta} p(D^*|\theta D B) \ p(\theta|D B) \ d\theta;
\]
often there’s no information about $D^*$ contained in $D$ if $\theta$ is known, in which case this expression simplifies to

$$p(D^*|DB) = \int_\Theta p(D^*|\theta B) p(\theta|DB) d\theta; \quad \text{(prediction)} \quad (2)$$

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

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

In view of Cox’s Theorem, the problem now becomes:

How can You specify the four ingredients $p(\theta|B)$, $p(D|\theta B)$, and \{A, U(a, \theta)\} well (in fact, can this be done optimally?)?

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).

For the rest of this talk I’ll concentrate on inference and prediction, which require specifying \{p(\theta|B), p(D|\theta B)\} — call such a specification a model $M$ for Your uncertainty about $\theta$. 
**The Calibration Principle**

As a profession, we currently don’t have a theorem, like Cox’s Theorem, that tells us how to specify Bayesian models optimally; the best we can do at present is appeal to a set of principles that can provide some guidance.

Here’s one that makes good sense to me:

**Calibration Principle:** In model specification, it’s helpful to 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.

**(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, 2012) 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.
The $M^*$ Approach

Having adopted the Calibration Principle, it makes sense to talk about an underlying data-generating model $M_{DG}$, which is unknown to You.

From now on I’ll focus on the sampling distribution $p(D|\theta \mathcal{B})$.

**Q:** How can You specify $p(D|\theta \mathcal{B})$ in a well-calibrated way?

**How not to do this:** 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 $M^*$ “=” $M_{DG}$.

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 (this is an example of what I mean by comparing actual performance with known truth).

A: One approach to solving this problem is calibration cross-validation (CCV):

- 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 $M$ to construct posterior predictive distributions for all of the data values in $V$ and see how the latter compare with the former).
Calibration Cross-Validation (CCV)

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 $M = \{M_1, \ldots, M_m\}$,

(c) see how well the models in $M$ 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 $M$ (or, better, use BMA) on the data in $(M \cup V)$, 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 \((M \cup V)\), 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, 2012) 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 $\mathcal{M}_{\text{current}} \leftarrow \{M_0\}$.

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

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

(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 question in (d):

$Q$: Is $M_1$ better than $M_2$?
The Modeling-As-Decision Principle

This question sounds fundamental but is not: better 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,
\[
\frac{p(M_2|DB)}{p(M_1|DB)} = \frac{p(M_2|B)}{p(M_1|B)} \cdot \frac{p(D|M_2 B)}{p(D|M_1 B)};
\]

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 (4) says

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

(Bayes factors seem to have first been considered by Turing and Good (∼ 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

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

(6)

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 \( p(M_j|\mathcal{B}) \) 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 \( M_1 \),

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

(7)
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

$$M_j : \left\{ \begin{array}{l} (\gamma_j | M_j B) \sim p(\gamma_j | M_j B) \\ (D | \gamma_j M_j B) \sim p(D | \gamma_j M_j B) \end{array} \right\} . \quad (8)$$

Here the quantity $p(D|M_j B)$ that defines the Bayes factor is
Integrated Likelihoods

\[ p(D|M_j \mathcal{B}) = \int p(D|\gamma_j M_j \mathcal{B}) p(\gamma_j | M_j \mathcal{B}) \, d\gamma_j ; \quad (9) \]

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

The first trouble is **technical**: the integral in (9) can be **difficult to compute**, and may not even be easy to **approximate**.

The second thing to **notice** is that (9) can be **rewritten** as

\[ p(D|M_j \mathcal{B}) = E_{(\gamma_j|M_j \mathcal{B})} \, p(D|\gamma_j M_j \mathcal{B}) . \quad (10) \]

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

**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

\[ M_1 = \text{Geometric}(\theta_1) \text{ likelihood with a } \text{Beta}(\alpha_1, \beta_1) \text{ prior on } \theta_1; \]
\[ M_2 = \text{Poisson}(\theta_2) \text{ likelihood with a } \text{Gamma}(\alpha_2, \beta_2) \text{ prior on } \theta_2. \]

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

\[
\frac{\Gamma(\alpha_1 + \beta_1) \Gamma(n + \alpha_1) \Gamma(n\bar{y} + \beta_1) \Gamma(\alpha_2) (n + \beta_2)^{n\bar{y} + \alpha_2} \left( \prod_{i=1}^{n} y_i! \right)}{\Gamma(\alpha_1) \Gamma(\beta_1) \Gamma(n + n\bar{y} + \alpha_1 + \beta_1) \Gamma(n\bar{y} + \alpha_2) \beta_2^{\alpha_2}}. \tag{11}
\]

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

\[
\frac{\Gamma(n + 1) \Gamma(n\bar{y} + 1) \Gamma(\epsilon) (n + \epsilon)^{n\bar{y} + \epsilon} \left( \prod_{i=1}^{n} y_i! \right)}{\Gamma(n + n\bar{y} + 2) \Gamma(n\bar{y} + \epsilon) \epsilon^{\epsilon}}. \tag{12}
\]

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 \(y\)).
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 \(\{\text{partial, intrinsic, fractional}\}\) 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

\[
p(D|M_j B) = \int p(D|\gamma_j M_j B) \ p(\gamma_j|M_j B) \ d\gamma_j ; \tag{13}
\]

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

\[
\log p(D | M_j B) = \log p(D | \hat{\gamma}_j M_j B) + \log p(\hat{\gamma}_j | M_j B)
+ \frac{k_j}{2} \log 2\pi - \frac{1}{2} \log |\hat{I}_j| + O \left( \frac{1}{n} \right) ; \quad (14)
\]

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):

$$\log p(y | M_j \mathcal{B}) = \log p(y | \hat{\gamma}_j M_j \mathcal{B}) - \frac{k_j}{2} \log n + O(1). \quad (15)$$

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

$$BIC(M_j | D \mathcal{B}) = -2 \log p(D | \hat{\gamma}_j M_j \mathcal{B}) + k_j \log n \quad (16)$$

(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 and the Unit-Information Prior

BIC is called an information criterion because it resembles AIC (Akaike, 1974), which was derived using information-theoretic reasoning:

$$AIC(M_j|DB) = -2 \log p(D|\hat{\gamma}_j M_j B) + 2 k_j.$$  \hspace{1cm} (17)

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

$$(\gamma_j|M_j B) \sim N_{k_j}(\hat{\gamma}_j, n\hat{I}_j^{-1}).$$  \hspace{1cm} (18)

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 20 that there are two generic utility-based model-comparison methods: Bayes factors and log scores.

- Log scores 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.

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

\[ M_j: \left\{ \begin{align*}
(\gamma_j | M_j \mathcal{B}) & \sim p(\gamma_j | M_j \mathcal{B}) \\
(y | \gamma_j, M_j \mathcal{B}) & \sim p(y | \gamma_j, M_j \mathcal{B})
\end{align*} \right\}. \tag{19} \]

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

\[ LS_{CV}(M_j | y \mathcal{B}) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | y_{-i} M_j \mathcal{B}), \tag{20} \]

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,
\[ LS_{FS}(M_j|y B) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|y M_j B), \quad (21) \]

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:

\[ KL[M_{DG}||p(\cdot|y M_j B)] = E_{M_{DG}} \log M_{DG} - E_{M_{DG}} \log p(\cdot|y M_j B) \]
\[ = E_{M_{DG}} \log M_{DG} - LS_{FS}(M_j|y B); \quad (22) \]

the first term on the right side of (22) 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 (e.g., intrinsic Bayes factors: Berger and Pericchi, 1996, and more recent cousins) tend to have model discrimination performance similar to that of BIC in calibration (repeated-sampling with known $M_{DG}$) environments; I’ll show results for BIC here.

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 $\theta$.

This is an example of the

- **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.

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, 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:

\begin{align*}
M_1: \left\{ \begin{array}{l}
(\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} \\
M_2: \left\{ \begin{array}{l}
(\theta|B) \sim \text{diffuse for } \theta > \Delta \\
(y_i|\theta B) \sim \text{IID } N(\theta, \sigma^2)
\end{array} \right\}, \quad (23)
\end{align*}
in which for simplicity I’ll take \( \sigma^2 \) to be known (the results are similar with \( \sigma^2 \) learned from the data).

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

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

- **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 
  \[ p(\theta > \Delta|y M^* B) > 0.5. \]

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

**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; Monte-Carlo approximations of the predictive ordinates in \( \text{LS}_{FS} \) based on 10,000 posterior draws.

The figures below give Monte-Carlo estimates of the probability that \( M_2 \) is chosen.
This exhibits all the monotonicities that it should, and correctly yields 0.5 for all $n$ with $\theta_{DG} = 15$. 

LS$_{FS}$ Results: Quantifying Improvement
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 to within $\pm 0.003$ (well within simulation noise with 1,000 replications).
Here BIC and the posterior-probability approach are algebraically identical, making the model-discrimination performance of all three approaches the same in this problem.
• (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

$$
[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})]
$$

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 $|\theta| \leq \lambda$, where $\lambda > 0$ 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**

\[
M_3: \begin{cases} 
(\theta|B) \sim \text{diffuse for } |\theta| \leq \lambda \\
(y_i|\theta B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{cases}
\]

and

\[
M_4: \begin{cases} 
(\theta|B) \sim \text{diffuse for } |\theta| > \lambda \\
(y_i|\theta B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{cases}
\]

in which $\sigma^2$ is again taken for **simplicity** to be **known**.

A **natural alternative** to **BIC** and $L_{FS}$ here is again based on **posterior probabilities**: as before, let

\[
M^* = \{(\theta|B) \sim \text{diffuse on } \mathbb{R}, (y_i|\theta B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)\},
\]

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

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

\{claiming the two drugs were bio-equivalent when they were, concluding that they were bio-equivalent when they were not, deciding that they were not bio-equivalent when they were, judging that they were not bio-equivalent when they were not\},

but here I’ll again simply compare the calibrative performance of \(LSFS\), posterior probabilities, and BIC.

Simulation experiment details, based on the SBP drug trial: \(\lambda = 5\); \(\sigma = 10\); \(n = 10, 20, \ldots, 100\); 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 \(LSFS\).

\[\text{NB} \quad \text{It has previously been established that when making the (unrealistic) sharp-null comparison } \theta = 0 \text{ versus } \theta \neq 0 \text{ in the context of } (y_i | \theta \mathcal{B}) \overset{\text{iid}}{\sim} N(\theta, \sigma^2), \text{ as } n \to \infty \text{ } LSFS \text{ selects the } \theta \neq 0 \text{ model with probability } \to 1 \text{ even when } \theta_{DG} = 0; \text{ this “inconsistency of log scores at the null model” has been used by some people as a reason to dismiss log scores as a model-comparison method.} \]
In this **more realistic setting**, comparing $|\theta| \leq \lambda$ versus $|\theta| > \lambda$ with $\lambda > 0$, $LS_{FS}$ has the **correct large-sample behavior**, both when $|\theta_{DG}| \leq \lambda$ and when $|\theta_{DG}| > \lambda$. 
The qualitative behavior of the $LS_{FS}$ and posterior-probability methods is identical, although there are some numerical differences (highlighted later).
In the quantifying-improvement case, the BIC and posterior-probability methods were algebraically identical; here they nearly coincide (differences of $\pm 0.001$ with 1,000 simulation repetitions).
If you call choosing $M_4$: $|\theta| > \lambda$ when $|\theta_{DG}| \leq \lambda$ a false-positive error and choosing $M_3$: $|\theta| \leq \lambda$ when $|\theta_{DG}| > \lambda$ a false-negative mistake, with $n = 10$ there's a trade-off: $LS_{FS}$ has more false positives and BIC has more false negatives.
By the time you reach $n = 50$ in this problem, $LS_{FS}$ and BIC are essentially equivalent.
An extreme example of the false-positive/false-negative differences between $LS_F$ and BIC in this setting may be obtained, albeit unwisely, by letting $\lambda \downarrow 0$.

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 $H_A: \theta \neq 0$.

It’s necessary to distinguish between problems in which there is or is not a structural singleton in the (continuous) set $\Theta$ of possible values of $\theta$: settings where it’s scientifically important to distinguish between $\theta = \theta_0$ and $\theta \neq \theta_0$ — an example would be discriminating between \{these two genes are on different chromosomes (the strength $\theta$ of their genetic linkage is $\theta_0 = 0$)\} and \{these two genes are on the same chromosome ($\theta > 0$)\}.

Sharp-null testing without structural singletons is always 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 $\theta \neq 0$ — with realistic values of $n$ — 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 ($i = 1, \ldots, n$)

$$M_5: \left\{ \begin{array}{l} (\sigma^2 | B) \sim \text{diffuse on } (0, \text{large}) \\ (y_i | \sigma^2, B) \sim_{\text{IID}} N(0, \sigma^2) \end{array} \right\} \quad \text{and} \quad (28)$$

$$M_6: \left\{ \begin{array}{l} (\theta, \sigma^2 | B) \sim \text{diffuse on } (-\text{large}, \text{large}) \times (0, \text{large}) \\ (y_i | \theta, \sigma^2, B) \sim_{\text{IID}} N(\theta, \sigma^2) \end{array} \right\}, \quad (29)$$

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.
Simulation experiment details: data-generating $\sigma_{DG} = 10$; $n = 10, 20, \ldots, 100$; data-generating $\theta_{DG} = \{0, 1, \ldots, 5\}$; 1,000 simulation replications, $M = 100,000$ Monte-Carlo draws for $LS_{FS}$; the figures below give Monte-Carlo estimates of the probability that $M_6$ is chosen.

As before, let’s call choosing $M_6$: $\theta \neq 0$ when $\theta_{DG} = 0$ a false-positive error and choosing $M_5$: $\theta = 0$ when $\theta_{DG} \neq 0$ a false-negative mistake.
In the limit as $\lambda \downarrow 0$, the $LS_{FS}$ approach makes hardly any false-negative errors but quite a lot of false-positive mistakes.
Interval ($\alpha = 0.05$) Results: Sharp-Null Testing

The behavior of the posterior interval approach is of course quite different: it makes many false-negative errors because its rate of false-positive mistakes is fixed at 0.05.
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.
BIC’s behavior is quite different from that of $LS_{FS}$ and fixed-$\alpha$ posterior intervals: its false-positive rate decreases as $n$ grows, but it suffers a high false-negative rate to achieve this goal.
When the interval method is modified so that \( \alpha \) matches the BIC behavior at \( \theta_{DG} = 0 \) (again letting \( \alpha \) vary with \( n \)), the two approaches have identical model-discrimination ability.
As another **model-comparison example**, suppose you have an **integer-valued** data set \( D = y = (y_1 \ldots y_n) \) and you wish to compare

\[ M_7 = \text{Geometric}(\theta_1) \text{ sampling distribution} \text{ with a} \]
\[ \text{Beta}(\alpha_1, \beta_1) \text{ prior on } \theta_1, \text{ and} \]

\[ M_8 = \text{Poisson}(\theta_2) \text{ sampling distribution} \text{ with a} \]
\[ \text{Gamma}(\alpha_2, \beta_2) \text{ prior on } \theta_2. \]

\( \text{LS}_{FS} \) and \( \text{BIC} \) both have **closed-form expressions** in this situation:

with \( s = \sum_{i=1} y_i \) and \( \hat{\theta}_1 = \frac{\alpha_1 + n}{\alpha_1 + \beta_1 + s + n}, \)

\[ \text{LS}_{FS}(M_7|y B) = \log \Gamma(\alpha_1 + n + \beta_1 + s) + \log \Gamma(\alpha_1 + n + 1) \]
\[ - \log \Gamma(\alpha_1 + n) - \log \Gamma(\beta_1 + s) \]
\[ + \frac{1}{n} \sum_{i=1}^{n} [\log \Gamma(\beta_1 + s + y_i)] \]
\[ - \log \Gamma(\alpha_1 + n + \beta_1 + s + y_i + 1), \] (30)

\[ \text{BIC}(M_7|y B) = -2[n \log \hat{\theta}_1 + s \log(1 - \hat{\theta}_1)] + \log n, \] (31)
Geometric Versus Poisson (continued)

\[ LS_{FS}(M_8|y \mathcal{B}) = (\alpha_2 + s) \log(\beta_2 + n) - \log \Gamma(\alpha_2 + s) \]
\[ - (\alpha_2 + s) \log(\beta_2 + n + 1) \]
\[ + \frac{1}{n} \sum_{i=1}^{n} [\log \Gamma(\alpha_2 + s + y_i) - y_i \log(\beta_2 + n + 1) \]
\[ - \log \Gamma(y_i + 1)] , \text{ and} \]
\[ BIC(M_8|y \mathcal{B}) = -2[s \log \hat{\theta}_2 - n \hat{\theta}_2 - \sum_{i=1}^{n} \log(y_i!)] + \log n , \] (33)

where \( \hat{\theta}_2 = \frac{\alpha_2 + s}{\beta_2 + n} \).

Simulation details: \( n = \{10, 20, 40, 80\} \), \( \alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 0.01 \), 1,000 simulation replications; it turns out that with \( (\theta_1)_{DG} = 0.5 \) (Geometric) and \( (\theta_2)_{DG} = 1.0 \) (Poisson), both data-generating distributions are monotonically decreasing and not easy to tell apart by eye.

Let's call choosing \( M_8 \) (Poisson) when \( M_{DG} = \text{Geometric} \) a false-Poisson error and choosing \( M_7 \) (Geometric) when \( M_{DG} = \text{Poisson} \) a false-Geometric mistake.
The table below records the Monte-Carlo probability that the Poisson model was chosen.

<table>
<thead>
<tr>
<th>n</th>
<th>LS.FS</th>
<th>BIC</th>
<th>n</th>
<th>LS.FS</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.8967</td>
<td>0.8661</td>
<td>10</td>
<td>0.4857</td>
<td>0.4341</td>
</tr>
<tr>
<td>20</td>
<td>0.9185</td>
<td>0.8906</td>
<td>20</td>
<td>0.3152</td>
<td>0.2671</td>
</tr>
<tr>
<td>40</td>
<td>0.9515</td>
<td>0.9363</td>
<td>40</td>
<td>0.1537</td>
<td>0.1314</td>
</tr>
<tr>
<td>80</td>
<td>0.9846</td>
<td>0.9813</td>
<td>80</td>
<td>0.0464</td>
<td>0.0407</td>
</tr>
</tbody>
</table>

Both methods make more false-Poisson errors than false-Geometric mistakes; the results reveal once again that neither BIC nor $LS_{FS}$ uniformly dominates — each has a different pattern of false-Poisson and false-Geometric errors ($LS_{FS}$ correctly identifies the Poisson more often than BIC does, but as a result BIC gets the Geometric right more often than $LS_{FS}$).
Properties of $LS_{FS}$

- Log scores are entirely free from the diffuse-prior problems bedeviling Bayes factors:

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

in which

$$p(y_i|y M_j B) = \int p(y_i|\gamma_j M_j B) p(\gamma_j|y M_j B) d\gamma_j$$

(34)

this expectation is over the posterior (not the prior) distribution for the parameter vector $\gamma_j$ in model $M_j$, and is therefore completely stable with respect to small variations in how prior diffuseness (if scientifically called for) is specified, even with only moderate $n$.

- Following the Modeling-As-Decision Principle, the decision-theoretic justification for Bayes factors involves not only the Bayes factors themselves but also the prior model probabilities, which can be hard to specify in a scientifically-meaningful way: under the Bayes-factor (possibly unrealistic) 0/1 utility structure,
You’re supposed to choose the model with the highest posterior probability, not the one with the biggest Bayes factor.

By contrast, specification of prior model probabilities doesn’t arise with log scores, which have a direct decision-theoretic justification based on the Prediction Principle.

• It may seem that log scores have no penalty for unnecessary model complexity, but this is not true: for example, if one of Your models carries around a lot of unnecessary parameters, this will needlessly inflate its predictive variances, making the heights of its predictive densities go down, thereby lowering its log score.

• It may also seem that the behavioral rule based on posterior Bayes factors (Aitkin 1991) is the same as the rule based on \( L_{FS} \), which favors model \( M_j \) over \( M_{j'} \) if

\[
n L_{FS}(M_j|y, B) > n L_{FS}(M_{j'}|y, B). \tag{35}
\]

But this is not true either: for example, in the common situation in which the data set \( D \) consists of observations \( y_i \) that are conditionally IID from \( p(y_i|\eta_j, M_j, B) \) under \( M_j \),
Summary

\[ nLS_{FS}(M_j | y, \mathcal{B}) = \log \prod_{i=1}^{n} \left[ \int p(y_i | \eta_j, M_j, \mathcal{B}) \ p(\eta_j | y, M_j, \mathcal{B}) \ d\eta_j \right], \quad (36) \]

and this is not the same as

\[ \log \int \left[ \prod_{i=1}^{n} p(y_i | \eta_j, M_j, \mathcal{B}) \right] \ p(\eta_j | y, M_j, \mathcal{B}) \ d\eta_j = \bar{L}_j^{PBF} \quad (37) \]

because the product and integral operators do not commute.

● Some take-away messages:

— In the bio-equivalence example, even when You (unwisely) let \( \lambda \downarrow 0 \), thereby testing a sharp-null hypothesis, the asymptotic behavior of log scores is irrelevant; what counts is the behavior of log scores and Bayes factors with Your sample size and the models being compared, and for any given \( n \) it’s not possible to say that the false-positive/false-negative trade-off built into Bayes factors is universally better for all applied problems than the false-positive/false-negative trade-off built into log scores,
or vice versa — You have to think it through in each problem.

For instance, the tendency of log scores to choose the “bigger” model in a nested-model comparison is exactly the right qualitative behavior in the following two examples (and many more such examples exist):

— Variable selection in searching through many compounds or genes to find successful treatments: here a false-positive mistake (taking an ineffective compound or gene forward to the next level of investigation) costs the drug company $C$, but a false-negative error (failing to move forward with a successful treatment, in a highly-competitive market) costs $kC$ with $k = 10–100$.

— In a two-arm clinical-trial setting, consider the random-effects Poisson regression model

\[
(y_i|\lambda_i, B) \overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i),
\]

\[
\log \lambda_i = \beta_0 + \beta_1 x_i + e_i
\]

\[
(e_i|\sigma_e^2, B) \overset{\text{IID}}{\sim} N(0, \sigma_e^2), \quad (\beta_0, \beta_1, \sigma_e^2) \sim \text{diffuse},
\]
where the $y_i$ are counts of a relatively rare event and $x_i$ is 1 for the treatment group and 0 for control; You would consider fitting this model instead of its fixed-effects counterpart, obtained by setting $\sigma_e^2 = 0$, to describe unexplainable heterogeneity (Poisson over-dispersion).

In this setting, Bayes factors will make the mistake of \{telling You that $\sigma_e^2 = 0$ when it’s not\} more often than log scores, and log scores will make the error of \{telling You that $\sigma_e^2 > 0$ when it’s actually 0\} more often than Bayes factors, but the former mistake is much worse than the latter, because You will underpropagate uncertainty about the fixed effect $\beta_1$, which is the whole point of the investigation.

- All through this discussion it’s vital to keep in mind that

the gold standard for false-positive/false-negative behavior is provided neither by Bayes factors nor by log scores but instead by Bayesian decision theory in Your problem.
• Asymptotic conclusions are often misleading: while it’s true that

**Old Theorem:** \( P_{\theta_{DG}=0}(LS_{FS} \text{ chooses } \theta = 0) \to 0 \text{ as } n \to \infty, \)

it’s also true that

**New Theorem** (Draper, 2011): for any \( \lambda > 0, \)

\[ P_{|\theta_{DG}| \leq \lambda}(LS_{FS} \text{ chooses } |\theta| \leq \lambda) \to 1 \text{ as } n \to \infty, \]

and the second theorem would seem to call the relevance of the first theorem into question.

• As a profession, we need to strengthen the progression

**Principles \to Axioms \to Theorems**

in optimal model specification; the Calibration Principle, the Modeling-As-Decision Principle, the Prediction Principle and the Decision-Versus-Inference Principle seem helpful in moving toward this goal.
Is $M_1$ Good Enough?

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 Krnjajić (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).
• 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

  \[
  \text{Principles } \rightarrow \text{ Axioms } \rightarrow \text{ Theorem}
  \]

  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

\[ \{A, U(a, \theta)\} \]

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 θ but (b) You’re also uncertain about how to quantify Your uncertainty about θ, i.e., You have model uncertainty.

- This acknowledgment of Your model uncertainty implies a willingness by You to consider two or more models in an ensemble \( \mathcal{M} = \{M_1, M_2, \ldots \} \), which gives rise immediately to two questions:

\[
Q_1: \text{Is } M_1 \text{ better than } M_2? \quad Q_2: \text{Is } M_1 \text{ good enough?}
\]
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.
Summary (continued)

- 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 $M_{new}$ 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 $M_{DG}$ is one or the other of $\mathcal{M} = \{M_1, M_2\}$ (apart from parameter estimation), and call 
  
  \{choosing $M_2$ when $M_{DG} = M_1$\} a false positive and 
  \{choosing $M_1$ when $M_{DG} = M_2$\} 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

  \{choosing $M_1$ when $M_{DG} = M_1$, choosing $M_1$ when $M_{DG} = M_2$, choosing $M_2$ when $M_{DG} = M_1$, choosing $M_2$ when $M_{DG} = M_2$\}

  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**.)