Topics in Bayesian Modeling:
(1) Log Scores for Model Comparison and
(2) a Bayesian Non-Parametric Look at the
Frequentist Bootstrap

David Draper

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

and eBay Research Labs (San José CA)

{draper@ams.ucsc.edu, davdraper@ebay.com}
www.ams.ucsc.edu/~draper

UC Irvine Statistics Seminar

15 May 2014
1. **Log Scores** for **Model Comparison**

2. A **Bayesian non-parametric look** at the **frequentist bootstrap**
There are two rather generic ways to perform model comparisons in the Bayesian paradigm: Bayes factors and log scores.

Some people who like Bayes factors have tried to claim that log scores are "not Bayesian."

In the first part of this talk I will

(a) refute this claim,

(b) demonstrate that neither method uniformly dominates the other in model discrimination ability, and therefore

(c) advocate for a flexible position in which Bayesians should use whichever of the two methods performs better, on a problem-specific basis.
In the **Bayesian statistical paradigm**, when You (Good, 1950: a person wishing to reason sensibly in the presence of uncertainty) are solving a problem $\mathbb{P}$ involving inference, prediction and/or decision-making, You begin with three ingredients induced by $\mathbb{P}$:

- **an unknown $\theta$ of principal interest** (think of a vector in $\mathbb{R}^k$),
- **a data set $D$ (think of a vector in $\mathbb{R}^n$) relevant** to decreasing Your uncertainty about $\theta$, and
- **a finite set of (true/false) propositions $\mathcal{B}$, all true, exhaustively describing the context** of the problem $\mathbb{P}$ and the **data-gathering process** that led to $D$.

With this setup, a foundational theorem — independently developed by Bruno de Finetti (1937) and the American physicist Richard T. Cox (1946), based on different conceptions of the meaning of probability — then says that, if You wish to quantify Your uncertainty about $\theta$ in a logically-internally-consistent manner, one way to accomplish this goal is to specify

$\mathbb{P} \rightarrow (\theta, D, \mathcal{B})$
(\theta, D, B) \rightarrow M = \{ p(\theta|B), p(D|\theta B), (A|B), U(a, \theta|B) \}

(a) two probability distributions for inference and prediction, namely Your prior distribution \( p(\theta|B) \) — to quantify Your information about \( \theta \) external to \( D \) — and Your sampling distribution \( p(D|\theta B) \) — which, when converted into Your likelihood function \( \ell_c(\theta|DB) = c \, p(D|\theta B) \) (for some \( c > 0 \)), quantifies Your information about \( \theta \) internal to \( D \), respectively, and

(b) two additional ingredients for decision-making, namely Your action space \( (A|B) \) (of possible behavioral choices \( a \)) and Your utility function \( U(a, \theta^*|B) \), which quantifies and trades off the costs and benefits arising from choosing action \( a \) if the unknown \( \theta \) took on the value \( \theta^* \).

Having specified these four ingredients, which collectively form Your model

\[
M = \{ p(\theta|B), p(D|\theta B), (A|B), U(a, \theta|B) \}
\]

for Your uncertainty about \( \theta \),

(1) the inference problem is solved with Bayes’s Theorem,

\[
p(\theta|DB) \propto p(\theta|B) \, \ell_c(\theta|DB) ,
\]
in which Your **posterior distribution** \( p(\theta|D B) \) summarizes the **totality** of Your **information** about \( \theta \);

(2) the **prediction problem** is **solved** with the equation

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

in which \( D^* \) is a **new data set** (e.g., future data), \( \Theta \) is the set of **all possible** \( \theta \) **values** and Your **posterior predictive distribution** \( p(D^*|D B) \) quantifies the **totality** of Your **information** about \( D^* \); and

(3) the **decision problem** is **solved** with the equation

\[
a^*_P = \arg\max_{a \in (A|B)} \int_{\Theta} U(a, \theta|B) p(\theta|D B) \, d\theta ,
\]

in which \( a^*_P \) is the **optimal action** in the **principal decision problem** (if any) at the **heart** of \( P \): in **other words**, find the action that **maximizes expected utility**, where the **expectation** is over Your **total-information distribution** \( p(\theta|D B) \).
The Main Substantive Problem: Model Uncertainty

Mapping 1
(typically unique)

\[ P \rightarrow (\theta, D, B) \]

Mapping 2
(rarely unique)

\[ \downarrow \]

\[ \begin{align*}
M &= \{ p(\theta|B), p(D|\theta B), \\
&\quad (A|B), U(a, \theta|B) \} 
\end{align*} \]

Mapping 1 is generally unique, but what about Mapping 2?

It would be nice if the context of the problem \( P \) You’re solving would uniquely determine \( M \) (this could be regarded as an instance of optimal Bayesian model specification; more later), but this is unfortunately rarely true.

In practice, given the current state of understanding of this issue in the statistics profession, You generally have to fall back on basic principles to aid You in the model-specification process, which will involve activities such as answering questions of the form

\[ Q_{MC_1}: \text{Is model } M_2 \text{ better than } M_1? \]
The Modeling-As-Decision Principle

In my view, three of these basic model-specification principles are as follows.

• The **Modeling-As-Decision Principle** (preamble). Questions such as $Q_{MC_1}$ above seem basic, but are actually not: deeper question

\[
\{ \text{model comparison, iteration (ii)} \} \quad Q_{MC_2}: \text{Is model } M_2 \text{ better than } M_1, \text{ for the purpose to which the modeling will be put?}
\]

It’s easy to think of situations (e.g., should the Challenger space shuttle have been launched at 31°F?) in which

(a) only **crude modeling** is needed to obtain a **definitive** and retrospectively correct answer,

(b) **two models**, $M_1$ and $M_2$, are **available**, with $M_2$ **fitting** the data **much better** than $M_1$, and yet

(c) $M_1$ and $M_2$ are **equally good** for the **purpose** to which the **modeling** will be **put** (deciding whether to **launch** at 31°F).
Three Modeling Principles

This gives rise to

The **Modeling-As-Decision Principle** (statement): Making clear the purpose of the modeling 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;

- The **Calibration Principle**: In model specification, it helps to know something about how often {the methods You’re using to choose one model over another} get the right answer, and this can be ascertained by

  (a) creating simulation environments (structurally similar to the setup of the problem \( P \) You’re currently solving) in which You know what the right answer is, and

  (b) seeing how often Your methods recover known truth; and

- The **Prediction Principle**: Good models make good predictions, and bad models make bad predictions; that’s one important way You know that a model is good or bad.
Consider first the (simplest) one-sample setting, in which $D = y = (y_1 \ldots y_n)$ for real-valued $y_i$ and the models to be compared are

$$M_j : \left\{ \begin{array}{l} (\gamma_j | M_j B) \sim p(\gamma_j | M_j B) \\ (y | \gamma_j M_j B) \sim p(y | \gamma_j M_j B) \end{array} \right\} . \quad (5)$$

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 (see, e.g., O’Hagan and Forster 2004) 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,

$$LS_{CV}(M_j | y B) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | y_{-i} M_j B) , \quad (6)$$

in which $y_{-i}$ is the $y$ vector with observation $i$ omitted.
Somewhat surprisingly, Draper and Krnjajić (2014; cf. Laud and Ibrahim, 1995) have shown that a full-sample log score that omits the leave-one-out idea,

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

made operational with the rule \{favor $M_2$ over $M_1$ if $LS_{FS}(M_2|yB) > LS_{FS}(M_1|yB)$\}, can have better small-sample model discrimination ability than $LS_{CV}$.

$LS_{FS}$ looks like it uses the data twice, but any such effect turns out to be negligible for even moderate $n$.

Utility justification for log scores. I assume now that the central tasks in $P$ do not include decision-making, so that Your model $M$ reduces to \{$p(\theta|B), p(D|\theta B)$\}.

For simplicity of exposition, let’s continue to consider the one-sample setting with no covariates, in which (i) $D = y = (y_1, \ldots, y_n)$ for $y_i \in \mathbb{R}$ and (ii) $y^*$ is a future $y$ value (generalizations are straightforward).
Before the data set $y$ arrives, your uncertainty about the $y_i$ is exchangeable (this is part of $B$), so by de Finetti’s Representation Theorem for continuous outcomes, the only models with non-zero prior model probability can be expressed (for $i = 1, \ldots, n$) as

$$
(F|B) \sim p(F|B)
$$

$$
(y_i|F B) \overset{\text{IID}}{\sim} F,
$$

in which $F$ is a continuous CDF on $\mathbb{R}$.

Without loss of generality (in the sense that the resulting posterior distributions are dense in the set $\mathcal{F}$ of all CDFs on $\mathbb{R}$), model (8) may be specialized to

$$
(F|\alpha_0 F_0 B) \sim DP(\alpha_0, F_0)
$$

$$
(y_i|F B) \overset{\text{IID}}{\sim} F,
$$

in which $DP(\alpha_0, F_0)$ is the Dirichlet-process (DP) prior with concentration parameter $\alpha_0 \geq 0$ and prior estimate $F_0$ of $F$.

By the usual DP conjugate updating, the posterior on $F$ (given $y$ and $B$) induced by (9) is
Steps (1) and (2) in the Argument

\[ (F|y B) \sim DP(\alpha^*, F^*), \]  

(10)

where \( \alpha^* = (\alpha_0 + n) \) and \( F^* = \frac{\alpha_0 F_0 + n \hat{F}_n}{\alpha_0 + n} \); here \( \hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq t) \) is the empirical CDF based on \( y \) and \( I(A) \) is 1 if proposition \( A \) is true and 0 otherwise.

Thus the posterior expectation of \( F \) (given \( y \) and \( B \)) is

\[ E(F|y B) = F^*, \]  

which reduces to \( E(F|y B) = \hat{F}_n \) when \( \alpha_0 \downarrow 0 \).

The utility-justification argument for log scores proceeds in the following seven steps.

1. Under the **Calibration Principle**, it's sensible to speak of an underlying data-generating model \( M_{DG} \), which corresponds in model (9) to a point-mass DP prior \((\alpha_0 \rightarrow \infty)\) on \( F \) at some CDF \( F_{DG} \); in other words, in this context, \( M_{DG} \equiv F_{DG} \) (a simple example would be \( M_{DG} : (y_i|B) \overset{\text{IID}}{\sim} N(0,1) \)).

2. Under the **Modeling-As-Decision Principle**, Your job in choosing between two models \( M_1 \) and \( M_2 \) is to formulate this model comparison as a decision problem, as follows.
• **In the setting** of equation (9), choosing a model corresponds to **specifying** \((\alpha_0, F_0)\), so the **action space** \((A|B)\) in this **subsidiary decision problem** consists of all possible choices of \((\alpha_0, F_0)\) for \(\alpha_0 \in \mathbb{R}^+ \) and \(F_0 \in \mathcal{F}\) (this **includes hierarchical specifications** such as \((F_0|\mu \sigma B) \sim N(\mu, \sigma^2)\) with a **prior** on \((\mu, \sigma))\).

• The **uncertain quantity** \(\theta\) in this **decision problem** is \(M_{DG} = F_{DG}\), so let the set \(\Theta\) of **possible values** of \(\theta\) be \(\Theta = \mathcal{F}\).

• The **utility function** in general decision problems has the form \(U(a, \theta|B)\); here, in this **subsidiary decision problem**, it **suffices** (for reasons that **will become clear** below) to define it **only** for \(\theta = M_{DG} = F_{DG}\), as \(U(M, F_{DG}|B)\), where \(M\) is a **particular choice** of \((\alpha_0, F_0)\).

• In the **maximization** of expected utility in this **subsidiary decision problem**, the **expectation** is over the **posterior distribution** \(p(M_{DG}|y B)\) for the **unknown** \(\theta = M_{DG} = F_{DG}\), **given** the data set \(y\) and the **background information** \(B\).

This **means** that, in this **context**, \(p(M_{DG}|y B) = p(F_{DG}|y B)\), which (as **noted** above) is the **DP(\alpha^*, F^*)** distribution.
Steps (3)–(5) in the Argument

(3) Each choice of a model $M$ induces a predictive distribution $p_M(y^*|y MB)$ for a new data value $y^*$; the corresponding predictive distribution under $M_{DG}$ is

\[ p_{M_{DG}}(y^*|y M_{DG} \mathcal{B}) = p(y^*|y F_{DG} \mathcal{B}) = p(y^*|F_{DG} \mathcal{B}), \]

which is just the sampling distribution under $F_{DG}$.

(4) Let the CDF corresponding to the predictive density

\[ p_{M_{DG}}(y^*|y M_{DG} \mathcal{B}) = p(y^*|F_{DG} \mathcal{B}) \]

be $F_{DG}(y^*)$ (suppressing the dependence on $\mathcal{B}$ for notational simplicity); then an integral such as

\[ \int_{\mathbb{R}} p(y^*|F_{DG} \mathcal{B}) \log p_M(y^*|y MB) \, dy^* \quad (11) \]

can equally well be expressed as \[ \int_{\mathbb{R}} \log p_M(y^*|y MB) \, dF_{DG}(y^*). \]

(5) Motivated by the Prediction Principle, now define

\[ U(M, F_{DG}|\mathcal{B}) = \int_{\mathbb{R}} \log p_M(y^*|y MB) \, dF_{DG}(y^*) - \int_{\mathbb{R}} \log p(y^*|F_{DG} \mathcal{B}) \, dF_{DG}(y^*) \quad (12) \]
\[
U(M, F_{DG}|B) \equiv \int_{\mathbb{R}} \log p_M(y^* | y M B) \, dF_{DG}(y^*) - \int_{\mathbb{R}} \log p(y^* | F_{DG} B) \, dF_{DG}(y^*)
\]

\[
= - \left[ \int_{\mathbb{R}} p(y^* | F_{DG} B) \log p(y^* | F_{DG} B) \, dy^* - \int_{\mathbb{R}} p(y^* | F_{DG} B) \log p_M(y^* | y M B) \, dy^* \right]
\]

\[
= -KL[p_M(y^* | y M B) \parallel p(y^* | F_{DG} B)] ; \quad (13)
\]

in other words, \(U(M, F_{DG}|B)\) is minus the Kullback-Leibler divergence of \{the predictive distribution for a new data value \(y^*\) under \(M\)\} from \{the corresponding predictive (sampling) distribution under \(F_{DG}\)\}.

\(\text{Now, recalling from above that } p(F_{DG}|y B) \text{ is the } DP(\alpha^*, F^*) \text{ distribution, it follows that } \alpha_0 \downarrow 0, E(F_{DG}|y B) = \hat{F}_n.\)

Thus, by Fubini’s theorem, for \(\alpha_0 \downarrow 0\), the expected utility is
Expected Utility = Difference of Log Scores

\[
E_{(F_{DG} | y B)} U(M, F_{DG} | B) = E_{(F_{DG} | y B)} \int_{\mathcal{R}} \log p_M(y^* | y M B) \, dF_{DG}(y^*) - \\
E_{(F_{DG} | y B)} \int_{\mathcal{R}} \log p(y^* | F_{DG} B) \, dF_{DG}(y^*) \\
= \int_{\mathcal{R}} E_{(F_{DG} | y B)} [\log p_M(y^* | y M B) \, dF_{DG}(y^*)] - \\
\int_{\mathcal{R}} E_{(F_{DG} | y B)} [\log p(y^* | F_{DG} B) \, dF_{DG}(y^*)] \\
= \int_{\mathcal{R}} \log p_M(y^* | y M B) \, d\hat{F}_n(y^*) - \\
\int_{\mathcal{R}} \log p(y^* | F_{DG} B) \, d\hat{F}_n(y^*) \\
= \frac{1}{n} \sum_{i=1}^{n} \log p_M(y_i | y M B) - \\
\frac{1}{n} \sum_{i=1}^{n} \log p_{M_{DG}}(y_i | y M_{DG} B) \\
\equiv LS_{FS}(M | y B) - LS_{FS}(M_{DG} | y B). \tag{14}
\]
Therefore, with this utility function in the subsidiary decision problem, model $M_2$ will maximize expected utility (when compared with model $M_1$) iff

$$\left[ LS_{FS}(M_2|yB) - LS_{FS}(M_{DG}|yB) \right] > \left[ LS_{FS}(M_1|yB) - LS_{FS}(M_{DG}|yB) \right];$$

in other words, iff

$$LS_{FS}(M_2|yB) > LS_{FS}(M_1|yB).$$

Thus the model-comparison rule

\{find the model with the largest full-sample log score\}

has a well-grounded basis in Bayesian model specification, as the solution to \{the model-comparison problem, when viewed as a subsidiary decision problem with a utility function that rewards predictive accuracy\}.

Now that log scores and Bayes factors are both Bayesian, how do they compare in their ability to correctly discriminate between models?
Strengths and weaknesses of Bayes factors and log scores.

Each of these approaches to answering the question

\[ Q_1 : \text{Is } M_1 \text{ better than } M_2? \]

has its advocates (Bayes factors: Berger, Pericchi, Bayarri, ...); log scores: Gelfand & Ghosh, Laud & Ibrahim, Draper, ...).

• A brief review of Bayes factors. It looks natural to compare models on the basis of their posterior probabilities; from Bayes’s Theorem in odds form,

\[
\frac{p(M_2|DB)}{p(M_1|DB)} = \left[ \frac{p(M_2|B)}{p(M_1|B)} \right] \cdot \left[ \frac{p(D|M_2B)}{p(D|M_1B)} \right]; \quad (17)
\]

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 the Bayes factor, so in plain language equation (17) says
Bayes Factors (continued)

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

(18)

(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 \text{ over } M_1 \text{ if the posterior odds for } M_2 \text{ are greater} \}

and

\{choose } M_2 \text{ over } M_1 \text{ if } p(M_2|D,B) > p(M_1|D,B) \}

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

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

(19)
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 $M_1$,

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

favoring $M_2$ if $BF(M_2 \text{ over } M_1|D, B) > 1$, 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

$$M_j: \begin{cases} 
(\gamma_j | M_j B) \sim p(\gamma_j | M_j B) \\
(D | \gamma_j M_j B) \sim p(D | \gamma_j M_j B)
\end{cases}. \quad (21)$$

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

$$p(D | M_j B) = \int p(D | \gamma_j M_j B) p(\gamma_j | M_j B) \, d\gamma_j; \quad (22)$$

this is called an **integrated likelihood** (or **marginal likelihood**) because it tells you to take a weighted average of the sampling distribution/likelihood $p(D | \gamma_j M_j B)$, 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 (22) can be **difficult to compute**, and may not even be easy to **approximate**.

The second thing to **notice** is that (22) can be **rewritten** as...
The Darkness of the Cloud Becomes Apparent

\[ p(D|M_j B) = E_{\gamma_j|M_j B} p(D|\gamma_j M_j B) . \] (23)

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 data set} \( D \)).

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

\textbf{Example:} Integer-valued data set \( D = (y_1 \ldots y_n) \); \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \);

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

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

\[ \frac{\Gamma(\alpha_1 + \beta_1) \Gamma(n + \alpha_1) \Gamma(n\bar{y} + \beta_1) \Gamma(\alpha_2) (n + \beta_2)^{n\bar{y}+\alpha_2} (\prod_{i=1}^{n} y_i !)}{\Gamma(\alpha_1) \Gamma(\beta_1) \Gamma(n + n\bar{y} + \alpha_1 + \beta_1) \Gamma(n\bar{y} + \alpha_2) \beta_2^{\alpha_2}} . \] (24)
Instability of Bayes Factors to Prior Specification

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

\[
\frac{\Gamma(n + 1) \Gamma(n \bar{y} + 1) \Gamma(\epsilon) (n + \epsilon)^{n \bar{y} + \epsilon} (\prod_{i=1}^{n} y_i !) \Gamma(n + n \bar{y} + 2) \Gamma(n \bar{y} + \epsilon) \epsilon^\epsilon}{\Gamma(n + n \bar{y} + 2) \Gamma(n \bar{y} + \epsilon) \epsilon^\epsilon}.
\] (25)

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 \{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

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

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

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

(27)

where $\hat{\gamma}_j$ 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 $\log p(\hat{\gamma}_j|M_j B)$, and this is a term that won’t go away with more data: as $n$ increases this term is $O(1)$. 

--

Laplace Approximation (continued)
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):

$$\log p(y|M_j B) = \log p(y|\hat{\gamma}_j M_j B) - \frac{k_j}{2} \log n + O(1).$$ \hfill (28)

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

$$BIC(M_j|D B) = -2 \log p(D|\hat{\gamma}_j M_j B) + k_j \log n$$ \hfill (29)

(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 $-2 \log p(D|\hat{\gamma}_j M_j B)$ term).

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

$$AIC(M_j|D B) = -2 \log p(D|\hat{\gamma}_j M_j B) + 2 k_j.$$ \hfill (30)
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

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

(\textbf{note that this only makes sense after transforming all the components of } \gamma_j \textbf{ to live on the entire real line}).

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.

OK, so now we have two Bayesian ways to compare models — Bayes factors and log scores — each supported by people who (by and large) have acted toward each other like warring factions.

I will now argue that neither approach dominates the other, which leads me to propose a peace treaty based on the recommendation

\{use each method when its strengths outweigh those of the other method\};

along the way in this argument, I’ll articulate the final Principle for Bayesian modeling in this talk.
• **Case 1:** $M_1$ and $M_2$ are both parametric, and the dimensions of their parameter spaces are the same.

**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 $(SBP_{before} - SBP_{after})$ in the population of patients to which it’s appropriate to generalize from the patients in Your trial, and let $D = y = (y_1 \ldots y_n)$,
The Decision-Versus-Inference Principle

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**: It’s good to 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 \((A|B) = (a_1, a_2) = (\text{don’t take the drug forward to phase III, do take it forward})\), and a sensible utility function \( U(a_j, \theta|B) \) 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 **60 mmHg**, beyond which the drug starts inducing **fainting spells**).

However, to **facilitate** a **comparison** between Bayes factors (and their special case BIC (Schwarz, 1978)) 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:

---

**Practical Significance Improvement Threshold**

---
\[
M_1: \left\{ \begin{array}{l}
(\theta | B) \sim \text{diffuse for } \theta \leq \Delta \\
(y_i | \theta B) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)
\end{array} \right\} \quad \text{and} \quad (32)
\]

\[
M_2: \left\{ \begin{array}{l}
(\theta | B) \sim \text{diffuse for } \theta > \Delta \\
(y_i | \theta B) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}, \quad (33)
\]

in which \textbf{for simplicity} I’ll take \( \sigma \) to be \textbf{known} (the \textbf{results} are \textbf{similar} with \( \sigma \) \textbf{learned} from the \textbf{data}).

This gives rise to \textbf{three model-selection methods} that can be \textbf{compared calibratively}:

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

- \textbf{Posterior probability:} let

\[
M^*: \left\{ \begin{array}{l}
(\theta | B) \sim \text{diffuse on } \mathbb{R} \\
(y_i | \theta B) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}, \quad (34)
\]

and choose \( M_2 \) if \( p(\theta > \Delta | y M^* B) > 0.5 \).

- \textbf{BIC:} choose \( M_2 \) if \( BIC(M_2 | y B) < BIC(M_1 | y B) \).
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 \( LS_{FS} \) 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 $\theta_{DG} = 15$. 
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

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

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 $|\theta| \leq \lambda$, 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

$$M_3: \left\{ \begin{array}{l}
(\theta|B) \sim \text{diffuse for } |\theta| \leq \lambda \\
y_i|\theta B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}$$

and

$$M_4: \left\{ \begin{array}{l}
(\theta|B) \sim \text{diffuse for } |\theta| > \lambda \\
y_i|\theta B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}, \quad (36)$$

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

A **natural alternative** to **BIC** and **LSFS** 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, 
\textit{deciding} that they were not bio-equivalent when they were, 
\textit{judging} that they were not bio-equivalent when they were not\},

but here I’ll again simply \textbf{compare} the calibrative performance of 
\textit{LS\textsubscript{FS}}, \textit{posterior probabilities}, and \textit{BIC}.

**Simulation experiment details**, based on the SBP drug trial: \( \lambda = 5; \)
\( \sigma = 10; \ n = 10, 20, \ldots, 100; \) \textbf{data-generating}
\( \theta_{DG} = \{-9, -7, -5, -3, -1, 0, 1, 3, 5, 7, 9\}; \) \( \alpha = 0.05; \) \textbf{1,000 simulation}
replications, \( M = 10,000 \) Monte-Carlo draws for \textit{LS\textsubscript{FS}}.
In this 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 ±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.
In the context of the quantifying-improvement example, the real-world purpose of the experiment was to decide whether or not to take the drug forward to phase III.

Suppose that You tried to solve this decision problem with a popular inferential tool: frequentist hypothesis-testing of $H_0: \theta \leq \Delta$ versus $H_A: \theta > \Delta$ at significance level $\alpha$.

Decision-theoretically this is already wrong; as noted back on page 83, the utility function should actually be continuous in $\theta$ rather than artificially dichotomizing $\Theta$ into $(-\infty, \Delta]$ and $(\Delta, \infty)$.

Even if You temporarily buy into this incorrect dichotomization, to solve the problem properly You’d 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>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$ (stop)</td>
<td>$\theta \leq \Delta$</td>
</tr>
<tr>
<td></td>
<td>$u_{11}$</td>
</tr>
<tr>
<td>$a_2$ (phase III)</td>
<td>$-u_{21}$</td>
</tr>
</tbody>
</table>
Decision-Theory (Not Inference) For Decision Problems

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

- $u_{11}$ is the **gain** from correctly not taking the drug forward to phase III;
- $u_{12}$ 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

$$P(\theta > \Delta | y B) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^*. \quad (38)$$

The **frequentist (hypothesis-testing) inferential approach** is equivalent to this only if
\[ \alpha = 1 - u^* = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}}. \]  

The implicit trade-off between false positives and false 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 \( \{u_{11}, u_{12}, u_{21}, u_{22}\} \).

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 \( u_{12} \gg u_{21} \) 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 \( LS_{FS} \).)
Case 2: Comparing Models of Different Dimensions

- **Case 2:** $M_1$ and $M_2$ are both parametric, but the dimension of the parameter space in $M_2$ is greater than that in $M_1$.

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 (back in the days before genome sequencing) 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 ($\theta > 0$)\}.

**The Structural Singleton Principle.** Comparing a model defined by $\theta = \theta_0$ with one defined by $\theta \neq \theta_0$ — which is equivalent to testing the sharp-null hypothesis $H_0: \theta = \theta_0$ — in settings without a structural singleton at $\theta_0$ is always unwise.

This is because

(a) You already know from scientific context, when the outcome variable is continuous, that $H_0$ is false, and (relatedly)
Comparing Models of Different Dimensions (continued)

(b) it’s silly from a measurement point of view: with a (conditionally) IID $N(\theta, \sigma^2)$ sample $y$ 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 its advertised weight.

In a setting in which $\theta = 0$ is a structural singleton, here are some results: here I’m comparing the models $(i = 1, \ldots, n)$

\[
M_5: \left\{ \begin{array}{l}
(\sigma | \mathcal{B}) \sim \text{diffuse on } (0, \text{large}) \\
(y_i | \sigma \mathcal{B}) \overset{\text{IID}}{\sim} N(0, \sigma^2)
\end{array} \right\} \quad \text{and} \quad (40)
\]

\[
M_6: \left\{ \begin{array}{l}
(\theta \sigma | \mathcal{B}) \sim \text{diffuse on } (-\text{large, large}) \times (0, \text{large}) \\
(y_i | \theta \sigma \mathcal{B}) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}, \quad (41)
\]

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.
Testing Sharp-Null Hypotheses (continued)

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 this structural-singleton setting, 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.
Interval ($\alpha$ Modified to $LS_{FS}$ Behavior) Results

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

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

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

$LS_{FS}$ and $BIC$ both have **closed-form expressions** in this situation: with $s = \sum_{i=1}^n y_i$ and $\hat{\theta}_1 = \frac{\alpha_1 + n}{\alpha_1 + \beta_1 + s + n}$,

$$LS_{FS}(M_7 | y B) = \log \Gamma(\alpha_1 + n + \beta_1 + s) + \log \Gamma(\alpha_1 + n + 1)$$
$$- \log \Gamma(\alpha_1 + n) - \log \Gamma(\beta_1 + s)$$
$$+ \frac{1}{n} \sum_{i=1}^n [\log \Gamma(\beta_1 + s + y_i)]$$
$$- \log \Gamma(\alpha_1 + n + \beta_1 + s + y_i + 1), \quad (42)$$

$$BIC(M_7 | y B) = -2[n \log \hat{\theta}_1 + s \log(1 - \hat{\theta}_1)] + \log n, \quad (43)$$
\[ LS_{FS}(M_8|y B) = (\alpha_2 + s) \log(\beta_2 + n) - \log \Gamma(\alpha_2 + s) \]
\[ - (\alpha_2 + s) \log(\beta_2 + n + 1) \]
\[ + \frac{1}{n} \sum_{i=1}^{n} [\log \Gamma(\alpha_2 + s + y_i) - y_i \log(\beta_2 + n + 1) - \log \Gamma(y_i + 1)] , \text{ and} \]
\[ BIC(M_8|y B) = -2[s \log \hat{\theta}_2 - n \hat{\theta}_2 - \sum_{i=1}^{n} \log(y_i!)] + \log n , \quad (45) \]

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 $LS_{FS}$).
**Q₁**: Is $M_1$ better than $M_2$?

As before, let’s agree to call **choosing $M_2$ when the structure of $M_1$ is correct** a false-positive error, and **choosing $M_1$ when the structure of $M_2$ is correct** a false-negative mistake.

It turns out that the log-score approach has model-discrimination characteristics similar to those of the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), but log scores avoid the DIC drawback of obtaining (sharply) different estimates of model complexity as a function of the parameterization used to define the deviance.

- **Case 1**: $M_1$ and $M_2$ are both parametric, and the dimensions of their parameter spaces are the same.

In this case, {Bayes factors/BIC} and {log scores/DIC} will often have similar false-positive and false-negative error rates; when they differ (e.g., with small samples), neither uniformly dominates, because lower false-positive rates are always accompanied by higher false-negative rates.
\[ Q_1: \text{Is } M_1 \text{ better than } M_2? \]

- **Case 2:** \(M_1\) and \(M_2\) are both parametric, but the dimension of the parameter space in \(M_2\) is greater than that in \(M_1\).

**Canonical example** \((i = 1, \ldots, n)\):

\[
M_5: \left\{ \begin{array}{c}
(\sigma | B) \sim \text{diffuse on } (0, \text{large}) \\
(y_i | \sigma, B) \overset{\text{iid}}{\sim} N(0, \sigma^2)
\end{array} \right\} \quad \text{and} \quad (46)
\]

\[
M_6: \left\{ \begin{array}{c}
(\theta, \sigma | B) \sim \text{diffuse on } (-\text{large}, \text{large}) \times (0, \text{large}) \\
(y_i | \theta, \sigma, B) \overset{\text{iid}}{\sim} N(\theta, \sigma^2)
\end{array} \right\}, \quad (47)
\]

In this setting, advocates of Bayes factors often point out the following **consistency results**: as \(n \to \infty\) with the models under consideration fixed at \(M_5\) and \(M_6\),

- \(P_{RS}(\text{Bayes factors choose } M_6|M_6 \text{ correct}) \to 1\)
- \(P_{RS}(\text{Bayes factors choose } M_5|M_5 \text{ correct}) \to 1\)
As \( n \to \infty \) with the models under consideration fixed at \( M_5 \) and \( M_6 \),

- \( P_{RS}(\text{Bayes factors choose } M_6 | M_6 \text{ correct}) \to 1 \)
- \( P_{RS}(\text{Bayes factors choose } M_5 | M_5 \text{ correct}) \to 1 \)
- \( P_{RS}(\text{log scores choose } M_6 | M_6 \text{ correct}) \to 1 \)
- \( P_{RS}(\text{log scores choose } M_5 | M_5 \text{ correct}) \to 0 \)

(We already saw this in the graph on page 105.)

This is correct (it’s a valid theorem), but for me it’s not a relevant theorem, for the following reasons:

- The asymptotics are unrealistic: as \( n \) grows, to better model the complexity of the real world, the models under comparison don’t stay fixed (they increase in complexity).

- Data-gathering often unfolds over time, in which case as \( n \) grows the IID assumption in \( M_5 \) and \( M_6 \) becomes less plausible, as the stationarity of the process You’re studying comes increasingly into question.
• Most importantly, when \( n = 71 \) in my problem, I don’t care what happens for \( n = \infty \): I want to know about the false-positive/false-negative tradeoffs of various model comparison methods with \( n = 71 \), and consistency tells me precisely nothing about that.

The right way to answer this question is either with closed-form calculations (if possible) or with simulation:

(1) **Hold** the **structure** of the **problem** and the **sample size fixed** to match the **real problem**, with **known data-generating values** of the parameters (similar to parameter estimates based on Your data), and evaluate the **false-positive** and **false-negative error rates** of the competing model-comparison methods (**no method will uniformly dominate**, for the reasons given above);

(2) **Think about** the **real-world consequences** of **false-positive** and **false-negative errors**; and

(3) **Choose** the **model-comparison method** with the **best performance** on the **type of error** that’s more important.
As a general rule in Case 2, Bayes factors were designed for consistency, so they tend to make more false-negative errors than log scores; and log scores were designed to make good predictions, so they make more false-positive errors than Bayes factors.

(Actually, by the Modeling-As-Decision Principle, 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, but the 3–step process on the previous page will often be a good approximation to the decision-theoretic solution.)

Peace treaty proposal: Advocates of \{Bayes factors/BIC\} and \{log scores/DIC\} should shake hands on the true proposition that neither approach uniformly dominates: for any $n < \infty$, both approaches make both false-positive and false-negative errors, and there’s no model-comparison method that simultaneously minimizes both error rates for fixed $n$; therefore, everybody should become well acquainted with both approaches, and use them flexibly according to the real-world severity of the two kinds of errors they make.
Examples of the real-world implications of false-positive and false-negative errors:

• In the structural-singleton genetic linkage example (back on page 100), from the point of view of scientific inference it’s arguably worse to {declare linkage between two genes when none exists} (a false-positive mistake) than to {fail to declare linkage when it’s present} (a false-negative error; cf. the usual Neyman-Pearson type I/type II argument), so Bayes factors would be better in this instance than log scores from an inferential scientific perspective.

• Variable selection in searching through many compounds or genes to find successful treatments to be developed by a drug company: 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 $\kappa C$ with $\kappa = 10–100$: log scores would be better here.
Lest You think that Bayes factors are always better for scientific inference and log scores are always superior for decision-making:

- In a two-arm clinical-trial setting (such as the IHGA case study), consider again the mixed-effects Poisson regression model $M_2$:

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

$$ \log \lambda_i = \beta_0 + \beta_1 x_i + e_i \quad (48) $$

$$ (e_i | \sigma_e \mathcal{B}) \overset{\text{IID}}{\sim} N(0, \sigma_e^2), \quad (\beta_0 \beta_1 \sigma_e | \mathcal{B}) \sim \text{diffuse}, $$

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 $M_1$, obtained by setting $\sigma_e = 0$, to describe unexplainable heterogeneity.

In this setting, Bayes factors will make the mistake of \{telling You that $\sigma_e = 0$ when it’s not\} more often than log scores, and log scores will make the error of \{telling You that $\sigma_e > 0$ 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.
Outline

(1) Log Scores for Model Comparison

(2) A Bayesian non-parametric look at the frequentist bootstrap
Case Study 1. (Krnjajić, Kottas, Draper 2008): In-home geriatric assessment (IHGA). In an clinical trial conducted in the 1980s (Hendriksen et al., 1984), 572 elderly people, representative of $\mathcal{P} = \{\text{all non-institutionalized elderly people in Denmark}\}$, 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>$n_C0$ $n_C1$ $\ldots$ $n_Cm$</td>
<td>$n_C = 287$</td>
<td>$\bar{y}_C$</td>
<td>$s_C$</td>
</tr>
<tr>
<td>Treatment</td>
<td>$n_T0$ $n_T1$ $\ldots$ $n_Tm$</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 $\mathcal{P}$ under the $C$ and $T$ conditions, respectively.

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

\[ Q_1: \] Was the mean number of hospitalizations per two years in the IHGA group different from that in control by an amount that was large in practical terms? [description involving \( \bar{y}_T - \bar{y}_C \)]

\[ Q_2: \] Did IHGA (causally) change the mean number of hospitalizations per two years by an amount that was large in statistical terms? [inference about \( \frac{\mu_T - \mu_C}{\mu_C} \)]

\[ Q_3: \] 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]

\[ Q_4: \] 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), ...).
**Definition.** In model specification, optimal = \{conditioning only on propositions rendered true by the context of the problem and the design of the data-gathering process, while at the same time ensuring that Your set \( B \) of conditioning propositions includes all relevant problem context\}.

**Q:** Can this **optimality goal** be achieved? **A:** Yes, sometimes.

**Example:** Optimal Analysis (1) of IHGA clinical trial:

<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_{Cm} )</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_{Tm} )</td>
<td>( n_T = 285 )</td>
<td>( \bar{y}_T )</td>
<td>( s_T )</td>
</tr>
</tbody>
</table>

**Before the data set arrives,** Your **uncertainty** about the control-group data values \( \{ C_i = \text{number of hospitalizations for control patient } i \} \) is **exchangeable**, meaning that Your predictive distribution \( p(C_1, \ldots, C_{n_C} | B) \) is the same no matter what order the \( C_i \) values are written down in.
Similarly, before the data set arrives, your uncertainty about the treatment-group data values \( \{T_j = \text{number of hospitalizations for treatment patient } j\} \) is also exchangeable.

These exchangeability judgments arise directly from problem context and are therefore part of \( \mathcal{B} \); in other words, basing the model on exchangeability is an example of optimal Bayesian model specification.

de Finetti (1937) proved a wonderful theorem with the following consequences in this clinical trial (and others like it).

Since the control patients were chosen to be representative of (like a random sample from)

\[
\mathcal{P}_C = \{\text{all elderly non-institutionalized Danish people in the early 1980s, receiving standard health care}\},
\]

and since the treatment patients were like a random sample from

\[
\mathcal{P}_T = \{\text{all elderly non-institutionalized Danish people in the early 1980s, receiving standard health care plus IHGA}\},
\]
and since there’s no logical or probabilistic linkage between the $C_i$ and $T_j$,

(a) it’s meaningful to think of $F_C$ and $F_T$ as the cumulative distribution functions (CDFs) of the control and treatment population hospitalization counts, and

(b) de Finetti’s theorem then says that the following model achieves optimal Bayesian model specification:

$$
(F_C | B) \sim DP[\alpha_C, F_{0C}] \quad \text{and} \quad (F_C | \alpha_T B) \sim DP[\alpha_T, F_{0T}]
$$

$$(C_i | F_C B) \overset{\text{IID}}{\sim} F_C \quad \text{and} \quad (T_j | F_T B) \overset{\text{IID}}{\sim} F_T,
$$

in which $DP(\alpha_C, F_{0C})$ is a member of the Dirichlet-Process class of Bayesian non-parametric priors on $F_C$, the set of all CDFs on $\mathbb{R}$, and similarly for $DP(\alpha_T, F_{0T})$.

Focusing for simplicity just on the control-group data and letting $C = (C_1, \ldots, C_{n_C})$, it’s a basic fact about DP priors that

$$
(F_C | C B) \sim DP \left( \alpha_C + n_C, \frac{\alpha_C F_{0C} + n_C \hat{F}_{n_C}}{\alpha_C + n_C} \right).
$$
where \( \hat{F}_{nc} \) is the empirical CDF of the control-group data values (and similarly for the treatment data).

With the \( DP(\alpha, F_0) \) prior, \( \alpha \) plays the role of the prior sample size and \( F_0 \) is the prior estimate of \( F \).

If (as is the case here) little is known about hospitalization rates for elderly non-institutionalized Danish people in the early 1980s with and without IHGA, this state of information can be captured with the choices \( (\alpha, F_0) = (0, \text{anything}) \), in which case the posterior distributions in control and treatment become

\[
(F_C | C B) \sim DP\left(n_C, \hat{F}_{nc}\right) \quad \text{and} \quad (F_T | T B) \sim DP\left(n_T, \hat{F}_{nt}\right). \quad (51)
\]

**Fact** (Draper, 2014). If \( \hat{F}_n \) is the empirical CDF based on \( y = (y_1, \ldots, y_n) \), then simulated draws from \( DP\left(n, \hat{F}_n\right) \) can be approximated to high accuracy, even with small \( n \), by making frequentist bootstrap draws from \( y \), and this analysis will be about 30 times faster than the conventional Bayesian Monte-Carlo method for DPs (the stick-breaking algorithm).
Thus a highly accurate, computationally fast, Monte-Carlo approximate optimal Bayesian analysis of this clinical trial is:

- Choose a large integer $M$ such as 100,000 or 1,000,000.

- For $m = 1, \ldots, M$,
  
  — draw $(C_1^*, \ldots, C_{n_C}^*)$ at random with replacement from $(C_1, \ldots, C_{n_C})$ and compute the mean $\bar{C}_m^*$ of these $C_i^*$ values;
  
  — draw $(T_1^*, \ldots, T_{n_T}^*)$ at random with replacement from $(T_1, \ldots, T_{n_T})$ and compute the mean $\bar{T}_m^*$ of these $T_j^*$ values;
  
  — compute $\theta_m^* = \frac{\bar{T}_m^* - \bar{C}_m^*}{\bar{C}_m^*}$; store this value at position $m$ in vector $\theta^*$.

- Draw a histogram or density trace of the $\theta^*$ values as Your approximate posterior distribution for $\theta = \frac{\mu_T - \mu_C}{\mu_C}$ given the data set $(C, T)$ and the background information $B$; calculate the mean and SD of the $\theta^*$ values as Your approximate posterior mean and SD for $\theta$ (respectively); compute the 2.5% and 97.5% quantiles of the distribution of the $\theta^*$ values as Your approximate 95% Bayesian interval estimate for $\theta$. 
This analysis plan should make everybody happy: it uses only the frequentist bootstrap to achieve a highly accurate approximate optimal Bayesian analysis (i.e., You frequentists out there can interpret the results in a Bayesian way, with direct probability statements), and with minimal computing time.

Optimal Analysis 2 (BQQI). Another approach to optimal Bayesian model specification in this clinical trial is provided by an approach that might be called Bayesian Qualitative/Quantitative Inference (BQQI).

Consider just the control group for a moment, and temporarily denote the data values $C_i$ in this group by $y = (y_1, \ldots, y_n)$.

Another of de Finetti’s Representation Theorems (generalizing the result for Bernoulli outcomes), not mentioned previously, permits a completely different analysis of the IHGA data, as follows.

• If the data vector $y = (y_1, \ldots, y_n)$ takes on $\ell$ distinct values $v = (v_1, \ldots, v_\ell)$ (real numbers or not) and I judge (my uncertainty about) the infinite sequence $(y_1, y_2, \ldots)$ to be exchangeable,
then a **desire for logical internal consistency compels** me

(i) to **think about** the quantities \( \phi = (\phi_1, \ldots, \phi_\ell) \), where \( \phi_j \) is the **limiting relative frequency** of the \( v_j \) **values** in the **infinite sequence**, and

(ii) to **adopt** the **Multinomial** model

\[
(\phi | B) \sim p(\phi | B)
\]

\[
p(y_i | \phi) = c \prod_{j=1}^{\ell} \phi_j^{s_j}, \quad (52)
\]

where \( s_j \) is the **number of** \( y_i \) **values equal** to \( v_j \);

- **If context suggests** a **diffuse** prior for \( \phi \) (as in the **IHGA case study**), a convenient (conjugate) choice is **Dirichlet** \( (\alpha) \) with \( \alpha = (\alpha_1, \ldots, \alpha_\ell) \) and **all** of the \( \alpha_j \) **positive but close to** 0; and

- with a **Dirichlet** \( (\alpha) \) **prior** for \( \phi \), the **posterior** is **Dirichlet** \( (\alpha') \), where \( s = (s_1, \ldots, s_\ell) \) and \( \alpha' = (\alpha + s) \).
Note, remarkably, that the $v_j$ values themselves make no appearance in the model; this modeling approach is natural with qualitative outcomes but can also be used when the $v_j$ are real numbers.

For example, for real-valued $y_i$, if (as in the IHGA case study) interest focuses on the (underlying population) mean in the infinite sequence $(y_1, y_2, \ldots)$, this is $\mu_y = \sum_{j=1}^{\ell} \phi_j v_j$, which is just a linear function of the $\phi_j$ with known coefficients $v_j$.

In the IHGA two-independent-samples setting, I can apply de Finetti’s Representation Theorem twice, in parallel, on the $C$ and $T$ data values.

I don’t know much about the underlying frequencies of $\{0, 1, \ldots, 7\}$ hospitalizations under $C$ and $T$ external to the data, so I’ll use a Dirichlet$(\epsilon, \ldots, \epsilon)$ prior for both $\phi_C$ and $\phi_T$ with $\epsilon = 0.001$, leading to a Dirichlet$(138.001, \ldots, 2.001)$ posterior for $\phi_C$ and a Dirichlet$(147.001, \ldots, 0.001)$ posterior for $\theta_T$ (other small positive choices of $\epsilon$ yield similar results).
library(MCMCpack)
alpha.C <- c(138.001, 77.001, 46.001, 12.001, 8.001, 4.001, 0.001, 2.001)
alpha.T <- c(147.001, 83.001, 37.001, 13.001, 3.001, 1.001, 1.001, 0.001)
set.seed(3141593)
phi.C.star <- rdirichlet(100000, alpha.C)
phi.T.star <- rdirichlet(100000, alpha.T)
mean.effect.C.star <- phi.C.star %*% (0:7)
mean.effect.T.star <- phi.T.star %*% (0:7)
theta.star <- (mean.effect.T.star - mean.effect.C.star) / mean.effect.C.star
print(posterior.mean.theta <- mean(theta.star))
# [1] -0.1809106
print(posterior.sd.theta <- sd(theta.star))
# [1] 0.08959087
quantile(theta.star, probs = c(0.0, 0.025, 0.5, 0.95,
                              0.975, 1.0))
  # 0% 2.5% 50% 95% 97.5% 100%
  # -0.495724757 -0.344056588 -0.185267638 -0.026189168 0.007791367 0.362005284
print( posterior.probability.ihga.beneficial <-
  mean( theta.star < 0 ) )
# [1] 0.97038
<table>
<thead>
<tr>
<th>Analysis</th>
<th>theta Posterior Mean</th>
<th>SD</th>
<th>Posterior Probability IHGA beneficial (theta &lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Non-parametric</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Frequentist Bootstrap]</td>
<td>-0.177</td>
<td>0.0891</td>
<td>0.963</td>
</tr>
<tr>
<td>2 BQQI [Bayesian Bootstrap]</td>
<td>-0.181</td>
<td>0.0896</td>
<td>0.970</td>
</tr>
</tbody>
</table>

The **Bayesian Qualitative/Quantitative Inferential (BQQI)** results, which are based on an instance of optimal model specification, coincide in this case with the more technically challenging Bayesian non-parametric analyses, and are achieved with no MCMC sampling and a computational clock time of less than 1 second.

The BQQI approach is an application of the **Bayesian bootstrap** (Rubin, 1981), which (for complete validity) includes the assumption that the observed $y_i$ values form an exhaustive set of \{all possible values the outcome $y$ could take on\}. 
Limits of Validity of BQQI

That assumption is met in the IHGA case study: possible data values of \{8, 9, \ldots \} can be added, each with Dirichlet prior weight of \( \epsilon \) and count 0, and the changes that result to the above analysis are negligible.

Caution: Not much is currently known about how well the BQQI approach works with conceptually continuous outcome variables; such outcomes are always discretized by the measuring process, so BQQI can technically always be applied, but — when there are many unattained discretized values between the attained values — it’s not yet clear what will happen in general.
• The **Modeling-As-Decision Principle** (page 9).

• The **Calibration Principle** (page 9).

• The **Prediction Principle** (page 9).

• **Full-sample log scores** \(LS_{FS}\), page 11 are a valid Bayesian way to compare models.

• **Bayes factors** (page 19) can be hideously sensitive to tiny details in the specification of diffuse priors on the parameters in the models being compared (page 23).

• When applicable, **BIC** (page 27) — which has built-in Unit-Information priors (page 28) — is a version of Bayes factors that often satisfactorily solves the sensitivity problem (but BIC is not applicable in, e.g., hierarchical models with random effects).

• The **Decision-Versus-Inference Principle** (page 31).

• The **Structural Singleton Principle** (page 49).
• Bayes factors do not uniformly dominate log scores in model discrimination ability, and log scores do not uniformly dominate Bayes factors: the two approaches just have different built-in false-positive and false-negative trade-offs (page 60).

• Therefore, instead of choosing one approach and heaping contempt upon the other, we should use whichever of the two methods performs better, on a problem-specific basis (page 64).

• With \{log scores, which are better than DIC, which is better than AIC\}, the goal is accurate out-of-sample prediction; to achieve this goal, these methods favor somewhat less parsimonious models.

• By contrast, with \{Bayes factors, BIC\}, the goal is consistency (page 61); to achieve this goal, these methods favor somewhat more parsimonious models.

• Optimal Bayesian model specification (new definition in the literature: page 70) is possible; Bayesian non-parametric (BNP) modeling can in some cases achieve this goal (page 75).
• The frequentist bootstrap accurately simulates draws from an important BNP posterior distribution — $DP(n, \hat{F}_n)$ — and does so about 30 times faster than the usual DP stick-breaking algorithm (page 73).

• Bayesian Qualitative/Quantitative Inference (BQQI; page 75), based on the Bayesian bootstrap, (a) can also achieve optimal Bayesian model specification and (b) is computationally extremely fast.