

Bayesian Model Specification

4: Dealing With Model Uncertainty

David Draper

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

and (1 Jul 2014–30 Sep 2015) *eBay Research Labs*

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

SHORT COURSE (DAY 3)
UNIVERSITY OF READING (UK)

27 Nov 2014

© 2014 David Draper (all rights reserved)

Getting From the Context and Design to the Model

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 the **set of conditioning propositions** includes **all relevant problem context**}.

This seems **hard to achieve**; for **example**, in the **IHGA case study**, **visualizing** the **data set before it arrives**, it would look like the **table shell** presented back on **page 2** of **Part 1** of the **Lecture Notes**:

Group	Number of Hospitalizations				n	Mean	SD
	0	1	...	k			
Control	n_{C0}	n_{C1}	...	n_{Ck}	$n_C = 287$	\bar{y}_C	s_C
Treatment	n_{T0}	n_{T1}	...	n_{Tk}	$n_T = 285$	\bar{y}_T	s_T

The **problem context** and **design** make this **table shell** something **You can condition on**, and the **lack of previous trials with IHGA** (this was the **first time** it was **implemented anywhere**) implies that **You can also condition** on a **diffuse choice** for $p(\theta|B)$ (with **572 observations**, it **won't matter much** how this **diffuseness** is **specified**), but **context** and **design** don't seem to have **anything to say** about the **predictive (sampling) distribution** $p(D|\theta B)$.

Model Uncertainty

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.****

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

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

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

An Ensemble \mathcal{M} of Models

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

Suppose **initially**, for the sake of **discussion**, that You’ve **identified** such an **ensemble** (I’ll present some **ideas** for how to do this later) and that it turns out to be **finite**: $\mathcal{M} = (M_1, \dots, M_k)$ for $2 \leq k < \infty$; **what next?**

Are You **supposed** to try to **choose** one of these **models** (the **model selection problem**) and **discard** the rest, or **combine** them in some way (if so, **how?**), or **what?**

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

Dealing With Model Uncertainty

(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 θ **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**.

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

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

I'm aware of **three approaches to improved assessment and propagation of model uncertainty**: **BMA, BNP, CCV**.

- **Bayesian model averaging (BMA)**: If **interest** focuses on **something** that has the **same meaning across all the models** in \mathcal{M} — for example, a set of **future data values** D^* to be **predicted** — **calculation** reveals (e.g., Leamer, 1978) that

$$p(D^*|D\mathcal{M}\mathcal{B}) = \int_{\mathcal{M}} p(D^*|D M \mathcal{B}) p(M|D\mathcal{M}\mathcal{B}) dM, \quad (1)$$

which is **eminently reasonable**: equation **(1)** tells You to form a **weighted average** of Your **conditional predictive distributions** $p(D^*|D M \mathcal{B})$, given particular **models** $M \in \mathcal{M}$, **weighted** by those models' **posterior probabilities** $p(M|D\mathcal{M}\mathcal{B})$.

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

- **Bayesian nonparametric (BNP) modeling**: The **BMA integral** in (1) can be thought of as an **approximation** to the **(unattainable?) ideal of averaging over all worthwhile models**; a **better approximation** to this **ideal** can often be achieved with **Bayesian nonparametric modeling**, which dates back to **de Finetti (1937)**.

Continuing the **Kaiser example** on page 14 (**Part 1**), suppose You also **observe** (for each of the $n = 112$ randomly-sampled patients from the **population** \mathbb{P} of $N = 8,561$ heart-attack patients) a **real-valued conceptually-continuous quality-of-care score** y_i , and (following **de Finetti**) You're thinking about Your **predictive distribution** $p(y_1 \dots y_n | \mathcal{B})$ for these scores **before any data have arrived**.

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

He (and later **Diaconis/Freedman**) went on to **prove** that, if Your judgment of **exchangeability** extends from $(y_1 \dots y_n)$ to $(y_1 \dots y_N)$ (as it certainly **should** here, given the **random sampling**) and $N \gg n$ (as is **true** here), then all **logically-internally-consistent predictive distributions** can **approximately** be expressed **hierarchically** as follows:

Bayesian Nonparametric (BNP) Modeling

letting F stand for the **empirical CDF** of the **population values** $(y_1 \dots y_N)$, the **hierarchical model** is (for $i = 1, \dots, n$)

$$\left\{ \begin{array}{l} (F|\mathcal{B}) \sim p(F|\mathcal{B}) \\ (y_i|F\mathcal{B}) \stackrel{\text{IID}}{\sim} F \end{array} \right\}.$$

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

IHGA Example, Revisited: Once again **visualizing** the **IHGA data set before it arrives**, here's the **table shell** one more time:

Group	Number of Hospitalizations				n	Mean	SD
	0	1	...	k			
Control	n_{C0}	n_{C1}	...	n_{Ck}	$n_C = 287$	\bar{y}_C	s_C
Treatment	n_{T0}	n_{T1}	...	n_{Tk}	$n_T = 285$	\bar{y}_T	s_T

BNP Case Study

Letting (as before) μ_C and μ_T be the **mean hospitalization rates** (per two years) in the **population \mathcal{P}** (of all elderly non-institutionalized people in Denmark in the early 1980s) under the C and T conditions, respectively, the **inferential quantity of main interest** is still $\theta = \frac{\mu_T - \mu_C}{\mu_C}$ (or this could be **redefined without loss** as $\theta = \frac{\mu_T}{\mu_C}$); how can You draw **valid and accurate inferences** about θ while **coping with Your uncertainty** about the **population C and T CDFs** — call them F_C and F_T , respectively — of **numbers of hospitalizations per person** (per two years)?

One approach: Bayesian qualitative-quantitative inference (Draper 2013): **exchangeability** implies a **multinomial sampling distribution** on the **qualitative outcome variable** with **category labels** $0, 1, \dots$, and this permits **optimal model specification** here (this approach treats the **hospitalization outcome categorically** but permits **quantitative inference** about θ).

Another approach: Bayesian nonparametric modeling — it turns out that **DP priors** put **all their mass** on **discrete distributions**, so **one BNP model** for this data set would involve placing **parallel DPs priors** on F_C and F_T ; see **KKD (2008)** for details on the **results**.

BNP Case Study (continued)

To serve as the **basis** of the M^* (**cheating**) **approach** (in which You **look at the data** for **inspiration** on which models to fit), here's a **table** of the **actual data values**:

Group	Number of Hospitalizations								n	Mean	SD
	0	1	2	3	4	5	6	7			
Control	138	77	46	12	8	4	0	2	287	0.944	1.24
Treatment	147	83	37	13	3	1	1	0	285	0.768	1.01

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

It's **natural** to think **initially** of **parallel Poisson**(λ_C) and **Poisson**(λ_T) modeling (M_1), but there's **substantial over-dispersion**: the C and T **variance-to-mean ratios** are $\frac{1.24^2}{0.944} \doteq \mathbf{1.63}$ and $\frac{1.01^2}{0.768} \doteq \mathbf{1.33}$.

Bayesian Parametric Modeling

Unfortunately we have **no covariates** to help **explain** the **extra-Poisson variability**, and there's **little information external** to the **data set** about the **treatment effect**; this latter **state of knowledge** is expressed in **prior distributions** on **parameters** by making them **diffuse** (i.e., ensuring they have **large variability** to express **substantial uncertainty**).

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

$$\begin{aligned}(y_i | \lambda_i M_3 \mathcal{B}) &\stackrel{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \gamma_0 + \gamma_1 x_i + \epsilon_i \\ (\epsilon_i | \sigma_\epsilon^2 M_3 \mathcal{B}) &\stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2) \\ (\gamma_0 \gamma_1 \sigma_\epsilon^2 | M_3 \mathcal{B}) &\sim \text{diffuse.}\end{aligned}\tag{2}$$

In this model the **unknown** of **main policy interest** is

BNP Example

$\theta = \frac{\text{population } \bar{\tau}}{\text{population } \bar{c}} = e^{\gamma_1}$; the **other parameters** can be collected in a **vector** $\eta = (\gamma_0, \sigma_\epsilon^2)$; and the **random effects** ϵ_i can be thought of as **proxying** for the **combined main effect** $\sum_{j=2}^J \gamma_j (x_{ij} - \bar{x}_j)$ of all the **unobserved relevant covariates** (age, baseline health status, ...).

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

Gaussian assumption:

$$\begin{aligned} (y_i | \lambda_i, M_4, \mathcal{B}) &\stackrel{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \gamma_0 + \gamma_1 x_i + \epsilon_i \\ (\epsilon_i | F, M_4, \mathcal{B}) &\stackrel{\text{iID}}{\sim} F \\ (F | \alpha, \sigma_\epsilon^2, M_4, \mathcal{B}) &\sim DP(\alpha, F_0), \quad F_0 = N(0, \sigma_\epsilon^2) \\ (\gamma_0, \gamma_1, \sigma_\epsilon^2 | M_4, \mathcal{B}) &\sim \text{diffuse}; \quad (\alpha | M_4) \sim \text{small positive}. \end{aligned} \tag{3}$$

Dirichlet-Process Mixture Modeling

Many **Bayesian prior distributions** $p(\theta|M_j; \mathcal{B})$ have **two user-friendly inputs**: a **quantity** θ_0 that acts like a **prior estimate** of the **unknown** θ , and a **number** n_0 that **behaves like a prior sample size** (i.e., a **measure of how tightly the prior is concentrated** around θ_0); **DP priors are no exception to this pattern.**

In equation (3), $DP(\alpha, F_0)$ is a **Dirichlet-process prior** on F with **prior estimate** $F_0 = N(0, \sigma_c^2)$ and a **quantity** (α) that behaves something like a **prior sample size**; this is referred to as **Dirichlet-process mixture modeling**, because (3) is a **mixture model** — each **person** in the study has her/his **own** λ , drawn from F_C (control) or F_T (treatment) — in which **uncertainty** about F_C and F_T is **quantified** via a **DP**.

NB **Bayesian model averaging** (BMA) with a **finite set of models** can be regarded as a **crude approximation** to what **Bayesian nonparametric** (BNP) modeling is **trying** to do, namely **average over Your uncertainty in model space** to provide an **honest representation** of Your **overall uncertainty** that **doesn't condition on things You don't know are true.**

Cross-Validation

- **Calibration cross-validation (CCV)**: The way the **IHGA** example unfolded looks a **lot** like the M^* **approach** I **condemned** previously: I used the **entire data set** to suggest which models to **consider**.

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

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

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

— **2CV** (**two-fold cross-validation**) is **frequently used**: You (a) **partition** the data into **modeling** (M) and **validation** (V) **subsets**, (b) use M to explore a **variety of models** until You've found a **"good"** one M^* , and (c) see how well M^* **validates** in V (a **useful Bayesian way** to do this is to **use the data** in M)

Calibration Cross-Validation (CCV)

to construct **posterior predictive distributions** for **all of the data values** in V and see how the **latter compare** with the **former**).

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

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

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

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

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

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

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

The **goal** with this **method** is both

- (1) a **good answer**, to the **main scientific question**, that has **paid a reasonable price** for **model uncertainty** (the **inferential answer** is based only on $M + V$, making Your **uncertainty bands wider**) and
- (2) an **indication** of how **well calibrated** {the **iterative fitting process** yielding the **answer** in (1)} is in C (a **good proxy** for **future data**).

You can use **decision theory** (Draper and Southwood, 2013) 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**).

Modeling Algorithm

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_{current} is good enough to stop (how decide?), return $\mathcal{M}_{\text{current}}$; else
- (c) Generate a new candidate model M_{new} (how choose?) and set $\mathcal{M}_{\text{current}} \leftarrow \mathcal{M}_{\text{current}} \cup M_{\text{new}}$.
- (d) If M_{new} is better than M_{current} (how decide?), set $M_{\text{current}} \leftarrow M_{\text{new}}$.
- (e) Go to (b).

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

Q_1 : Is M_1 **better** than M_2 ?

Q_2 : Is M_1 **good enough**?

The Modeling-As-Decision Principle

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

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

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

Two such **methods** are **Bayes factors** and **log scores**.

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

$$\frac{p(M_2|D\mathcal{B})}{p(M_1|D\mathcal{B})} = \left[\frac{p(M_2|\mathcal{B})}{p(M_1|\mathcal{B})} \right] \cdot \left[\frac{p(D|M_2\mathcal{B})}{p(D|M_1\mathcal{B})} \right]; \quad (4)$$

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

$$\left(\begin{array}{c} \text{posterior} \\ \text{odds} \\ \text{for } M_2 \\ \text{over } M_1 \end{array} \right) = \left(\begin{array}{c} \text{prior odds} \\ \text{for } M_2 \\ \text{over } M_1 \end{array} \right) \cdot \left(\begin{array}{c} \text{Bayes factor} \\ \text{for } M_2 \\ \text{over } M_1 \end{array} \right). \quad (5)$$

(**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 $p(M_2|D\mathcal{B}) > p(M_1|D\mathcal{B})$ } are **equivalent**.

Decision-Theoretic Basis for Bayes Factors

This approach does have a **decision-theoretic basis**, but it's rather **odd**: if You pretend that the **only possible data-generating mechanisms** are

$\mathcal{M} = \{M_1, \dots, M_m\}$ 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**

$$U(M, M_{DG}) = \begin{cases} 1 & \text{if } M = M_{DG} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

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 ,

$$BF(M_2 \text{ over } M_1 | D \mathcal{B}) = \frac{p(D|M_2 \mathcal{B})}{p(D|M_1 \mathcal{B})}, \quad (7)$$

Parametric Model Comparison

favoring M_2 if $BF(M_2 \text{ over } M_1 | D \mathcal{B}) > 1$, i.e., if $p(D|M_2 \mathcal{B}) > p(D|M_1 \mathcal{B})$, 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: $p(D|M_j \mathcal{B})$ 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** γ_j (of length k_j), where $\gamma_j = (\theta, \eta_j)$, and is **specified** by

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

Here the quantity $p(D|M_j \mathcal{B})$ 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 γ_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 γ_j in model M_j (evaluated at the **observed data set** D).

A few **additional words** about **prior distributions** on **parameters**:

Instability of Bayes Factors

A **distribution (density)** for a **real-valued parameter** θ that summarizes the **information**

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

will have **most of its mass** concentrated **near** θ_0 ,
whereas the **information**

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

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

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

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

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

Instability of Bayes Factors (continued)

$M_1 = \mathbf{Geometric}(\theta_1)$ likelihood with a **Beta** (α_1, β_1) prior on θ_1 ;

$M_2 = \mathbf{Poisson}(\theta_2)$ likelihood with a **Gamma** (α_2, β_2) prior on θ_2 .

The **Bayes factor** in favor of M_1 over M_2 turns out to be

$$\frac{\Gamma(\alpha_1 + \beta_1) \Gamma(n + \alpha_1) \Gamma(n\bar{y} + \beta_1) \Gamma(\alpha_2) (n + \beta_2)^{n\bar{y} + \alpha_2} (\prod_{i=1}^n y_i!)}{\Gamma(\alpha_1) \Gamma(\beta_1) \Gamma(n + n\bar{y} + \alpha_1 + \beta_1) \Gamma(n\bar{y} + \alpha_2) \beta_2^{\alpha_2}}. \quad (11)$$

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

$$\frac{\Gamma(n + 1) \Gamma(n\bar{y} + 1) \Gamma(\epsilon) (n + \epsilon)^{n\bar{y} + \epsilon} (\prod_{i=1}^n y_i!)}{\Gamma(n + n\bar{y} + 2) \Gamma(n\bar{y} + \epsilon) \epsilon^\epsilon}. \quad (12)$$

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** (α_2, β_2) away from 0 and let $(\alpha_1, \beta_1) \downarrow 0$, You can **completely reverse** this and make the evidence in **favor** of the **Poisson model** over the **Geometric** as **large** as You want (for **any** y).

Approximating Integrated Likelihoods

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

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

Approximating integrated likelihoods. The goal is

$$p(D|M_j \mathcal{B}) = \int p(D|\gamma_j M_j \mathcal{B}) p(\gamma_j|M_j \mathcal{B}) d\gamma_j; \quad (13)$$

maybe there's an **analytic approximation** to this that will suggest how to **avoid trouble**.

Laplace Approximation

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** $P^*(\gamma_j) \equiv p(D|\gamma_j M_j \mathcal{B}) p(\gamma_j|M_j \mathcal{B})$ in $p(D|M_j \mathcal{B})$ is an **un-normalized version** of the **posterior distribution** $p(\gamma_j|D M_j \mathcal{B})$, 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

$$\begin{aligned} \log p(D|M_j \mathcal{B}) &= \log p(D|\hat{\gamma}_j M_j \mathcal{B}) + \log p(\hat{\gamma}_j|M_j \mathcal{B}) \\ &\quad + \frac{k_j}{2} \log 2\pi - \frac{1}{2} \log |\hat{I}_j| + O\left(\frac{1}{n}\right); \quad (14) \end{aligned}$$

here $\hat{\gamma}_j$ is the **maximum likelihood estimate** of the **parameter vector** γ_j under **model** M_j and \hat{I}_j is the **observed information matrix** under M_j .

Notice that the **prior** on γ_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:

$$AIC(M_j|D\mathcal{B}) = -2 \log p(D|\hat{\gamma}_j M_j \mathcal{B}) + 2 k_j. \quad (17)$$

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

$$(\gamma_j|M_j \mathcal{B}) \sim N_{k_j}(\hat{\gamma}_j, n\hat{l}_j^{-1}). \quad (18)$$

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 .

Bayes Factors; Log Scores

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 18** 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 $D = y = (y_1 \dots y_n)$ for real-valued y_i and the **models** to be **compared** are (as before)

$$M_j: \left\{ \begin{array}{l} (\gamma_j | M_j \mathcal{B}) \sim p(\gamma_j | M_j \mathcal{B}) \\ (y | \gamma_j M_j \mathcal{B}) \sim p(y | \gamma_j M_j \mathcal{B}) \end{array} \right\}. \quad (19)$$

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**,

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

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**,

Full-Sample Log Score

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

made **operational** with the **rule** {favor M_2 over M_1 if $LS_{FS}(M_2|y \mathcal{B}) > LS_{FS}(M_1|y \mathcal{B})$ }, 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 $p(\cdot|y M_j \mathcal{B})$, in which M_{DG} is **approximated** by the **empirical CDF**:

$$\begin{aligned} KL[p(\cdot|y M_j \mathcal{B})||M_{DG}] &= E_{M_{DG}} \log M_{DG} - E_{M_{DG}} \log p(\cdot|y M_j \mathcal{B}) \\ &\doteq E_{M_{DG}} \log M_{DG} - LS_{FS}(M_j|y \mathcal{B}); \quad (22) \end{aligned}$$

the **first term** on the **right side** of (22) is **constant** in $p(\cdot|y M_j \mathcal{B})$, so **minimizing** $KL[p(\cdot|y M_j \mathcal{B})||M_{DG}]$ is **approximately the same** as **maximizing** LS_{FS} .

Bayes Factors/BIC Versus Log Scores

What follows is a **sketch of recent results** (Draper, 2013) 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 $n \rightarrow \infty$ 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**

Clinical Trial to Quantify Improvement

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 θ 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 $D = y = (y_1 \dots y_n)$. where y_i is the **observed difference** ($SBP_{before} - SBP_{after}$) for **patient** i ($i = 1, \dots, n$).

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**

Decision, Not Inference

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 θ , but **it's not**; it's a **decision problem** that **involves** θ .

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 $\mathcal{A} = (a_1, a_2) =$ (**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 θ over a **broad range** of **positive** θ 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 θ range,

Models For Quantifying Improvement

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 θ is not exactly 0 (because the **outcome variable in this experiment is conceptually continuous**).

What **matters** here is whether $\theta > \Delta$, where Δ 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 θ **external** to this **experimental data set**, what **counts** in this **situation** is the **comparison** of the following **two models**:

$$M_1: \left\{ \begin{array}{l} (\theta|\mathcal{B}) \sim \text{diffuse for } \theta \leq \Delta \\ (y_i|\theta\mathcal{B}) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \end{array} \right\} \text{ and} \quad (23)$$

$$M_2: \left\{ \begin{array}{l} (\theta|\mathcal{B}) \sim \text{diffuse for } \theta > \Delta \\ (y_i|\theta\mathcal{B}) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \end{array} \right\}, \quad (24)$$

Quantifying Improvement: Model Comparison Methods

in which **for simplicity** I'll take σ^2 to be **known** (the **results** are **similar** with σ^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 $LS_{FS}(M_2|y \mathcal{B}) > LS_{FS}(M_1|y \mathcal{B})$.

- **Posterior probability**: let

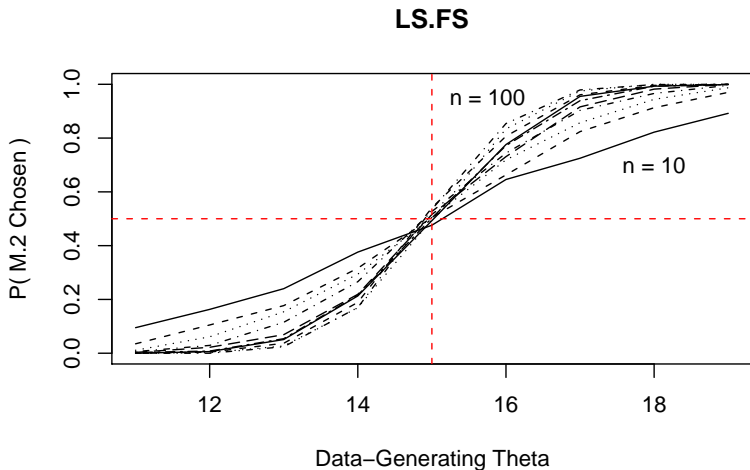
$M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \mathfrak{R}, (y_i|\theta \mathcal{B}) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)\}$ and **choose** M_2 if $p(\theta > \Delta|y M^* \mathcal{B}) > 0.5$.

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

Simulation experiment details, based on the **SBP drug trial**: $\Delta = 15$;
 $\sigma = 10$; $n = 10, 20, \dots, 100$; **data-generating** $\theta_{DG} = 11, 12, \dots, 19$;
 $\alpha = 0.05$; **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**.

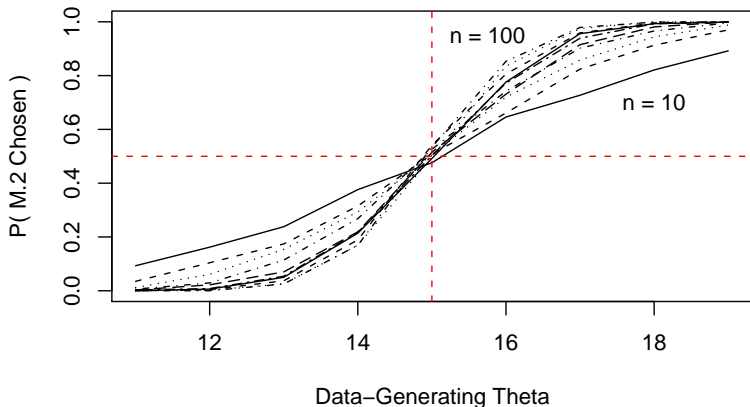
LS_{FS} Results: Quantifying Improvement



This exhibits all the **monotonicities** that it **should**, and **correctly yields 0.5** for all n with $\theta_{DG} = 15$.

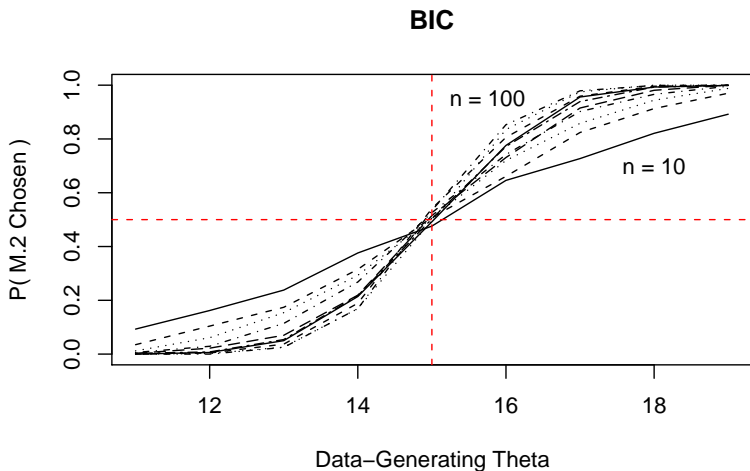
Posterior Probability Results: Quantifying Improvement

Posterior Probability



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 ± 0.003 (well within simulation noise with **1,000** replications).

BIC Results: Quantifying Improvement



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

- **(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 θ stand for the **mean difference**

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \quad (25)$$

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 ($i = 1, \dots, n$).

Again in this **setting** there's **nothing special** about $\theta = 0$, and as **before** You **know scientifically** that θ is **not exactly 0**;

Bio-Equivalence Modeling

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 θ **external** to this **experimental data set**, what **counts** here is a **comparison** of

$$M_3: \left\{ \begin{array}{l} (\theta|\mathcal{B}) \sim \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta \mathcal{B}) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \quad (26)$$

$$M_4: \left\{ \begin{array}{l} (\theta|\mathcal{B}) \sim \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta \mathcal{B}) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \end{array} \right\}, \quad (27)$$

in which σ^2 is again taken for **simplicity** to be **known**.

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

$$M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \mathfrak{R}, (y_i|\theta \mathcal{B}) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)\}, \text{ but this time favor } M_4 \text{ over } M_3 \text{ if } p(|\theta| > \lambda | y M^* \mathcal{B}) > 0.5.$$

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

Bio-Equivalence Model Comparison

costs and benefits of

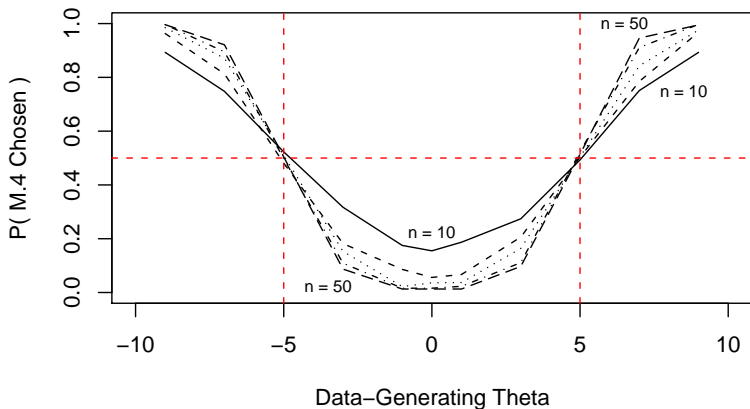
{**claiming** the two drugs were **bio-equivalent** when they **were**,
concluding that they were **bio-equivalent** when they **were not**,
deciding that they were **not bio-equivalent** when they **were**,
judging that they were **not bio-equivalent** when they were **not**},

but here I'll again simply **compare** the **calibrative performance** of
 LS_{FS} , **posterior probabilities**, and **BIC**.

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

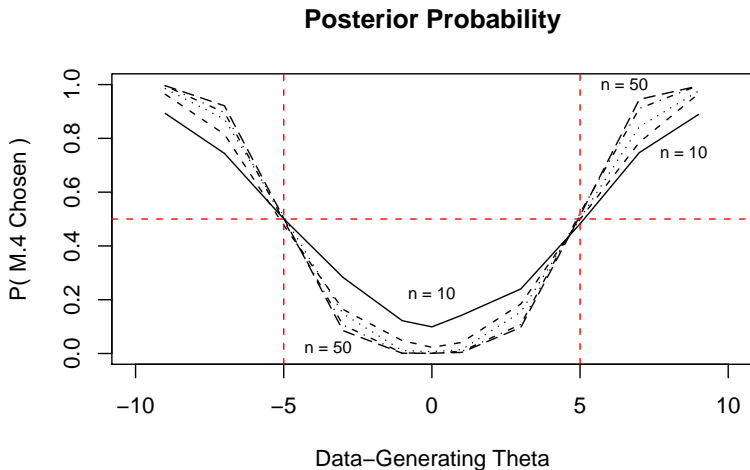
NB It has **previously been established** that when **making** the
(unrealistic) sharp-null comparison $\theta = 0$ versus $\theta \neq 0$ in the **context**
of $(y_i | \theta) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$, as $n \rightarrow \infty$ LS_{FS} **selects** the $\theta \neq 0$ **model** with
probability $\rightarrow 1$ even when $\theta_{DG} = 0$; 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**.

LS.FS



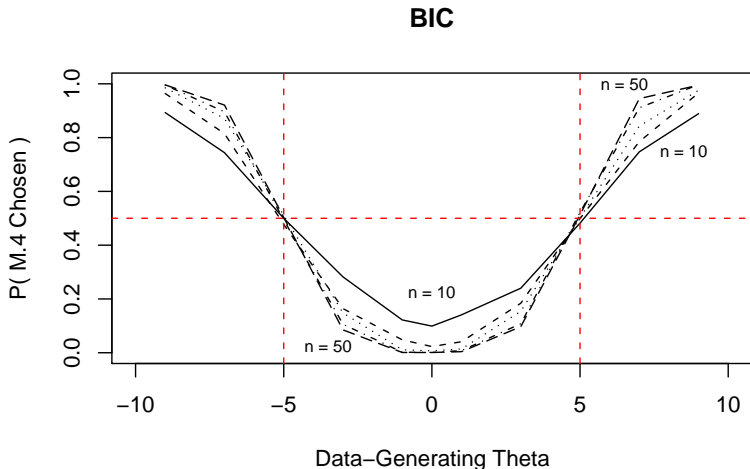
In this **more realistic setting**, comparing $|\theta| \leq \lambda$ versus $|\theta| > \lambda$ with $\lambda > 0$, LS_{FS} has the **correct large-sample behavior**, **both** when $|\theta_{DG}| \leq \lambda$ and when $|\theta_{DG}| > \lambda$.

Posterior Probability Results: Bio-Equivalence



The **qualitative behavior** of the LS_{FS} and **posterior-probability methods** is **identical**, although there are some **numerical differences** (**highlighted** later).

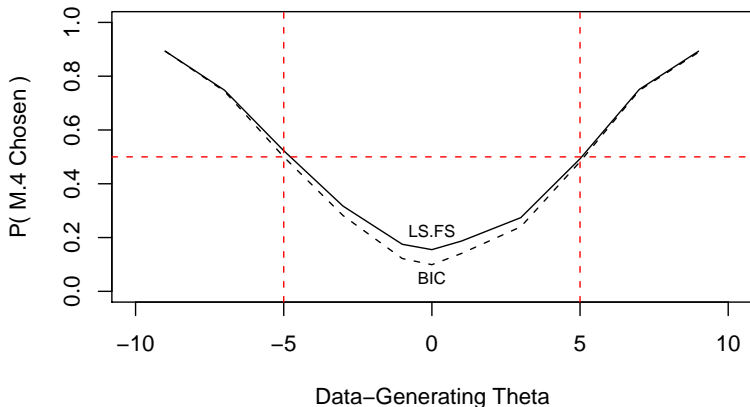
BIC Results: Bio-Equivalence



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).

LS_{FS} Versus BIC Results: Bio-Equivalence

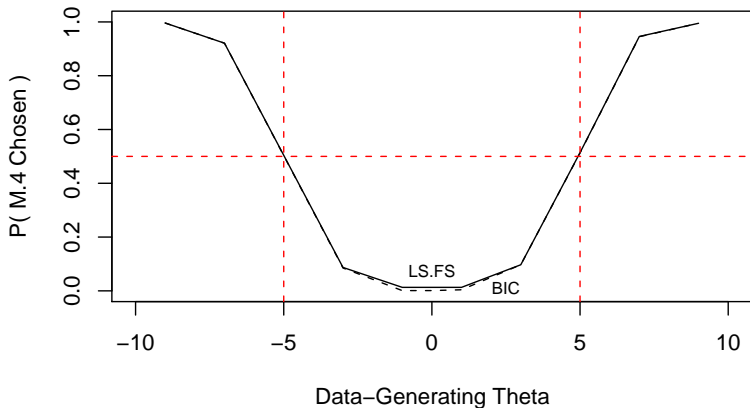
LS.FS Versus BIC (n = 10)



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**.

LS_{FS} Versus BIC Results: Bio-Equivalence

LS.FS Versus BIC (n = 50)



By the time You **reach** $n = 50$ in **this problem**, LS_{FS} and BIC are **essentially equivalent**.

The Decision-Versus-Inference Principle, Revisited

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** α .

Decision-theoretically this is **already wrong**; as **noted** back on **page 34**, the **utility function** should **actually** be **continuous** in θ rather than **artificially dichotomizing** Θ into $(-\infty, \Delta]$ and (Δ, ∞) .

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 $u_{ij} \geq 0$):

<u>Action</u>	<u>Truth</u>	
	$\theta \leq \Delta$	$\theta > \Delta$
a_1 (stop)	u_{11}	$-u_{12}$
a_2 (phase III)	$-u_{21}$	u_{22}

Decision-Theory (Not Inference) For Decision Problems

<u>Action</u>	<u>Truth</u>	
	$\theta \leq \Delta$	$\theta > \Delta$
a_1 (stop)	u_{11}	$-u_{12}$
a_2 (phase III)	$-u_{21}$	u_{22}

- u_{11} is the **gain** from **correctly not taking the drug forward** to **phase III** (this is clearly **0**);
- u_{12} is the **loss** from **incorrectly failing to take the drug forward** to **phase III**;
- u_{21} is the **loss** from **incorrectly taking the drug forward** to **phase III**;
- u_{22} is the **gain** from **correctly taking the drug forward** to **phase III**.

The **optimal Bayesian decision** turns out to be:
choose a_2 (go forward to phase III) iff

$$P(\theta > \Delta | y \mathcal{B}) \geq \frac{u_{21}}{u_{12} + u_{21} + u_{22}} = u^*. \quad (28)$$

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

Optimal Decision-Making in Phase-II Trials

$$\alpha = 1 - u^* = \frac{u_{12} + u_{22}}{u_{12} + u_{21} + u_{22}}. \quad (29)$$

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

In **phase-II clinical trials** or **micro-array experiments**, when You're **screening many drugs** or **genes** for those that **may lead** to an **effective treatment** and — from the **drug company's point of view** — a **false-negative error** (of **failing to move forward** with a **drug** or **gene** that's actually **worth further investigation**) can be **much more costly** than a **false-positive mistake**, this **corresponds** to $u_{12} \gg u_{21}$ and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use (**much**) **larger** α **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 α **values** on the order of **0.45**, which is **close** to the **implicit trade-off** in **BIC** and LS_{FS} .)

For People Who Like to Test Sharp-Null Hypotheses

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

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 Θ of **possible values** of θ : **settings** where it's **scientifically important** to **distinguish** between $\theta = \theta_0$ and $\theta \neq \theta_0$ — an **example** would be **discriminating** between $\{\text{these two genes are on different chromosomes (the strength } \theta \text{ of their genetic linkage is } \theta_0 = 0)\}$ and $\{\text{these two genes are on the same chromosome } (\theta > 0)\}$.

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)**

Testing Sharp-Null Hypotheses (continued)

(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** ($i = 1, \dots, n$)

$$M_5: \left\{ \begin{array}{l} (\sigma^2 | \mathcal{B}) \sim \text{diffuse on } (0, \text{large}) \\ (y_i | \sigma^2 \mathcal{B}) \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \end{array} \right\} \text{ and} \quad (30)$$

$$M_6: \left\{ \begin{array}{l} (\theta | \sigma^2 \mathcal{B}) \sim \text{diffuse on } (-\text{large}, \text{large}) \times (0, \text{large}) \\ (y_i | \theta \sigma^2 \mathcal{B}) \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2) \end{array} \right\}, \quad (31)$$

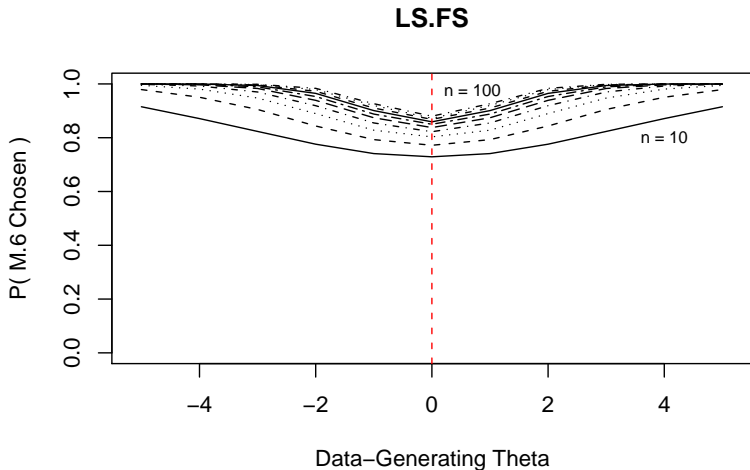
In **this case** a **natural Bayesian competitor** to **BIC** and LS_{FS} would be to **construct** the **central** $100(1 - \alpha)\%$ **posterior interval** for θ under M_6 and **choose** M_6 if **this interval doesn't contain 0**.

Testing Sharp-Null Hypotheses (continued)

Simulation experiment details: data-generating $\sigma_{DG} = 10$; $n = 10, 20, \dots, 100$; data-generating $\theta_{DG} = \{0, 1, \dots, 5\}$; **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.

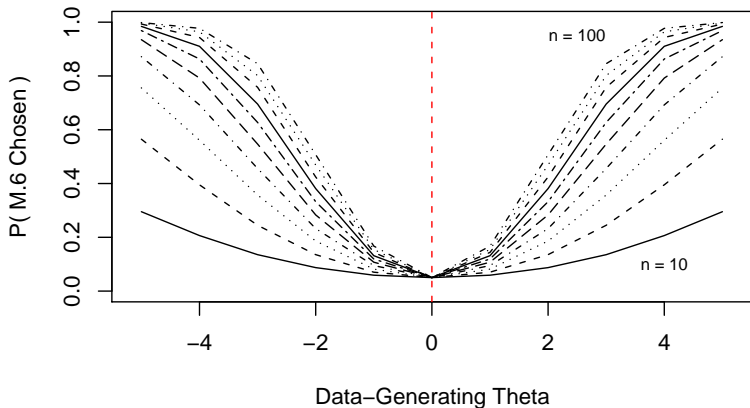
LS_{FS} Results: Sharp-Null Testing



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

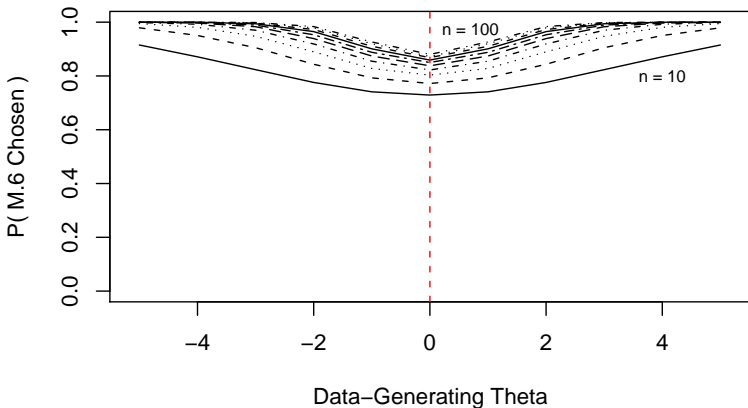
Posterior Interval (alpha = 0.05)



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**.

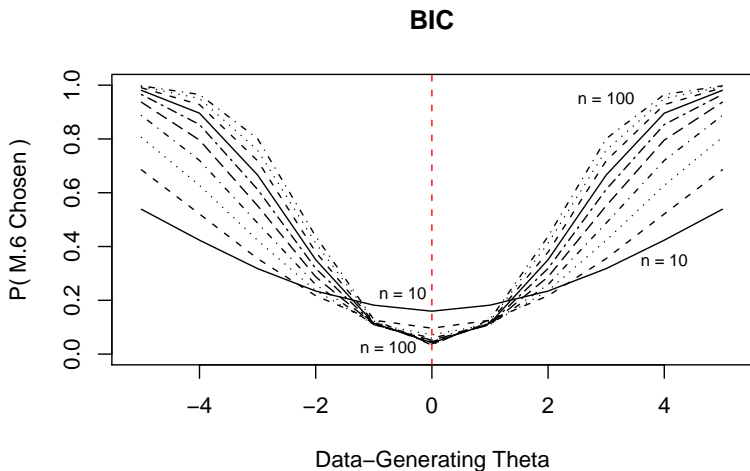
Interval (α Modified to LS_{FS} Behavior) Results

Posterior Interval (alpha Modified to LS.FS Behavior)



When the **interval method** is **modified** so that α **matches** the LS_{FS} **behavior** at $\theta_{DG} = 0$ (letting α **vary** with n), the **two approaches** have **identical model-discrimination ability**.

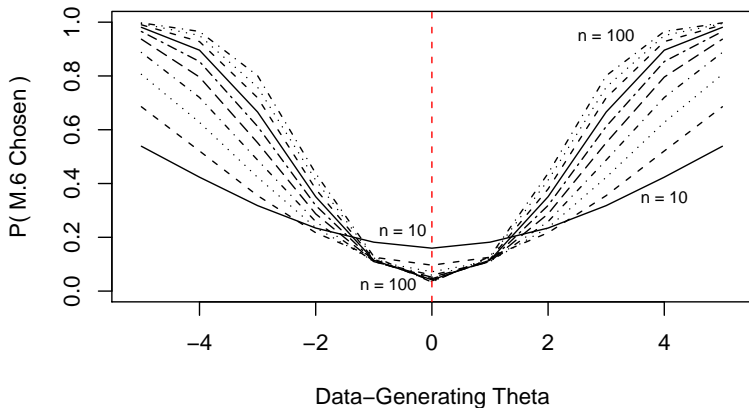
BIC Results: Sharp-Null Testing



BIC's behavior is quite different from that of LS_{FS} and fixed- α posterior intervals: its false-positive rate decreases as n grows, but it suffers a high false-negative rate to achieve this goal.

Interval (α Modified to BIC Behavior) Results

Posterior Interval (alpha Modified to BIC Behavior)



When the **interval method** is **modified** so that α **matches** the **BIC behavior** at $\theta_{DG} = 0$ (again letting α **vary** with n), the **two approaches** have **identical model-discrimination ability**.

LS_{FS} Versus BIC: Geometric Versus Poisson

As another **model-comparison example**, suppose You have an **integer-valued** data set $D = y = (y_1 \dots y_n)$ and You wish to **compare**

$M_7 =$ **Geometric**(θ_1) **sampling distribution** with a **Beta**(α_1, β_1) **prior** on θ_1 , and

$M_8 =$ **Poisson**(θ_2) **sampling distribution** with a **Gamma**(α_2, β_2) **prior** on θ_2 .

LS_{FS} and **BIC** both have **closed-form expressions** in this **situation**:

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

$$\begin{aligned} LS_{FS}(M_7|y \mathcal{B}) &= \log \Gamma(\alpha_1 + n + \beta_1 + s) + \log \Gamma(\alpha_1 + n + 1) \\ &\quad - \log \Gamma(\alpha_1 + n) - \log \Gamma(\beta_1 + s) \quad (32) \\ &\quad + \frac{1}{n} \sum_{i=1}^n [\log \Gamma(\beta_1 + s + y_i) \\ &\quad - \log \Gamma(\alpha_1 + n + \beta_1 + s + y_i + 1)], \end{aligned}$$

$$BIC(M_7|y \mathcal{B}) = -2[n \log \hat{\theta}_1 + s \log(1 - \hat{\theta}_1)] + \log n, \quad (33)$$

Geometric Versus Poisson (continued)

$$\begin{aligned}LS_{FS}(M_8|y\mathcal{B}) &= (\alpha_2 + s) \log(\beta_2 + n) - \log \Gamma(\alpha_2 + s) \\ &\quad - (\alpha_2 + s) \log(\beta_2 + n + 1) \\ &\quad + \frac{1}{n} \sum_{i=1}^n [\log \Gamma(\alpha_2 + s + y_i) - y_i \log(\beta_2 + n + 1) \\ &\quad - \log \Gamma(y_i + 1)], \text{ and}\end{aligned}\tag{34}$$

$$BIC(M_8|y\mathcal{B}) = -2[s \log \hat{\theta}_2 - n \hat{\theta}_2 - \sum_{i=1}^n \log(y_i!)] + \log n,\tag{35}$$

$$\text{where } \hat{\theta}_2 = \frac{\alpha_2 + s}{\beta_2 + n}.$$

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 $M_{DG} = \mathbf{Geometric}$ a **false-Poisson** error and **choosing** M_7 (Geometric) when $M_{DG} = \mathbf{Poisson}$ a **false-Geometric** mistake.

Geometric Versus Poisson (continued)

The **table below** records the **Monte-Carlo probability** that the **Poisson model** was chosen.

M.DG = Poisson			M.DG = Geometric		
n	LS.FS	BIC	n	LS.FS	BIC
10	0.8967	0.8661	10	0.4857	0.4341
20	0.9185	0.8906	20	0.3152	0.2671
40	0.9515	0.9363	40	0.1537	0.1314
80	0.9846	0.9813	80	0.0464	0.0407

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}).

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

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

in which

$$\begin{aligned} p(y_i|y M_j \mathcal{B}) &= \int p(y_i|\gamma_j M_j \mathcal{B}) p(\gamma_j|y M_j \mathcal{B}) d\gamma_j & (36) \\ &= E_{(\gamma_j|y M_j \mathcal{B})} p(y_i|\gamma_j M_j \mathcal{B}); \end{aligned}$$

this **expectation** is over the **posterior (not the prior) distribution** for the **parameter vector** γ_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**,

Properties of LS_{FS} (continued)

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 LS_{FS} , which **favors model M_j over $M_{j'}$** if

$$n LS_{FS}(M_j|y, \mathcal{B}) > n LS_{FS}(M_{j'}|y, \mathcal{B}). \quad (37)$$

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 $p(y_i|\eta_j, M_j, \mathcal{B})$ under M_j ,

$$nLS_{FS}(M_j|y, \mathcal{B}) = \log \prod_{i=1}^n \left[\int p(y_i|\eta_j, M_j, \mathcal{B}) p(\eta_j|y, M_j, \mathcal{B}) d\eta_j \right], \quad (38)$$

and this is **not the same as**

$$\log \int \left[\prod_{i=1}^n p(y_i|\eta_j, M_j, \mathcal{B}) \right] p(\eta_j|y, M_j, \mathcal{B}) d\eta_j = \bar{L}_j^{PBF} \quad (39)$$

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**,

Summary (continued)

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 $\$k C$ with $k = 10\text{--}100$.

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

$$\begin{aligned}(y_i | \lambda_i, \mathcal{B}) &\stackrel{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log \lambda_i &= \beta_0 + \beta_1 x_i + e_i \\ (e_j | \sigma_e^2, \mathcal{B}) &\stackrel{\text{iID}}{\sim} N(0, \sigma_e^2), \quad (\beta_0, \beta_1, \sigma_e^2) \sim \text{diffuse},\end{aligned}\tag{40}$$

Summary (continued)

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 β_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**.

Summary (continued)

- **Asymptotic conclusions are often misleading:** while it's **true** that

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

it's **also true** that

New Theorem (Draper, 2013): for any $\lambda > 0,$
 $P_{|\theta_{DG}| \leq \lambda}(LS_{FS} \text{ chooses } |\theta| \leq \lambda) \rightarrow 1 \text{ as } n \rightarrow \infty,$

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 \rightarrow **Axioms** \rightarrow **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

Is M_1 Good Enough? (continued)

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

Principles → **Axioms** → **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

Summary (continued)

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

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

* in addition $\{\mathcal{A}, U(a, \theta)\}$ for **decision**, where \mathcal{A} is **Your set of available actions** and $U(a, \theta)$ is **Your utility function** (mapping from **actions a** and unknown θ 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** $\mathcal{M} = \{M_1, M_2, \dots\}$, which gives rise immediately to **two questions**:

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

Summary (continued)

- 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 $\mathcal{M}_{\text{current}} \leftarrow \{M_0\}$.
- (b) If M_{current} is good enough to stop (how decide?), return $\mathcal{M}_{\text{current}}$; else
- (c) Generate a new candidate model M_{new} (how choose?) and set $\mathcal{M}_{\text{current}} \leftarrow \mathcal{M}_{\text{current}} \cup M_{\text{new}}$.
- (d) If M_{new} is better than M_{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**.

Summary (continued)

- 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 $\{\text{choosing } M_2 \text{ when } M_{DG} = M_1\}$ a **false positive** and $\{\text{choosing } M_1 \text{ when } M_{DG} = M_2\}$ 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 $\{\text{choosing } M_1 \text{ when } M_{DG} = M_1, \text{ choosing } M_1 \text{ when } M_{DG} = M_2, \text{ choosing } M_2 \text{ when } M_{DG} = M_1, \text{ choosing } M_2 \text{ when } M_{DG} = M_2\}$ and **maximize expected utility**.

Summary (continued)

— 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 (2013) 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 \mathcal{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**.)

Cromwell's Rule, Part 1: Inference and Prediction

The following two facts are easy consequences of the definition of conditional probability: for any two propositions A and B and any background information \mathcal{B} :

- **Cromwell's Rule, Part 1(a)** If $P(A|\mathcal{B}) = 0$ then $P(A|B\mathcal{B}) = 0$ for all B for which $P(A|B\mathcal{B})$ is defined (i.e., for which $P(B|\mathcal{B}) > 0$).
- **Cromwell's Rule, Part 1(b)** If $P(A|\mathcal{B}) = 1$ then $P(A|B\mathcal{B}) = 1$ for all B for which $P(A|B\mathcal{B})$ is defined (i.e., for which $P(B|\mathcal{B}) > 0$).

To see the implications of these facts, let A be a proposition about something unknown to You, such as $(\theta < 0)$, and let B be a proposition about Your data set D , such as $(y_1 = 3, y_2 = -0.4, \dots, y_n = 6.9)$.

Then Part 1(a) of Cromwell's Rule says that any proposition about the unknown θ to which You give prior probability 0 **must** have posterior probability 0, no matter how the data set D comes out, and Part 1(b) of Cromwell's Rule says the same thing with 0 replaced by 1.

Cromwell's Rule Part 1 (continued)

Bayes's Theorem is supposed to be a piece of machinery that permits **You** to learn, about unknowns from new data, in an **optimal** way; **Cromwell's Rule Part 1** says that if **You** dogmatically place prior probability **0** or **1** on something, no learning is possible when new data values arrive.

This is obviously a way to break the **Bayes's Theorem** learning machine, so the practical consequence of **Cromwell's Rule Part 1** is captured in the following piece of advice:

You should try hard never to put prior probability 0 or 1 on anything that might later have posterior probability between 0 and 1, depending on how new information comes out.

This has direct consequences for **Bayesian model specification**: for example, consider the **NB10** data set from **Day 1**, for which (by exchangeability) **Your basic sampling model** for the data values $y = (y_1, \dots, y_n)$ before **You see the data** is $(y_i | F \mathcal{B}) \stackrel{\text{iid}}{\sim} F$ for some (unknown) CDF F .

Cromwell's Rule Part 1 (continued)

If, before You see the data, You put all Your modeling eggs in the Gaussian basket, so that You replace $(y_i|F\mathcal{B}) \stackrel{\text{iid}}{\sim} F$ with $(y_i|\mu\sigma\mathcal{B}) \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, this means that You've placed all of your prior probability on the Gaussian family, thereby implicitly giving prior probability 0 to all non-Gaussian behavior (such as multimodality, skewness and/or heavy tails).

Now what do You do when the data set arrives and — as in the case of the NB10 data — demonstrates much heavier tails than the Gaussian family can accommodate?

Strictly speaking, by Cromwell's Rule Part 1(a), **all such non-Gaussian behavior must have posterior probability 0**, and yet the data set clearly makes You wish that You hadn't been so dogmatic in Your “prior on model space.”

Going back and changing Your prior on model space based on how the data set came out is a clear violation of the dichotomization {information internal to D , information external to D }, which was part of the axiomatization in the Day 3 (Part 1) Lecture Notes:

Cromwell's Rule, Part 2: Decision-Making

in effect, **when You do this You're using the data twice** — once in **specifying the prior on model space**, and **again in updating that prior with the data set D** — and the **typical consequence** will be **understatement of Your actual uncertainty**.

I see only two ways out of this dilemma:

- **Bayesian nonparametric methods**, which — when **used properly** — give **positive prior probability to all possible CDFs F** , and
- **Calibration cross-validation**, which (1) **allows You to “cheat” by looking at the data and changing Your prior on model space** but (2) **forces You to pay an appropriate price for having done so**.

Cromwell's Rule, Part 2: Part 1 of Cromwell's Rule is about **inference and prediction**; Part 2 concerns **decision-making**, and it also has a **part (a)** and a **part (b)**, which **give advice on how to specify Your action space \mathcal{A} and Your utility function $U(a, \theta)$ (respectively)**.

Cromwell's Rule, Part 2 (continued)

- **Cromwell's Rule, Part 2(a):** In enumerating the possible actions $\{a_1, a_2, \dots\}$ while specifying Your action space \mathcal{A} in the problem \mathbb{P} on which You're working, You should try hard not to omit any action a_i that might turn out to be optimal if it's included in \mathcal{A} .

The point is that Bayesian decision theory only optimizes over the possible actions You remember to put in \mathcal{A} : if You forget a feasible action a_i that (unknown to You) is better than all of the actions in Your current \mathcal{A} , maximization of expected utility cannot protect You from this omission.

Example: HIV screening with *ELISA* and Western Blot (Day 1, Lecture Notes Part 1). In that case study it was tempting to think that the only two possible actions were $a_1 = \{\text{test the blood sample with ELISA}\}$ and $a_2 = \{\text{test the blood sample with Western Blot}\}$, but a third action — $a_3 = \{\text{test half of the blood sample with ELISA; if negative, declare HIV negative; if positive, test the other half of the blood sample with Western Blot}\}$ — turned out to dominate the others.

Cromwell's Rule, Part 2 (continued)

Example: People at eBay are constantly running randomized controlled trials on the eBay web experience, looking for variations on things like (i) search algorithms and (ii) {presentation of items for sale to the users} that may create a better marketplace.

Having run an experiment in which the control group gets the current best eBay website and the treatment group gets {the current best eBay website plus a particular intervention I }, thereby obtaining a data set D , the unknown θ is {what the future would be like, as far as important outcome variables (such as user satisfaction) are concerned, if intervention I is or is not implemented}, and it appears that there are only two possible actions: $a_1 = \{\text{implement } I\}$ and $a_2 = \{\text{don't implement } I\}$.

However, as with the HIV case study, there's a third possible action that has an **adaptive** flavor: perhaps there's still too much uncertainty, on the basis of D , to make a good choice between a_1 and a_2 , so (if You're not in a tremendous hurry to choose) why not include $a_3 = \{\text{get more data before deciding}\}$ in Your \mathcal{A} ?

Cromwell's Rule, Part 2 (continued)

Bayesian sequential experimental design and analysis has this adaptive character — **get some data, see if the optimal choice is clearcut yet, if so make it, if not get more data** — and have been shown to yield results with **good false-positive and false-negative rates that involve collecting (far) less data than approaches with sample sizes that are fixed at the design stage.**

- **Cromwell's Rule, Part 2(a):** In specifying Your utility function for the problem \mathbb{P} on which You're working, You should try hard to ensure that
 - (1) Your vector of unknowns θ contains all relevant unknowns, and
 - (2) Your utility function $U(a, \theta)$ captures all relevant costs and benefits to be balanced against each other.

The point is that

- any relevant unknown that You mistakenly omit from θ has **no opportunity to influence Your decision**; the result will often be decisions that **don't hedge sufficiently against uncertainty,**

Cromwell's Rule, Part 2 (continued)

because **omitting** a **relevant unknown** is **tantamount** to **pretending that it's known**; and

- **any relevant cost or benefit that You mistakenly omit** from $U(a, \theta)$ **has no opportunity** to **influence** the **optimal trade-off** between **costs and benefits**, and **(dramatically) sub-optimal decisions can result when this happens.**

In **enumerating** the **relevant costs and benefits**, **You'll have to fight against** the following **three basic human tendencies**:

- (1) **Things that are easy to measure tend to get measured; things that are hard to measure tend to get ignored.**

With respect to a particular cost or benefit, the **classifications** $C_1 = \{\text{important, not important}\}$ and $C_2 = \{\text{hard to quantify, easy to quantify}\}$ **have nothing to do with each other; including or omitting costs and benefits solely on the basis of C_2 provides no assurance that the included costs and benefits are correct with respect to C_1 .**

Cromwell's Rule, Part 2 (continued)

(2) People with **optimistic world views** tend to **exaggerate benefits** and **downplay costs**; **pessimists** tend to **make the opposite mistakes**; **not many people get this right unless they're on the lookout for it.**

(3) If **costs and benefits** are **balanced against each other additively** in **Your utility function**, You'll typically find it **relatively easy to choose the scale on which to quantify the costs** (e.g., in **monetary terms**), but it may then be **quite difficult to quantify the benefits** on the same scale.

Example: The new drug *Olysio* (*simeprevir*, approved for use in the **U.S. in Nov 2013**) has been **shown to be quite effective at reducing the viral load of Hepatitis C patients, permitting their livers to begin to heal, as long as the drug is taken for at least three months.**

The **drug company** that **markets *Olysio*, Janssen**, has **set the wholesale price of a 12-week supply of this drug at \$66,360** (about **£41,000**).

Q: Should the **NHS** decide to approve *Olysio* for routine treatment of **Hepatitis C** in the **UK**?

The **cost term** in the **NHS's utility function** clearly comes out in £; to **trade this off** against the **health benefits** (e.g., **lengthened life span, better quality of life**), the **NHS** has to be **prepared to measure those benefits in monetary terms**, leading to **unpleasant questions** such as **“What's the monetary value of human life?”**

Example: Draper (1995) examines an attempt by economists in **1980** to **predict future oil prices** over the **horizon 1981–2020**; many **companies** and **governments** routinely make investment decisions based on **such predictions**.

The **OPEC oil embargo** of **1973–74** created a **large spike in oil prices**, because **demand stayed constant** and **supply dramatically dropped**; this was the **first time anything like that** had ever occurred.

Cromwell's Rule, Part 2 (continued)

If You had undertaken the 1980 predictive exercise mentioned above prior to 1973, You would have had no straightforward way to include {Will another OPEC oil embargo occur, and if so when?} as part of the unknowns in Your θ vector, but You would have no excuse for omitting this unknown after 1973–74, and indeed this omission could lead to a dramatic understatement of Your future uncertainty about the price of oil, causing You to make decisions that fail to hedge sufficiently against the totality of Your uncertainty.

Example: You're about to make a long drive by car, and You're wondering about the optimal driving speed: the faster You drive the quicker You get to Your destination (good), but undesirable outcomes increase in probability as You speed up (bad).

The action space clearly consists of possible speeds (that was easy), but what about θ and $U(a, \theta)$?

As a first pass, You might include in θ only the unknown {will You get a speeding ticket?}, in which case Your utility function would have only two terms, which could be combined additively:

Cromwell's Rule, Part 2 (continued)

a **benefit (quantified in monetary terms)** based on **how short the journey is**, and a **cost (also expressed in money)** based on **what happens if You get a ticket (You have to pay a fine, and Your insurance costs may rise)**.

I've found that this formulation typically leads to a recommendation to **drive quite rapidly**.

However, a number of additional relevant unknowns have been omitted from this first-pass specification of \mathcal{A} and $U(a, \theta)$:

- {Will You get into an accident?} If so, {How serious is the accident?} {How badly is Your car damaged?} {Does the accident injure or kill You?} If other people are involved in the accident, {How badly is their vehicle damaged?} {Does the accident injure or kill any of them?}

Increasingly unfavorable answers to all of these questions will all result in additional cost terms in Your utility function, with the result that Your optimal driving speed will monotonically decrease as You increase the realism of Your utility specification.