Bayesian Model Specification: Toward a *Theory of Applied Statistics*

David Draper

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

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

Short Course: San Francisco Chapter of the American Statistical Association

6 Nov 2011

Outline

(1) An **axiomatization** of **statistics** (Draper 2011).

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

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

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

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

(b) Too much **ad hockery** in **model specification**: still lacking **Principles** $\rightarrow$ **Axioms** $\rightarrow$ **Theorems**.

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

(5) The **Modeling-As-Decision Principle**, the **Prediction Principle** and the **Decision-Versus-Inference Principle** help with **3 (b)**.

An Example, to Fix Ideas

Example (Krnjajić, Kottas, Draper [KKD] 2008): *In-home geriatric assessment (IHGA)*. In an **experiment** conducted in the **1980s** (Hendriksen et al. 1984), **572 elderly people**, **representative** of $\mathcal{P} =$ {all **non-institutionalized elderly people** in **Denmark**}, were **randomized**, **287** to a **control** (C) group (who received **standard health care**) and **285** to a **treatment** (T) group (who received **standard care plus IHGA**: a kind of **preventive medicine** in which each person's **medical** and **social needs** were assessed and acted upon **individually**).

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

	Number of Hospitalizations						
Group	0	1	...	k	n	Mean	SD
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

Let μ_C and μ_T be the **mean hospitalization rates** (per two years) in $\mathcal{P}$ under the C and T **conditions**, respectively.

Here are **four statistical questions** that **arose** from **this study**:

The Four Principal Statistical Activities

Q_1: Was the **mean number of hospitalizations per two years** in the IHGA group **smaller** than that in **control** by an **amount** that was **large** in **practical** terms? $\left[\textbf{description} \text{ involving } \left(\frac{\bar{y}_T - \bar{y}_C}{\bar{y}_C}\right)\right]$

Q_2: Did IHGA **reduce** the **mean number of hospitalizations per two years** by an **amount** that was **large** in **statistical** terms? $\left[\textbf{inference} \text{ about } \left(\frac{\mu_T - \mu_C}{\mu_C}\right)\right]$

Q_3: On the **basis** of **this study**, how **accurately** can You **predict** the **total decrease** in **hospitalizations** over a period of N years if **IHGA** were **implemented throughout Denmark**? **[prediction]**

Q_4: On the **basis** of **this study**, is the **decision** to **implement IHGA** throughout Denmark **optimal** from a **cost-benefit** point of view? **[decision-making]**

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

An Axiomatization of Statistics

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

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

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

Examples: You may be **uncertain** about the **truth status** of

- $A =$ (**Barack Obama** will be **re-elected U.S. President** in **2012**), or
- $B =$ (the **in-hospital mortality rate** for patients at **hospital** H admitted in **calendar 2010** with a principal diagnosis of **heart attack** was **between 5% and 25%**).

4 **(implication)** It **follows** from 1–3 that **statistics** concerns **Your information** (**NOT** Your **beliefs**) about $A, B, \ldots$

Axiomatization (continued)

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

6 **(axiom)** These **assumptions** and **judgments**, which are themselves a form of **information**, can always be **expressed** in a **set** $\mathcal{B}$ of **propositions**.

Examples of $\mathcal{B}$: In the **IHGA study**, based on the **experimental design**, $\mathcal{B}$ would include the **propositions**

- (**Subjects** were **representative of [like a random sample from]** $\mathcal{P}$),
- (**Subjects** were **randomized** into one of two groups, **treatment** (**standard care + IHGA**) or **control** (**standard care**)).

7 **(definition)** Call the **"something of interest to You"** θ; in **applications** θ is often a **vector** (or **matrix**, or **array**) of **real numbers**, but **in principle** it could be **almost anything** (a **function**, an **image** of the **surface of Mars**, a **phylogenetic tree**, ...).

Axiomatization (continued)

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

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

9 **(implication)** The **presence** of D creates a **dichotomy**:

- **Your information** about θ {**internal, external**} to D.

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

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

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

Foundational Question

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

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

and using **all Your information** about θ (given $\mathcal{B}$) to make

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

(5) **Decisions** about how to **act sensibly**, even though **Your information** about θ may be **incomplete**.

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

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

Let's look **first** at the **probability foundations**.

Theory of Probability: Kolmogorov

From the **1650s (Fermat, Pascal)** through the **18th century (Bayes, Laplace)** to the period **1860–1930 (Venn, Boole, von Mises)**, **three different approaches** for how to think about **uncertainty quantification** — **classical, Bayesian**, and **frequentist probability** — were put forward in an **intuitive** way, but no one ever tried to prove a **theorem** of the form {**given these premises, there's only one sensible way to quantify uncertainty**} until **Kolmogorov**, **de Finetti**, and **RT Cox**.

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

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

Frequentist Probability: Kolmogorov

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

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

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

What's Ω? **This patient** is not the result of a **uniquely-specifiable repeatable "random" process**, he's just a guy who **walked into Your doctor's office**, and — throughout the **repetitions** of whatever **repeatable phenomenon** anyone might **imagine** — his **HIV status** is **not fluctuating "randomly"**: he's either **HIV positive** or he's **not**.

Theory of Probability: de Finetti

The **closest** You can come to making **Kolmogorov's approach work** here is to **imagine** the set Ω of **all people** {**similar to this patient in all relevant ways**} and ask **how often** You'd get an **HIV-positive person** if You **repeatedly chose** one person **at random** from Ω, but to **make this operational** You have to **specify** what You mean by **"similar to, in all relevant ways,"** and if You **try** to do this You'll notice that it's **not possible** to do so **uniquely** (in such a way that **all other reasonable people** would **unanimously agree** with You).

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

(**de Finetti** proved a **theorem** that's **equivalent** to the one **developed** by **Cox**; if You **prefer**, You can get to the **same place** (**probability** as an **operator** on **propositions** of **uncertain truth status**) with **de Finetti's betting primitive**, but (for me) **science** is about **information**, not **betting**.)

Theory of Probability: RT Cox

— **RT Cox** (1946): following **Laplace**, **probability** is a **quantification** of **information** about the **truth** of one or more **propositions**, constrained to obey **axioms** guaranteeing **internal logical consistency**; this is both **fundamental to science** and **as general as You can get**.

Cox's goal was to identify what **basic rules** $pl(A|B)$ — the **plausibility** (**weight of evidence** in favor) of (the **truth** of) A given B — should follow so that $pl(A|B)$ behaves **sensibly**, where A and B are **propositions** with B **assumed** by You to be **true** and the truth status of A **unknown** to You.

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

• Suppose You're **willing** to represent **degrees of plausibility** by **real numbers** (i.e., $pl(A|B)$ is a function from propositions A and B to $\Re$);

• You insist that **Your reasoning** be **logically consistent**:

— If a **plausibility assessment** can be arrived at in **more than one way**, then **every possible way** must lead to the **same value**.

Cox's Principles and Axioms

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

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

From these **principles** Cox derived a set of **axioms**:

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

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

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

Cox's Theorem

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

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

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

• **(normalization)** $P(A|B) + P(\overline{A}|B) = 1$, where $\overline{A} = (\text{not } A)$.

• **(the product rule)**:

$$P(A\,B|C) = P(A|C) \cdot P(B|A\,C) = P(B|C) \cdot P(A|B\,C).$$

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

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

- **(the sum rule)**:

$$P(A \text{ or } B|C) \equiv P(A+B|C) = P(A|C) + P(B|C) - P(A\,B|C).$$

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

$$P(A\,B\,C|D) = P(A|D) \cdot P(B|A\,D) \cdot P(C|A\,B\,D) \text{ and}$$

$$\begin{aligned} P(A+B+C|D) \;\; = \;\; & P(A|D) + P(B|D) + P(C|D) - P(A\,B|D) \\ & -P(A\,C|D) - P(B\,C|D) + P(A\,B\,C|D)\,. \end{aligned}$$

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

— **Example:** You're studying **quality of care** at the **17 Kaiser Permanente (KP) northern California hospitals** in **2003–7**, before the era of **electronic medical records**; during that time there was a **population** $\mathcal{P}$ of $N =$ **8,561 patients** at these facilities with a **primary admission diagnosis** of **heart attack**.

Inference About a Population Parameter

You take a **simple random sample** of $n = \mathbf{112}$ of these admissions and **record** whether or not each patient had an **unplanned transfer to the intensive care unit (ICU)**, observing $s = \mathbf{4}$ who did; θ is the **proportion** of such **unplanned transfers** in all of $\mathcal{P}$; here $\Theta = \{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N}{N}\}$, which can be **conveniently approximated** by $\Theta' = [0, 1]$.

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

In this setup θ is usually called a **(population) parameter**, and is **not itself the result of any sampling experiment** (**random** or **otherwise**); for this reason, it's **not possible** to **(directly) quantify uncertainty** about θ from the **Kolmogorov (set-theoretic)** point of view, but it makes **perfect sense** to do so from the **RT Cox (propositional)** point of view.

Optimal Reasoning About a Continuous θ

You could now **more generally define** a **function** $F_{(\theta|D\,\mathcal{B})}(q) = P(\theta \leq q|D\,\mathcal{B})$ and call it the **cumulative distribution function (CDF)** **for (not of)** $(\theta|D\,\mathcal{B})$, which is **shorthand** for

the **CDF** for **Your uncertainty about** θ given D and $\mathcal{B}$.

If $F_{(\theta|D\,\mathcal{B})}(q)$ turns out to be **continuous** and **differentiable** in q (I haven't said yet how to **calculate** F), it will be **convenient** to write

$$F_{(\theta|D\,\mathcal{B})}(b) - F_{(\theta|D\,\mathcal{B})}(a) = P(a < \theta \leq b|D\,\mathcal{B}) = \int_a^b p_{(\theta|D\,\mathcal{B})}(q)\,dq\,, \quad (1)$$

where the **(partial) derivative** $p_{(\theta|D\,\mathcal{B})}(q)$ of $F_{(\theta|D\,\mathcal{B})}$ with respect to q can be called the **density** **for (not of)** **(Your uncertainty about)** θ given D and $\mathcal{B}$.

In a **small abuse of notation** it's **common** to **write** $F(\theta|D\,\mathcal{B})$ and $p(\theta|D\,\mathcal{B})$ instead of $F_{(\theta|D\,\mathcal{B})}(q)$ and $p_{(\theta|D\,\mathcal{B})}(q)$ (respectively), letting the **argument** θ of $F(\cdot|D\,\mathcal{B})$ and $p(\cdot|D\,\mathcal{B})$ serve as a **reminder** of the **uncertain quantity** in question.

Ontology and Epistemology

NB In the **Kolmogorov approach** a **random variable** X is a **function** from Ω to some **outcome space** O, and if $O = \Re$ You'll often find it **useful** to **summarize** X's **behavior** through the **CDF of** X: $F_X(x) = P(\text{the set of } \omega \in \Omega \text{ such that } X(\omega) \leq x)$, usually written in **propositional-style shorthand** as $F_X(x) = P(X \leq x)$.

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

Jaynes (2003) makes a **worthwhile distinction**: the **statements**

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

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

Talking about "the **density of** θ" would be to **confuse ontology** and **epistemology**;

The Mind-Projection Fallacy

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

Returning to the **corollaries** of **Cox's Theorem**,

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

(a) It's **natural** (and indeed **You must be prepared** in this approach) to specify **two conditional probability distributions**:

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

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

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

Optimal Inference, Prediction and Decision

$$p(\theta|D\,\mathcal{B}) = c\,p(\theta|\mathcal{B})\,p(D|\theta\,\mathcal{B})\,, \qquad \textbf{(inference)} \qquad (2)$$

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

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

$$p(D^*|D\,\mathcal{B}) = \int_\Theta p(D^*|\theta\,D\,\mathcal{B})\,p(\theta|D\,\mathcal{B})\,d\theta\,;$$

often there's **no information** about D^* contained in D if θ is known, in which case this expression **simplifies** to

$$p(D^*|D\,\mathcal{B}) = \int_\Theta p(D^*|\theta\,\mathcal{B})\,p(\theta|D\,\mathcal{B})d\theta\,; \qquad \textbf{(prediction)} \qquad (3)$$

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

(i) the set $\mathcal{A}$ of **feasible actions** among which You're **choosing**, and

(ii) a **utility function** $U(a, \theta)$, taking values on $\Re$ and **quantifying** Your **judgments** about the **rewards** (monetary or otherwise) that would ensue if You chose **action** a and the **unknown** actually took the value θ — **without loss of generality** You can take **large values** of $U(a, \theta)$ to be **better than small values**;

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

$$a^* = \underset{a \in \mathcal{A}}{\text{argmax}}\, E_{(\theta|D\,\mathcal{B})}\, U(a, \theta) = \underset{a \in \mathcal{A}}{\text{argmax}} \int_{\Theta} U(a, \theta)\, p(\theta|D\,\mathcal{B})\, d\theta\,. \qquad (4)$$

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

Logical Consistency → Bayesian Reasoning Justified

Cox's Theorem is **equivalent** to the **assertion**

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

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

Advantages include a **unified probabilistic framework**: e.g., in my earlier **ICU example**, **Kolmogorov's non-Bayesian approach** does not permit **direct probability statements** about a **population parameter**, but **Cox's Theorem permits You** to make such statements (summarizing **all relevant available information**) in a natural way.

The Specification Burden

It's **worth noting**, however, that **there really is a theorem here**, of the form $A \to B$, from which $\overline{B} \to \overline{A}$; this **comes close to the assertion**

If You employ **non-Bayesian reasoning** then You're **open to the possibility** of **logical inconsistency**,

and indeed there have been some **embarrassing moments** in **non-Bayesian inference** over the past **100 years** (e.g., **negative estimates** for quantities that are **constrained** to be **non-negative**).

Challenges: These **corollaries** to **Cox's theorem** solve problems (3–5) above (page 8) — they leave **no ambiguity** about how to draw **inferences**, and make **predictions** and **decisions**, in the presence of **uncertainty** — but problems (1) and (2) are still **unaddressed**: to **implement** this **logically-consistent approach** in a given application, You have to **specify**

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

The Specification Burden (continued)

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

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

The results of **implementing** this approach are

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

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

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

To summarize: **Inference** and **prediction** require You to **specify** $p(\theta|\mathcal{B})$ and $p(D|\theta\,\mathcal{B})$; **decision-making** requires You to **specify** the same

two **ingredients** plus $\mathcal{A}$ and $U(a, \theta)$; how should this be done in a **sensible** way?

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

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

Developing (1) **principles**, (2) **axioms** and (3) **theorems** about **optimal specification** could be regarded as creating a **Theory of Applied Statistics**, which we **need** but **do not yet have**.

$p(\theta|\mathcal{B})$, $p(D|\theta\,\mathcal{B})$ and $\{\mathcal{A}, U(a, \theta)\}$ are all **important**; I'll **focus** here on the **problem** of **specifying** $\{p(\theta|\mathcal{B}), p(D|\theta\,\mathcal{B})\}$ — call such a **specification** a **model** M for **Your uncertainty** about θ (I'll have one **brief comment** about **decision theory** at the end).

What I Mean By Optimal Model Specification

How should M be **specified**? Where is the **progression**

Principles $\rightarrow$ **Axioms** $\rightarrow$ **Theorems**

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

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

Making **progress** on this **problem** requires **defining** the phrase **"optimal model specification;"** for this **purpose** the following **two-step argument** is **helpful**:

- All **Bayesian reasoning** under **uncertainty** is based on $P(A|B) = \frac{P(A\,B)}{P(B)}$ for **true/false propositions** A and B, and this is **undefined** if B is **false**; therefore

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

This **motivates** the following

Getting From the Context and Design to the Model

Definition: In model **specification**, **optimal** = {to come **as close as possible** to the **goal** of **conditioning** only on **propositions** rendered **true** by the **context** of the **study** and the **design** of the **data-gathering process**}.

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 3**:

	Number of Hospitalizations						
Group	0	1	...	k	n	Mean	SD
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 $p(\theta|\mathcal{B})$ (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** $p(D|\theta\,\mathcal{B})$.

The Calibration Principle

This is where a **good set of principles** starts to **help**: as a small **contribution** to **closing the gap** between **ad-hoc practice** and **lack of theory**, I'll focus in the rest of this **presentation** on **four principles** worth considering, the **first** of which is the

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

The **reasoning** behind the **Calibration Principle** is as follows:

(axiom) You want to **help positively advance** the **course of science**, and **repeatedly getting the wrong answer** runs **counter** to this desire.

(remark) There's **nothing** in the **Bayesian paradigm** to **prevent** You from making **one or both** of the following **mistakes** — (a) choosing $p(D|\theta\,\mathcal{B})$ **badly**; (b) inserting {**strong information** about θ **external** to D} into the **modeling process** that turns out **after the fact** to have been (badly) **out of step with reality** — and **repeatedly** doing this **violates the axiom** above.

Reasoning Behind the Calibration Principle

(remark) Paying attention to **calibration** is a **natural activity** from the **frequentist** point of view, but a **desire** to be **well-calibrated** can be given an **entirely Bayesian justification** via **decision theory**:

Taking a **broader perspective** over **Your career**, not just within any **single attempt** to solve an **inferential/predictive problem** in collaboration with **other investigators**, Your desire to take part **positively** in the **progress of science** can be **quantified** in a **utility function** that **incorporates** a **bonus** for being **well-calibrated**, and in this context (Draper, 2011) **calibration-monitoring** emerges as a **natural and inevitable Bayesian activity**.

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

But **the Calibration Principle** is **not enough**: in problems of **realistic complexity** You'll generally **notice** that (a) You're **uncertain** about θ

Model Uncertainty

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} (in a sense I'll make **precise** below).

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

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

To move toward an **answer** to this **question**, suppose (continuing the **Kaiser example** on page 15) that You also **observe** (for each of the $n =$ **112 randomly-sampled patients** from the **population** $\mathcal{P}$ of $N =$ **8,561 heart-attack patients**) a **real-valued conceptually-continuous**

Model Uncertainty (continued)

non-negative quality-of-care score y_i, and **inferential interest** focuses on the **mean** θ of these **scores** in $\mathcal{P}$; here the **data set** D is just $y = (y_1 \ldots y_n)$.

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

$$M_1\text{:} \left\{ \begin{array}{ccc} (\theta\, \sigma^2 | M_1\, \mathcal{B}) & \sim & p(\theta\, \sigma^2 | M_1\, \mathcal{B}) \\ (y_i | \theta\, \sigma^2\, M_1\, \mathcal{B}) & \overset{\text{IID}}{\sim} & \text{Gaussian}\big(\theta, \sigma^2\big) \end{array} \right\}, \tag{5}$$

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

$$M_2\text{:} \left\{ \begin{array}{ccc} (\theta\, \tau^2 | M_2\, \mathcal{B}) & \sim & p(\theta\, \tau^2 | M_2\, \mathcal{B}) \\ (y_i | \theta\, \tau^2\, M_2\, \mathcal{B}) & \overset{\text{IID}}{\sim} & \text{Lognormal}\big(\theta, \tau^2\big) \end{array} \right\}, \tag{6}$$

with the **Lognormal distribution parameterized** so that θ and τ^2 are the **mean** and **variance** on the y (rather than $\log y$) **scale**.

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

The Data-Generating Model M_{DG}

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

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

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

with (e.g.) $\left(\theta_{DG}, \sigma^2_{DG}\right) = (50, 10^2)$; I'll use the **notation** $\gamma_{DG} = (\theta_{DG}, \eta_{DG})$ for the **parameter vector** of M_{DG}.

Note that M_{DG} is a **single model** (e.g., $N(50, 10^2)$), not a **parametric family** of **single models** (e.g., $N(\mu, \sigma^2)$ with $-\infty < \mu < \infty$ and $0 \leq \sigma^2 < \infty$).

Rule 1, Revisited

The fact that M_{DG} is **unknown** to You presents a **challenge** in both **Bayesian** and **non-Bayesian** paradigms; the **form** this challenge takes in the **Bayesian approach** can be seen by recalling **Rule 1:**

• Choosing a **specific model** M_j amounts to **conditioning on it**; in other words, in practice You may **want** to compute $p(\theta|D\,\mathcal{B})$, but by choosing M_j You're **really computing** $p(\theta|D\,M_j\,\mathcal{B})$.

• Having **chosen** a particular **model** M_j (say), this makes me **wonder** what happens when $M_j \neq M_{DG}$, because in that case **choosing** M_j sounds like **conditioning on a false proposition**.

• However, it's **not quite meaningful** to write something like $M_j \neq M_{DG}$, because the **sampling-distribution** part of M_j actually contains **many models** from an M_{DG} **perspective**; in the **Gaussian-Lognormal** example above, for instance, M_{DG} specifies the **single model** $N(50, 10^2)$ but $p(y_i|\theta\,\sigma^2\,M_1\,\mathcal{B})$ specifies $N(\theta, \sigma^2)$ for **all** (θ, σ^2) in the **support** of the prior $p(\theta, \sigma^2|M_1\,\mathcal{B})$ (i.e., all (θ, σ^2) such that $p(\theta, \sigma^2|M_1\,\mathcal{B}) \neq 0$).

Asymptotic Consistency of Bayesian Inference

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

• This **theorem** motivates the following

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

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

So now the **correct wording of the question** is: what **happens** if I **choose** M_j but (**unknown to me**) $M_j \stackrel{c}{\neq} M_{DG}$?

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

Model Mis-Specification

Theorem (Berk, 1964): if $M_j \overset{c}{\neq} M_{DG}$, then as n **increases**, the **posterior distribution** $p(\theta|D\, M_j\, \mathcal{B})$ becomes **more and more concentrated** around {**point mass** at θ^*}, where $\gamma_j^* = (\theta^*, \eta_j^*)$ and θ^* is such that $p(D|\gamma_j^*\, M_j\, \mathcal{B})$ is as **close as possible** to $p(D|\gamma_{DG}\, M_{DG})$ (here **closeness** is measured by **Kullback-Leibler (KL) divergence**: for **densities** p and q, $D_{KL}(p||q) = \int p \log \frac{p}{q}$).

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

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

Having **introduced** this idea of a model M_j being **consistent** (or not) with an **underlying data-generating mechanism** M_{DG}, it would be

Dealing With Model Uncertainty

nice — from a **calibration** point of view — to be able to **compute** $p(\theta|D\,\mathcal{M}_c\,\mathcal{B})$, where $\mathcal{M}_c$ **includes all models** M_j such that $M_j \stackrel{c}{=} M_{DG}$;

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

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

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

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

(b) draw **inferences** about θ **pretending** that $M^* \stackrel{c}{=} M_{DG}$.

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; Draper, 1995) that

$$p(D^*|D\,\mathcal{M}\,\mathcal{B}) = \int_{\mathcal{M}} p(D^*|D\,M\,\mathcal{B})\,p(M|D\,\mathcal{M}\,\mathcal{B})\,dM\,, \qquad (8)$$

which is **eminently reasonable**: equation **(8)** 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**; for **implementation**, there are two `R` **packages** at `CRAN` — **BMA** and **BMS** — that perform **Bayesian model averaging** with a **wide variety** of **data configurations**.

- **Bayesian nonparametric (BNP) modeling:** The **BMA integral** in (8) can be thought of as an **approximation** to the **(unattainable?) ideal** of **averaging over all worthwhile models**; a **better approximation** to this **ideal** can often be achieved with **Bayesian nonparametric modeling**, which dates back to **de Finetti (1937)**.

Continuing the **Kaiser example** on page 15, suppose You also **observe** (for each of the $n =$ **112 randomly-sampled patients** from the **population** $\mathcal{P}$ of $N =$ **8,561 heart-attack patients**) a **real-valued conceptually-continuous quality-of-care score** y_i, and (following **de Finetti**) You're thinking about Your **predictive distribution**

Exchangeability

$p(y_1 \ldots y_n|\mathcal{B})$ for these scores **before any data have arrived**.

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

He (and later **Diaconis/Freedman**) went on to **prove** that, if Your judgment of **exchangeability** extends from $(y_1 \ldots y_n)$ to $(y_1 \ldots y_N)$ (as it certainly **should** here, given the **random sampling**) and $N \gg n$ (as is **true** here), then all **logically-internally-consistent predictive distributions** can **approximately** be expressed **hierarchically** as follows: letting G stand for the **empirical CDF** (see page 58) of the **population values** $(y_1 \ldots y_N)$, the **hierarchical model** is (for $i = 1, \ldots, n$)

$$\left\{ \begin{array}{rcl} (G|\mathcal{B}) & \sim & p(G|\mathcal{B}) \\ (y_i|G\,\mathcal{B}) & \overset{\text{IID}}{\sim} & G \end{array} \right\} . \tag{9}$$

This requires placing a **scientifically-appropriate prior distribution** $p(G|\mathcal{B})$ on the **set** $\mathcal{G}$ **of all CDFs** on $\Re$, which **de Finetti** didn't know

Bayesian Nonparametric (BNP) Modeling

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:

	Number of Hospitalizations						
Group	0	1	...	k	n	Mean	SD
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

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 1980**s) 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** (BQQI; Draper 2011): **exchangeability** implies a **multinomial sampling distribution** on the **qualitative outcome variable** with **category labels** 0, 1, . . . , and this permits **optimal model specification** here (this **approach** treats the **hospitalization outcome categorically** but permits **quantitative inference** about θ).

BQQI in the IHGA case study. de Finetti's **most basic theorem** about **exchangeability** says that if You're about to **observe** a **binary data set** $y = (y_1, \ldots, y_n)$ and You're willing to regard y as part of an **infinitely exchangeable sequence** (meaning that You judge all **finite subsets** exchangeable; this is like **thinking** of the y_i as having been **randomly sampled** from the **population** $(y_1, y_2, \ldots)$), then to be **logically internally consistent** Your joint predictive distribution $p(y_1, \ldots, y_n | \mathcal{B})$ must have the simple **hierarchical** form

$$(\theta | \mathcal{B}) \sim p(\theta | \mathcal{B}), \qquad (y_i | \theta \, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \tag{10}$$

Model = Prior (Sometimes)

where $\theta = P(y_i = 1|\mathcal{B})$ is the **limiting value** of the **mean** of the y_i in the **infinite sequence**.

Writing $s = (s_1, s_2)$ where s_1 and s_2 are the **numbers of 0s and 1s**, respectively, in $(y_1, \ldots, y_n)$, this is **equivalent** to the **model**

$$\begin{aligned} (\theta_2|\mathcal{B}) &\sim p(\theta_2|\mathcal{B}) \\ (s_2|\theta_2\,\mathcal{B}) &\sim \text{Binomial}(n, \theta_2), \end{aligned} \tag{11}$$

where (in a slight change of **notation**) $\theta_2 = P(y_i = 1|\mathcal{B})$; i.e., in this **simplest case** the **form** of the **likelihood function** (Binomial(n, θ_2)) is **determined** by a **desire** for **logical internal consistency**.

The **likelihood function** for θ_2 in this model is

$$l(\theta_2|y) = c\,\theta_2^{s_2}(1-\theta_2)^{n-s_2} = c\,\theta_1^{s_1}\theta_2^{s_2}, \tag{12}$$

from which it's **evident** that the **conjugate prior** for the **Bernoulli/Binomial likelihood** (the **choice of prior** having the **property** that the **posterior** for θ_2 has the same **mathematical form** as the **prior**) is the **family** of **Beta**(α_1, α_2) **densities**

$$p(\theta_2|\mathcal{B}) = c\,\theta_2^{\alpha_2-1}(1-\theta_2)^{\alpha_1-1} = c\,\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}, \tag{13}$$

BQQI (continued)

for some $\alpha_1 > 0, \alpha_2 > 0$.

With this **prior** the **conjugate updating rule** is **evidently**

$$\left\{ \begin{array}{c} (\theta_2|\mathcal{B}) \sim \text{Beta}(\alpha_1, \alpha_2) \\ (s_2|\theta_2\,\mathcal{B}) \sim \text{Binomial}(n, \theta_2) \end{array} \right\} \rightarrow (\theta_2|y\,\mathcal{B}) \sim \text{Beta}(\alpha_1 + s_1, \alpha_2 + s_2), \tag{14}$$

where s_1 (s_2) is the **number of 0s (1s)** in the **data set** $y = (y_1, \ldots, y_n)$.

Moreover, given that the **likelihood** represents a **(sample) data set** with s_1 0s and s_2 1s and a **data sample size** of $n = (s_1 + s_2)$, it's clear that

(a) the **Beta**(α_1, α_2) prior acts like a **(prior) data set** with α_1 0s and α_2 1s and a **prior sample size** of $(\alpha_1 + \alpha_2)$, and

(b) to achieve a relatively **diffuse (low-information-content)** prior for θ_2 (if that's what **context** suggests You should aim for) You should try to specify α_1 and α_2 **not far from 0**.

Here's an **easy generalization** of all of this: **suppose** the y_i take on $J \geq 2$ **distinct values** $v = (v_1, \ldots, v_J)$, and let $s = (s_1, \ldots, s_J)$ be the **vector** of **counts** ($s_1 = \#(y_i = v_1)$ and so on).

Multinomial Likelihood

If You **judge** the y_i to be part of an **infinitely exchangeable sequence**, then to be **logically internally consistent** Your **joint predictive distribution** $p(y_1, \ldots, y_n|\mathcal{B})$ must have the **hierarchical** form

$$\begin{aligned} (\theta|\mathcal{B}) &\sim p(\theta|\mathcal{B}) \\ (s|\theta\,\mathcal{B}) &\sim \text{Multinomial}(n, \theta), \end{aligned} \tag{15}$$

where $\theta = (\theta_1, \ldots, \theta_J)$ and θ_j is the **limiting relative frequency** of v_j values in the **infinite sequence**.

The **likelihood** for (vector) θ in this case has the **form**

$$l(\theta|y) = c \prod_{j=J}^{I} \theta_j^{s_j}, \tag{16}$$

from which it's **evident** that the **conjugate prior** for the **Multinomial likelihood** is of the **form**

$$p(\theta|\mathcal{B}) = c \prod_{j=1}^{J} \theta_j^{\alpha_j - 1}, \tag{17}$$

for some $\alpha = (\alpha_1, \ldots, \alpha_J)$ with $\alpha_j > 0$ for $j = 1, \ldots, J$;

Dirichlet Conjugate Prior

this is the **Dirichlet**(α) distribution, a **multivariate generalization** of the **Beta family**.

Here the **conjugate updating rule** is

$$\left\{ \begin{array}{c} (\theta|\mathcal{B}) \sim \text{Dirichlet}(\alpha) \\ (s|\theta\,\mathcal{B}) \sim \text{Multinomial}(n,\theta) \end{array} \right\} \rightarrow (\theta|y\,\mathcal{B}) \sim \text{Dirichlet}(\alpha + s), \quad (18)$$

where $s = (s_1, \ldots, s_J)$ and s_j is the **number of** v_j **values** $(j = 1, \ldots, J)$ in the data set $y = (y_1, \ldots, y_n)$.

Furthermore, by **direct analogy** with the $J = 2$ case,

(a) the **Dirichlet**(α) prior acts like a **(prior) data set** with α_j v_j values $(j = 1, \ldots, J)$ and a **prior sample size** of $\sum_{j=1}^{J} \alpha_j$, and

(b) to achieve a relatively **diffuse (low-information-content)** prior for θ (if that's what **context** suggests You should aim for) You should try to choose all of the α_j **not far from 0**.

To **summarize**:

(A) if the **data vector** $y = (y_1, \ldots, y_n)$ takes on J **distinct** values

Dirichlet-Multinomial Modeling

$v = (v_1, \ldots, v_J)$ (**real numbers or not**) and You judge (Your uncertainty about) the **infinite sequence** $(y_1, y_2, \ldots)$ to be **exchangeable**, then (by a **representation theorem** of de Finetti) **logical internal consistency** compels You (i) to **think about** the **quantities** $\theta = (\theta_1, \ldots, \theta_J)$, where θ_j is the **limiting relative frequency** of the v_j values in the **infinite sequence**, and (ii) to **adopt** the **Multinomial model**

$$\begin{aligned} (\theta|\mathcal{B}) &\sim p(\theta|\mathcal{B}) \\ p(y_i|\theta\,\mathcal{B}) &= c\prod_{j=1}^{J}\theta_j^{s_j}, \end{aligned} \tag{19}$$

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

(B) if **context** specifies a **diffuse** prior for θ, a convenient (**conjugate**) choice is **Dirichlet**(α) with $\alpha = (\alpha_1, \ldots, \alpha_J)$ and all of the α_j **positive but close to 0**; and

(C) with a **Dirichlet**(α) prior for θ the **posterior** is **Dirichlet**(α'), where $s = (s_1, \ldots, s_J)$ and $\alpha' = (\alpha + s)$.

Parametric Modeling on the y_i Not Required Here

Note, **remarkably**, that the v_j values themselves **make no appearance** in the model; this **modeling approach** is **natural** with **categorical outcomes** but **can also be used** when the v_j are **real numbers**.

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

This **fact** makes it **possible** to draw an **analogy** with the **distribution-free** methods that are at the heart of **frequentist non-parametric** inference: when Your **outcome variable** takes on a **finite number** of **real** values v_j, **exchangeability** compels a **Multinomial likelihood** on the **underlying frequencies** with which the v_j occur; You're **not required** to build a **parametric model** (e.g., normal, lognormal, ...) on the y_i **values** themselves.

In this **sense**, therefore, **model (19)** — particularly with the **conjugate Dirichlet** prior — can serve as a kind of **Bayesian qualitative-quantitative inference** (this is related to the **Bayesian bootstrap** (Rubin 1981)).

Dirichlet Sampling

Moreover, if You're **in a hurry** and You're **already familiar** with `WinBUGS`, You can **readily** carry out **inference** about quantities like μ_y above in this **environment**, but there's **no need to do MCMC** here: **ordinary Monte Carlo** (MC) sampling from the **Dirichlet**(α') **posterior distribution** is perfectly **straightforward**, e.g., in `R`, based on the following **fact**:

To **generate** a **random draw** $\theta = (\theta_1, \ldots, \theta_J)$ from the **Dirichlet**(α') distribution, with $\alpha' = (\alpha'_1, \ldots, \alpha'_J)$, **independently draw**

$$g_j \overset{\text{indep}}{\sim} \Gamma(\alpha'_j, \beta), \quad j = 1, \ldots, J \tag{20}$$

(where $\Gamma(a, b)$ is the **Gamma distribution** with parameters a and b) and **compute**

$$\theta_j = \frac{g_j}{\sum_{m=1}^{J} g_j} \, . \tag{21}$$

Any $\beta > 0$ **will do** in this **calculation**; $\beta = 1$ is a **good choice** that leads to **fast random number generation**.

Dirichlet Sampling (continued)

The **downloadable version** of R doesn't have a **built-in function** for making **Dirichlet draws** (although **packages** like MCMCpack at CRAN do have such **functions**), but it's **easy** to **write** one:

```
rdirichlet <- function( n.sim, alpha ) {
  J <- length( alpha )
  theta <- matrix( 0, n.sim, J )
  for ( j in 1:J ) {
    theta[ , j ] <- rgamma( n.sim, alpha[ j ], 1 )
  }
  theta <- theta / apply( theta, 1, sum )
  return( theta )
}
```

The **Dirichlet**(α) distribution has the following **moments**: if $\theta \sim$ Dirichlet(α) then

$$E(\theta_j) = \frac{\alpha_j}{\alpha_0}, \; V(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}, \; C(\theta_j, \theta_{j'}) = -\frac{\alpha_j \alpha_{j'}}{\alpha_0^2(\alpha_0 + 1)}, \quad (22)$$

where $\alpha_0 = \sum_{j=1}^{J} \alpha_j$ (note the **negative correlation** between components of θ).

Dirichlet Sampling (continued)

This can be **used** to **test** the **function** above:

```
alpha <- c( 5.0, 1.0, 2.0 )
alpha.0 <- sum( alpha )
test <- rdirichlet( 100000, alpha ) # 7 seconds (1.6 Unix GHz)
apply( test, 2, mean )
[1] 0.6258544 0.1247550 0.2493905
alpha / alpha.0
[1] 0.625 0.125 0.250
apply( test, 2, var )
[1] 0.02603293 0.01216358 0.02071587
alpha * ( alpha.0 - alpha ) / ( alpha.0^2 * ( alpha.0 + 1 ) )
[1] 0.02604167 0.01215278 0.02083333
cov( test )
              [,1]          [,2]          [,3]
[1,]  0.026032929 -0.008740319 -0.017292610
[2,] -0.008740319  0.012163577 -0.003423259
[3,] -0.017292610 -0.003423259  0.020715869
```

BQQI Analysis of IHGA Data

```
- outer( alpha, alpha, "*" ) / ( alpha.0^2 * ( alpha.0 + 1 ) )
              [,1]         [,2]         [,3]
[1,] -0.043402778 -0.008680556 -0.017361111
[2,] -0.008680556 -0.001736111 -0.003472222 # ignore diagonals
[3,] -0.017361111 -0.003472222 -0.006944444
```

BQQI analysis of the IHGA data: recall that the **policy** and **clinical interest** focused on $\eta = \frac{\mu_E}{\mu_C}$; here's the **data**:

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

In this **two-independent-samples** setting You can apply de Finetti's representation theorem **twice, in parallel**, on the C and E data.

Not much is known about the **underlying frequencies** of $0, 1, \ldots, 7$ hospitalizations under C and E **external** to the data, so You can use a **Dirichlet**$(\epsilon, \ldots, \epsilon)$ **prior** for both θ_C and θ_E with $\epsilon = 0.001$, leading to a

BQQI Analysis of IHGA Data (continued)

Dirichlet$(138.001, \ldots, 2.001)$ **posterior** for θ_C and a **Dirichlet**$(147.001, \ldots, 0.001)$ **posterior** for θ_E (other small positive choices of ϵ yield **similar results**).

```
alpha.C <- c( 138.001, 77.001, 46.001, 12.001, 8.001,
  4.001, 0.001, 2.001 )
alpha.E <- c( 147.001, 83.001, 37.001, 13.001, 3.001,
  1.001, 1.001, 0.001 )
theta.C <- rdirichlet( 100000, alpha.C ) # 8 sec (1.6 Unix GHz)
theta.E <- rdirichlet( 100000, alpha.E ) # also 8 sec
print( post.mean.theta.C <- apply( theta.C, 2, mean ) )
[1] 4.808015e-01 2.683458e-01 1.603179e-01 4.176976e-02
[5] 2.784911e-02 1.395287e-02 3.180905e-06 6.959859e-03
print( post.SD.theta.C <- apply( theta.C, 2, sd ) )
[1] 0.0294142963 0.0261001259 0.0216552661 0.0117925465
[5] 0.0096747630 0.0069121507 0.0001017203 0.0048757485
print( post.mean.theta.E <- apply( theta.E, 2, mean ) )
[1] 5.156872e-01 2.913022e-01 1.298337e-01 4.560130e-02
[5] 1.054681e-02 3.518699e-03 3.506762e-03 3.356346e-06
print( post.SD.theta.E <- apply( theta.E, 2, sd ) )
[1] 0.029593047 0.026915644 0.019859213 0.012302252
[5] 0.006027157 0.003501568 0.003487824 0.000111565
```

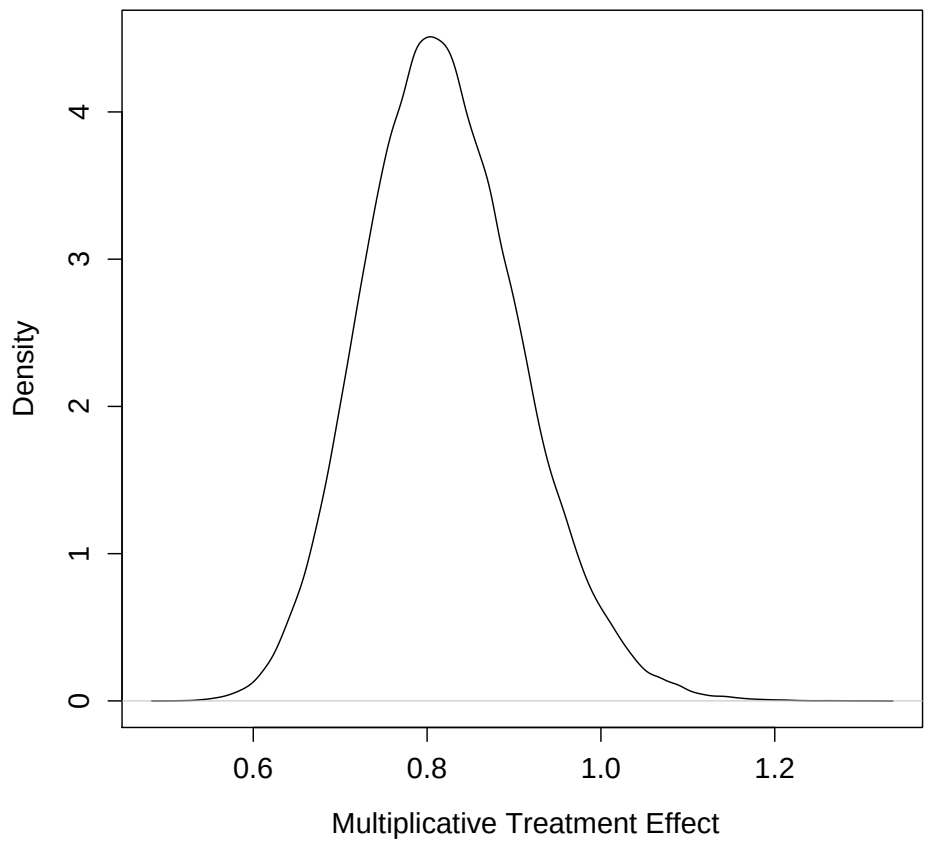
BQQI IHGA Results

```
mean.effect.C <- theta.C %*% ( 0:7 )
mean.effect.E <- theta.E %*% ( 0:7 )
mult.effect <- mean.effect.E / mean.effect.C
print( post.mean.mult.effect <- mean( mult.effect ) )
[1] 0.8189195
print( post.SD.mult.effect <- sd( mult.effect ) )
[1] 0.08998323
quantile( mult.effect, probs = c( 0.0, 0.025, 0.5,
  0.975, 1.0 ) )
      0%      2.5%       50%     97.5%      100%
0.5037150 0.6571343 0.8138080 1.0093222 1.3868332
mean( mult.effect < 1 )
[1] 0.9706
pdf( "bqqi-mult-effect.pdf" )
plot( density( mult.effect, n = 2048 ), type = 'l',
  cex.lab = 1.25, cex.axis = 1.25, main = '',
  xlab = 'Multiplicative Treatment Effect' )
dev.off( )
```

You would **estimate** that **IHGA reduces** the **mean number of hospitalizations** per two years (in the **population** $\mathcal{P}$ of **all elderly**

BQQI IHGA Results (continued)

Danish non-institutionalized people) by about $100(1 - 0.8189195)\%$ $\doteq$ **18%**, give or take about $100(0.08998323)\%$ $\doteq$ **9%**, with a **95% interval estimate** of **(–0.9,34.2)%**; the **posterior probability** that **IHGA would be beneficial in** $\mathcal{P}$ is **estimated** to be about $100(0.9706)$ $\doteq$ **97%**; and the **posterior distribution** for the **multiplicative effect** of **IHGA** is:



Optimal Bayesian Model Specification

In my view **this analysis completely satisfies** the **criterion** for **optimal Bayesian model specification**: it **conditions** only on **propositions rendered true** by the **study design** and **data-gathering process**.

NB I don't yet know much (and I don't think other people do either) about the **generalizability** of this finding, except to say that **more care** may be required to choose an **appropriate function** of the θ values when the y_i are **closer to continuous**.

Another approach: Bayesian nonparametric modeling. **Recall** back on **page 40** that if You have a **real-valued data set** $y = (y_1, \ldots, y_n)$ **drawn exchangeably (like a random sample)** from $\mathcal{P} = (y_1, y_2, \ldots)$, then all **logically-internally-consistent models** are of the **form**

$$\left\{ \begin{array}{rcl} (G|\mathcal{B}) & \sim & p(G|\mathcal{B}) \\ (y_i|G\,\mathcal{B}) & \overset{\text{IID}}{\sim} & G \end{array} \right\} \tag{23}$$

for some **prior distribution** $p(G|\mathcal{B})$ on the **set** $\mathcal{G}$ of **all CDFs** on $\Re$; how can such a **prior distribution** be **specified** in a way that's **responsive** to the **science** of the **problem**?

Dirichlet Process Priors

This **question** was **answered** by Ferguson (1973), who created **Dirichlet Process** priors; here's his **reasoning**.

• **Conjugate priors** are **nice**, for **two reasons**:

(a) by **definition** the **prior** and **posterior** have the **same form**, and

(b) the **prior** is **driven exclusively** by **two ingredients** —

(i) a **prior estimate** (in this case G_{prior}) of the **unknown quantity** G and

(ii) a **prior sample size** n_{prior} indicating the **amount** of available **information external** to the **sample data set** (the **prior acts** like a **prior data set** which, when **merged** with the **sample data set**, yields a **"posterior data set"** that can be **analyzed** by **maximum-likelihood methods** to produce the **same answer** as the **Bayesian analysis** with the **indicated prior** and **sample data set**).

• It would be **nice**, therefore, to be able to **create** a **prior** $NPP(\cdot, \cdot)$ on $\mathcal{G}$ that has this **conjugate behavior**:

The Empirical CDF

$$\left\{ \begin{array}{c} (G|\mathcal{B}) \sim NPP(n_{prior}, G_{prior}) \\ (y_i|G\,\mathcal{B}) \overset{\text{IID}}{\sim} G \end{array} \right\} \rightarrow (G|y\,\mathcal{B}) \sim NPP(n_{posterior}, G_{posterior}) \,,$$

where $n_{posterior} = (n_{prior} + n)$ and where $G_{posterior}$ is **related** in some **natural way** to (i) G_{prior} and (ii) a **good estimate** of G based **solely** on the **data**.

- With $(y_i|G\,\mathcal{B}) \overset{\text{IID}}{\sim} G$ it can be **shown** that the **nonparametric maximum likelihood estimator** of G is the **empirical CDF**

$$\hat{G}_n(t) = \frac{(\text{number of } y_i) \leq t}{n} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq t) \,, \tag{24}$$

where $I(A) = 1$ if **proposition** A is **true** and 0 **otherwise**; this can serve as the **"good estimate of G based solely on the data."**

- If You **take an IID sample** of size n from $\hat{G}_n$ (i.e., an **IID sample** from $y = (y_1, \ldots, y_n)$ **itself**; this should **remind** You of the **bootstrap**), obtaining $y^* = (y_1^*, \ldots, y_n^*)$, and You **keep track** of **how many replicates** of **each observation** You see, the **result** will follow a **multinomial distribution**, as follows: **sort** the y_i into $k \leq n$ **bins**

Dirichlet Process Priors (continued)

$(b_1, \ldots, b_k)$ ($k < n$ if there are **ties**) and let $n_j =$ (the **number** of y_i in bin b_j); then

$$(y^*|\hat{G}_n\,\mathcal{B}) \sim \hat{G}_n \quad \text{iff} \quad p(y^*|\hat{G}_n\,\mathcal{B}) = c\,\theta_1^{n_1}\cdots\theta_k^{n_k}, \tag{25}$$

for some $\theta = (\theta_1, \ldots, \theta_k)$ **such that** $\theta_j \geq 0$ for all $1 \leq j \leq k$ and $\sum_{j=1}^{k} \theta_j = 1$.

• As **noted** in the **BQQI section**, the **conjugate prior** for the **multinomial sampling distribution** is the **Dirichlet distribution**: with $\alpha = (\alpha_1, \ldots, \alpha_k)$ **such that** $\alpha_j \geq 0$ for all $1 \leq j \leq k$,

$$(\theta|\mathcal{B}) \sim \textbf{Dirichlet}(\alpha) \quad \text{iff} \quad p(\theta|\mathcal{B}) = c\,\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}. \tag{26}$$

• **Choose** any $(k-1)$ **distinct points** $-\infty < r_1 < \cdots < r_{k-1} < \infty$ on the **real line** — these define a **partition** $(A_1, \ldots, A_k)$ of the **line**, where $A_1 = (-\infty, r_1]$, $A_2 = (r_1, r_2]$, and so on — and imagine **dropping** all of the **data values** $(y_1, \ldots, y_n)$ onto the **line** and **counting** how many **fall** in $(A_1, \ldots, A_k)$: if You now **think** of a **density** g on $\Re$ and **compute** θ_j as the **mass** g **assigns** to **partition element** A_j, the **counts** will follow a **multinomial distribution** with **probabilities** $(\theta_1, \ldots, \theta_k)$.

Dirichlet Process Definition and Conjugate Updating

- To **create** a **probability distribution** such that **random draws** from it are **CDFs**, Ferguson therefore **defined** the **Dirichlet process** as follows, e.g., for **CDFs** on $\Re$ and with G having a **density** g:

Definition: CDF $G \sim DP(\alpha)$ (i.e., G follows a **Dirichlet process** with **parameter** α, where α is itself a **distribution**) iff for any **partition** $(A_1, \ldots, A_k)$ of $\Re$, the **random vector** $[G(A_1), \ldots, G(A_k)]$ follows a **Dirichlet distribution** with **parameter** $[\alpha(A_1), \ldots, \alpha(A_k)]$, where $G(A_j)$ means the **mass** assigned to A_j by g.

It's **useful** to **express** α in the **form** $\alpha(\cdot) = c\, G_0(\cdot)$, where G_0 is the **centering** or **base distribution** (the **prior estimate**) — in the **sense** that $E_{DP(cG_0)}(G) = G_0$ — and c acts like a **prior sample size**.

With this way of **writing** α, **conjugate updating** becomes **clear**:

$$\left\{ \begin{array}{c} (G|\mathcal{B}) \sim DP(c\, G_0) \\ (y_i|G\,\mathcal{B}) \overset{\text{IID}}{\sim} G \end{array} \right\} \rightarrow (G|y\,\mathcal{B}) \sim DP(c^*\, G^*)\,, \qquad (27)$$

where $c^* = (c + n)$ and $G^* = \frac{c\, G_0 + n\, \hat{G}_n}{c+n}$; thus G^* is a **weighted average** of G_0 and $\hat{G}_n$, with **weights** given by the **prior sample size** and

data sample size, respectively.

It **turns out** that **DP priors** put **all their mass** on **discrete CDFs**, which would be **OK** for the **IHGA data**; **one possible Bayesian nonparametric (BNP) model** for this **data set** (Krnjajić, Kottas and Draper 2008, on the **course web page**) would involve placing **parallel DPs priors** on the **population** C and T CDFs (below I'll describe **another BNP model** in the **IHGA example**).

However, You can get **as close as You like** to **any continuous CDF** through a **mixture of discrete CDFs**; this **observation** has **given rise** to **Dirichlet Process mixture modeling (DPMM)**, which is **more common** than just **putting a DP prior** directly on the **scale** of the **data**.

In fact, **mixture modeling** is **needed anyway** to make **DP priors** truly **useful**, as follows.

Consider the **Kaiser study** mentioned earlier, where of n **patient records** were **chosen randomly**, each of which **yielded** a **real-valued quality of care score** y_i.

DP Mixture Modeling (continued)

Before the **data** arrives Your **uncertainty** about $y = (y_1, \ldots, y_n)$ is **exchangeable**, so (switching **notation**, here and below: the **prior sample size hyperparameter** in the **DP prior** is usually **denoted** (a bit **confusingly**) by α) You want to **use** the **model**

$$\begin{aligned} (G|\mathcal{B}) &\sim DP(\alpha\, G_0) \\ (y_i|G\,\mathcal{B}) &\overset{\text{IID}}{\sim} G\,, \end{aligned} \tag{28}$$

but **what** should You **use** for α and G_0?

Suppose that **previous studies** suggest that **quality of care scores** may be **approximately Gaussian**, but (in case that's **not true** for **Your data set**) You **don't want** to be **dogmatic** about this; then You should **take** $G_0 =$ **Gaussian** and α **positive** but **rather small**; however, **which Gaussian**?

You **don't want** to take **something like** $G_0 = N(141, 63.7^2)$ because (at **design time**) You **don't know** what the **mean** and **SD** will be; You would **rather take** $G_0 = N(\mu, \sigma^2)$ with μ and σ^2 **unknown** and to be **learned** from the **data**.

Pólya Trees

DP mixture modeling allows You to **do this**: the **model** becomes

$$\begin{aligned} (G|\mathcal{B}) &\sim DP(\alpha\, G_0), \quad G_0 = N(\mu, \sigma^2) \\ (y_i|G\,\mathcal{B}) &\overset{\text{IID}}{\sim} G \\ (\mu, \sigma^2|\mathcal{B}) &\sim p(\mu, \sigma^2|\mathcal{B}) \\ (\alpha|\mathcal{B}) &\sim p(\alpha|\mathcal{B})\,. \end{aligned} \tag{29}$$

Models like this (and **considerably more complicated models** with **unknown CDFs**) are **fit via MCMC**, for **instance** with the `CRAN` **package** `DPpackage` in `R` (I'll give an **example** of the **use** of **this package** below).

It can be **non-trivial** to **choose** $p(\mu, \sigma^2|\mathcal{B})$ and $p(\alpha|\mathcal{B})$ to **get** both (a) **well-calibrated results** and (b) **MCMC chains** that **mix well**; the **best way** to **solve this problem** is through **experience** in **settings** where You **know the right answer**.

Another approach: Pólya trees. **Lavine (1992)** developed **another approach** to creating **priors on CDFs** called **Pólya trees**, which turn out to **include DP priors** as a **special case**; here's the **idea**.

Example: NB10. In **1962** and **1963** (Freedman, Pisani and Purves, 1978), two **employees** of the U.S. **National Bureau of Standards** (now called the **National Institute of Standards and Technology**) made $n = 100$ **weighings** of a **block of metal** called **NB10** — given this **name** because it was **supposed** to weigh **10 grams** — under **conditions** that were **as close as humanly possible** to the **statistical ideal** of **independent, identically distributed (IID) sampling** from the **population** $\mathcal{P}_{NB10} = \{$**all possible weighings** of **NB10** with the **given apparatus**$\}$.

Here the **unknown** θ of **principal interest** is **evidently** the **"true" weight** of **NB10**, by which I mean the **population mean** of $\mathcal{P}_{NB10}$; D consists of the **100 weighings** $y = (y_1, \ldots, y_n)$; and $\mathcal{B}$ **contains** the **proposition** (y is an **IID sample** from $\mathcal{P}_{NB10}$) (along with **background propositions** known to be **true** from the **context** of the **problem**, such as $(\theta > 0\}$ and $\{\theta$ is **close** to 10 grams)).

In a **situation** where You would **use** a $DP(\alpha\, G_0)$ prior, **Lavine's approach** instead **yields** the **Pólya tree prior** $PT(\Pi, \mathcal{A}_\alpha)$, where (a bit **confusingly**) $\mathcal{A}_\alpha$ **plays** the **role** of α and Π **acts** like G_0.

Pólya Trees (continued)

For the **NB10 data** a **natural Pólya-tree model** would be

$$\begin{aligned} (G|\mathcal{B}) &\sim PT(\Pi, \mathcal{A}_\alpha), \quad N(\mu, \sigma^2) \text{ determines } \Pi \\ (y_i|G\,\mathcal{B}) &\overset{\text{IID}}{\sim} G \\ (\mu, \sigma^2|\mathcal{B}) &\sim p(\mu, \sigma^2|\mathcal{B}) \\ (\alpha|\mathcal{B}) &\sim p(\alpha|\mathcal{B})\,. \end{aligned} \tag{30}$$

Here (a) $\Pi = \{B_\epsilon\}$ is a **binary tree partition** of the **real line**, where ϵ is a **binary sequence** that **locates** the **set** B_ϵ in the **tree**.

You get to **choose** these **sets** B_ϵ in a way that **centers the Pólya tree** on **any distribution you want**, in this case a **Gaussian** with **unknown mean and SD**.

This is done by **choosing** the **cutpoints** on the **line**, which **define** the **partitions**, based on the **quantiles** of $N(\mu, \sigma^2)$; for **example**, with $G_0 = N(0, 1)$, You get the **table** at the **top** of the **next page**, in which Φ is the $N(0, 1)$ CDF.

In practice this process has to stop somewhere; people use a tree **defined** down to **level** M, which is like working with **random histograms**,

Pólya Trees (continued)

each with 2^M **bars**.

Level	Sets	Cutpoint(s)
1	(B_0, B_1)	$\Phi^{-1}\left(\frac{1}{2}\right) = 0$
2	$(B_{00}, B_{01}, B_{10}, B_{11})$	$\Phi^{-1}\left(\frac{1}{4}\right) = -0.674, \Phi^{-1}\left(\frac{1}{2}\right) = 0, \Phi^{-1}\left(\frac{3}{4}\right) = +0.674$
$\vdots$	$\vdots$	$\vdots$

And (b) **Walker** et al. (1998; **emphasis** added):

"A **helpful image** is that of a **particle cascading through the partitions** B_ϵ. It **starts** [on the real line] and **moves** into B_0 with **probability** C_0 or into B_1 with **probability** $C_1 = 1 - C_0$. In **general**, on **entering** B_ϵ the **particle** could either **move** into $B_{\epsilon 0}$ or into $B_{\epsilon 1}$; let it **move** into the **former** with **probability** $C_{\epsilon 0}$ or into the **latter** with **probability** $C_{\epsilon 1} = 1 - C_{\epsilon 0}$. For **Pólya trees**, these **probabilities** are **Beta random variables**, $(C_{\epsilon 0}, C_{\epsilon 1}) \sim$ **Beta**$(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ with **non-negative** $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$. If we **denote** the **collection** of αs by $\mathcal{A}_\alpha$, a particular **Pólya tree distribution** is **completely defined** by Π and $\mathcal{A}_\alpha$."

To make a **Pólya tree prior** choose a **continuous distribution** with **probability 1**, the αs have to **grow quickly** as the **level** m of the **tree**

Pólya Trees (continued)

increases; following **Walker** et al. (1998) it's **common** to **take**

$$\alpha_\epsilon = \alpha\, m^2 \text{ whenever } \epsilon \text{ defines a set at level } m, \tag{31}$$

and this **defines** $\mathcal{A}_\alpha$.

As with **DP priors**, $\alpha > 0$ acts like a **prior sample size**: with **small** α the **posterior distribution** for G will be **based almost completely** on $\hat{G}_n$, the **empirical CDF** (the **"data distribution"**), whereas with **large** α the posterior will be **based almost completely** on the **prior centering distribution**, in this **case** $N(\mu, \sigma^2)$.

NB When the **data-generating distribution** is **multi-modal**, (somewhat **confusingly**) the α **hyper-parameter** in **DP priors** acts both as a **prior sample size** and a **prior indication** of **how many clusters (local modes)** the **distribution** has, but this **interpretational confusion doesn't occur** with **PT priors**.

Prior to posterior updating is **easy** with **Pólya trees**: if

$$\begin{aligned} (G|\mathcal{B}) &\sim PT(\Pi, \mathcal{A}_\alpha) \\ (y_i|G\,\mathcal{B}) &\overset{\text{IID}}{\sim} G \end{aligned} \tag{32}$$

Pólya Trees (continued)

and (say) y_1 is **observed**, then the **posterior** $p(G|y_1)$ for G given y_1 is **also a Pólya tree** — so the **PT priors** are again **conjugate** — with

$$(\alpha_\epsilon | y_1 \mathcal{B}) = \left\{ \begin{array}{cc} \alpha_\epsilon + 1 & \text{if } y_1 \in B_\epsilon \\ \alpha_\epsilon & \text{otherwise} \end{array} \right\} . \tag{33}$$

In other words the **updating** follows a **Pólya urn scheme** (e.g., Feller, 1968): at each **level** of the **tree**, if y_1 **falls** into a particular **partition set** B_ϵ, then **1 is added** to the α for that **set**.

The **graphs** on the **next few pages**

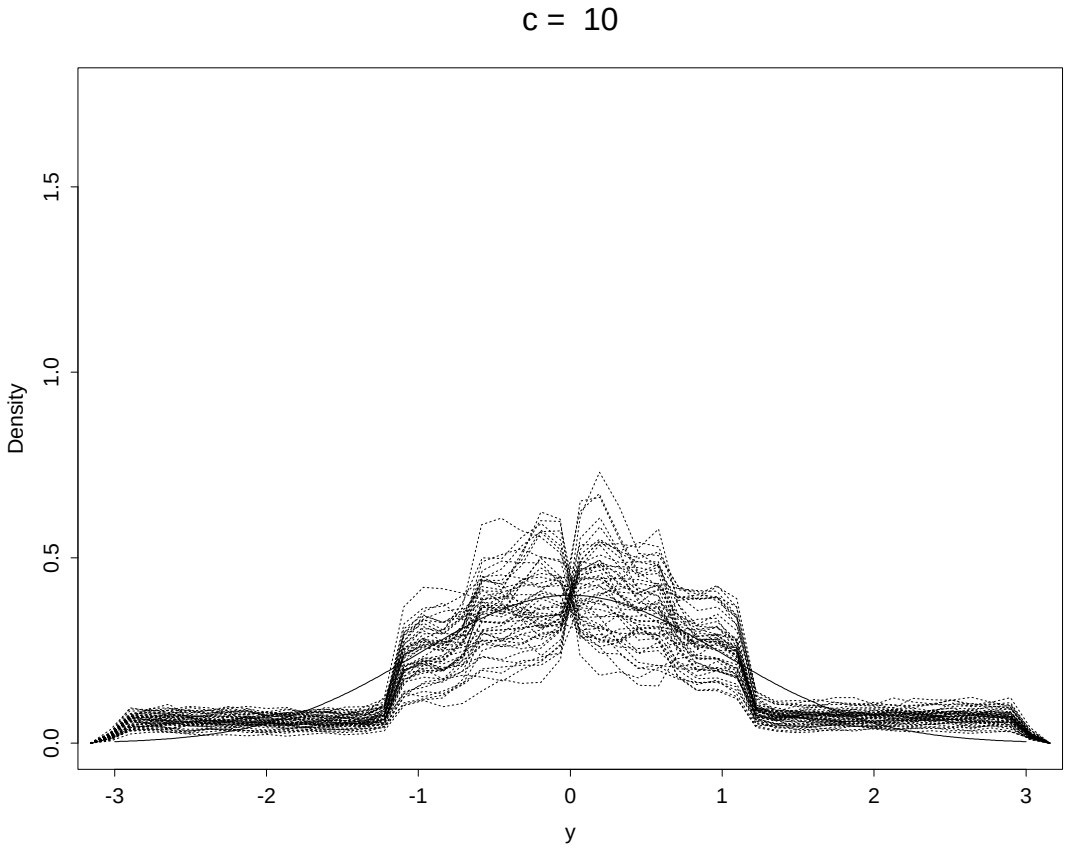
(i) show the **variation** around $N(0,1)$ **obtained** by **sampling** from a $PT(\Pi_{N(0,1)}, \mathcal{A}_\alpha)$ prior for G as α varies from 10 down to 0.1, and

(ii) illustrate **prior-to-posterior updating** for the **same range** of α with a **fairly skewed data set**.

`R` **code** to **perform** these **Pólya-tree simulations** is on the **course web page**.

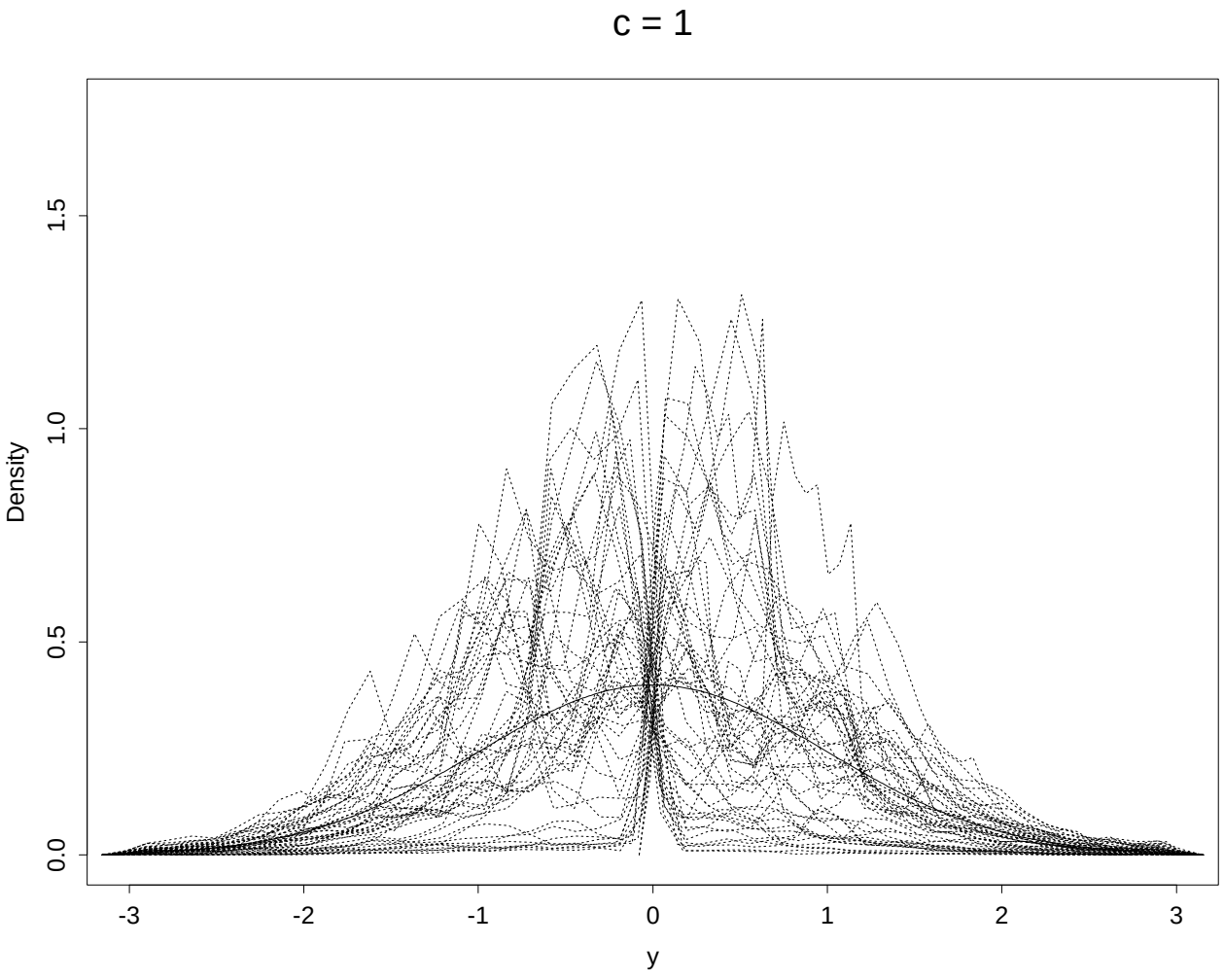
Sampling From a PT Prior

NB For the **next few pages**, the **main titles** of the **plots** say c when they **mean** α.



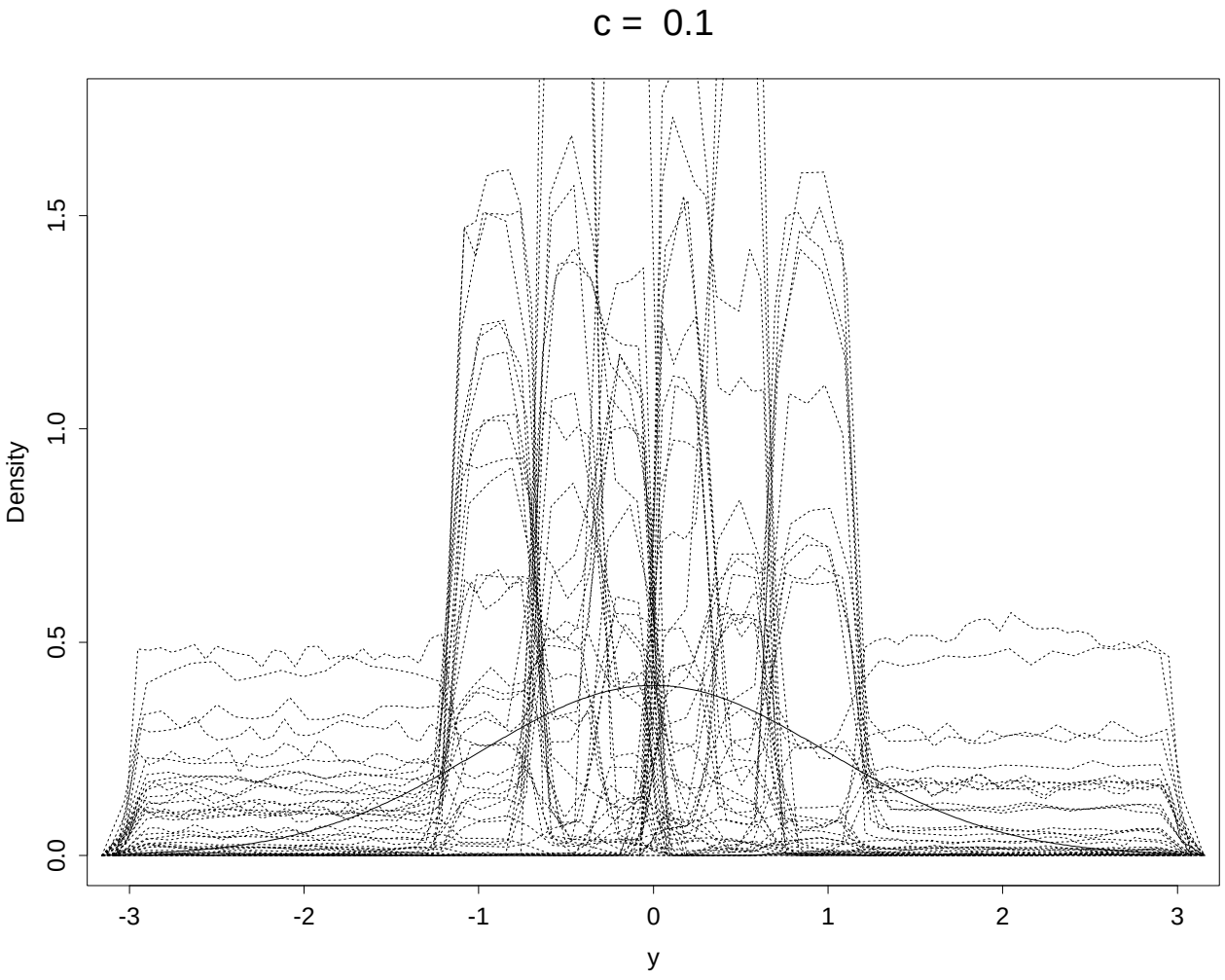
Sampling from a $PT(\Pi, \mathcal{A}_\alpha)$ **prior** for G **centered** at $N(0,1)$ **(solid line)** with $\alpha = 10$; for **large** α the **sampled distribution** follows the **prior** pretty **closely**.

Sampling From a PT Prior (continued)



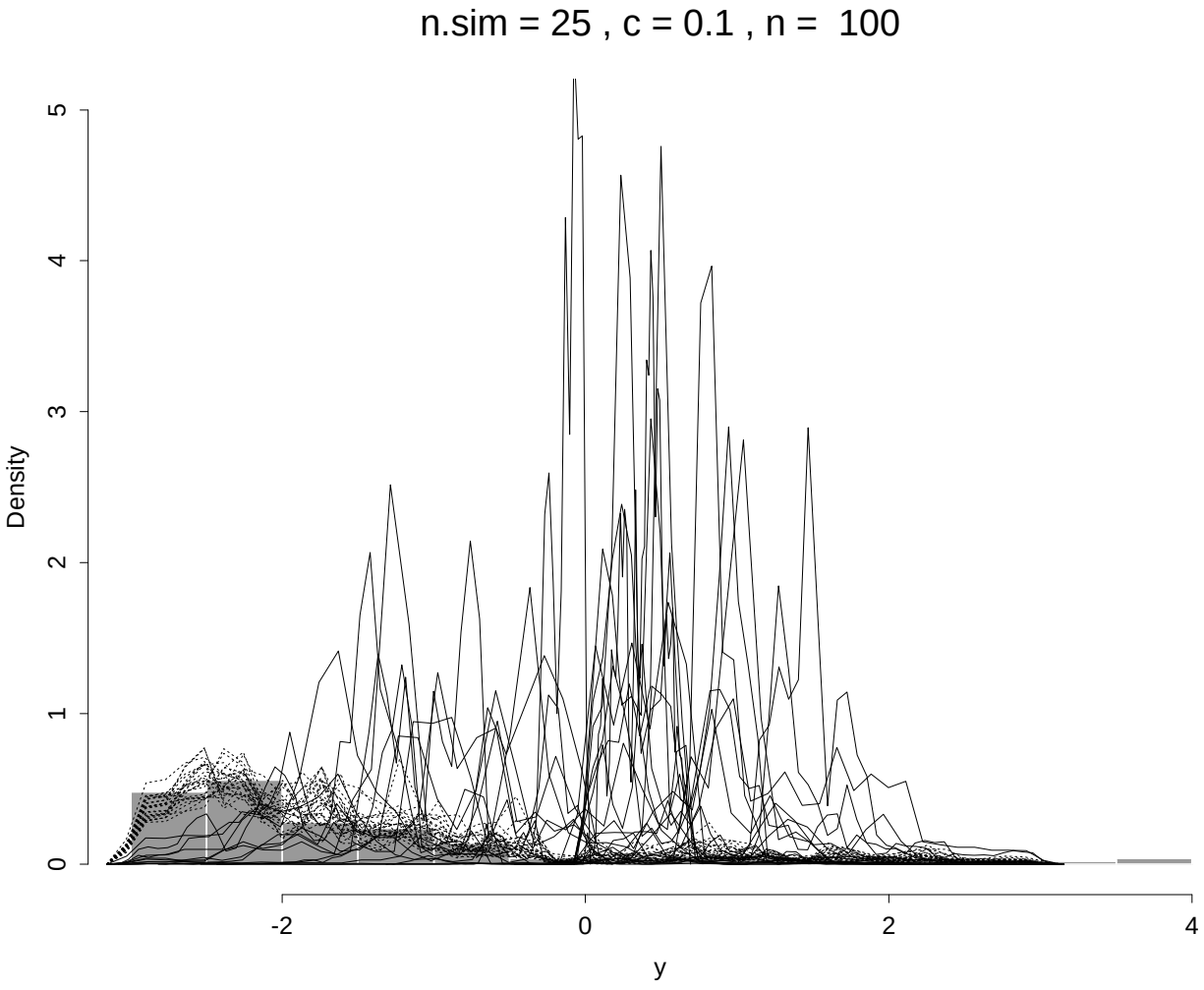
Sampling from a $PT(\Pi, \mathcal{A}_\alpha)$ **prior** for G **centered** at $N(0,1)$ **(solid line)** with $\alpha = 1$; the **sampled** Gs **vary more** around $N(0,1)$ with a **smaller** α.

Sampling From a PT Prior (continued)



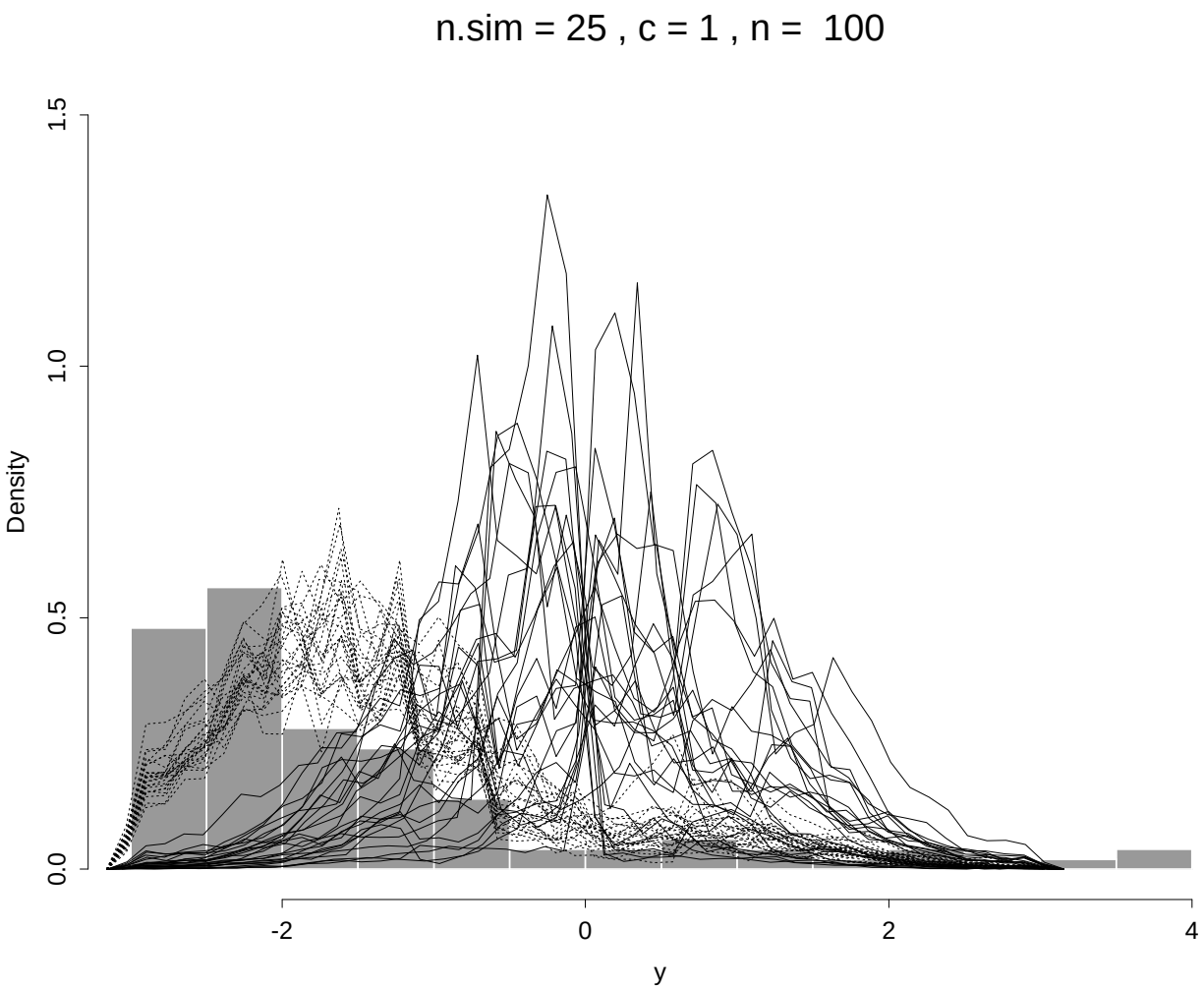
Sampling from a $PT(\Pi, \mathcal{A}_\alpha)$ **prior** for G **centered** at $N(0,1)$ **(solid line)** with $\alpha = 0.1$; with **small** α the **sampled** G bears **little relation** to the **centering distribution**.

PT Prior-To-Posterior Updating



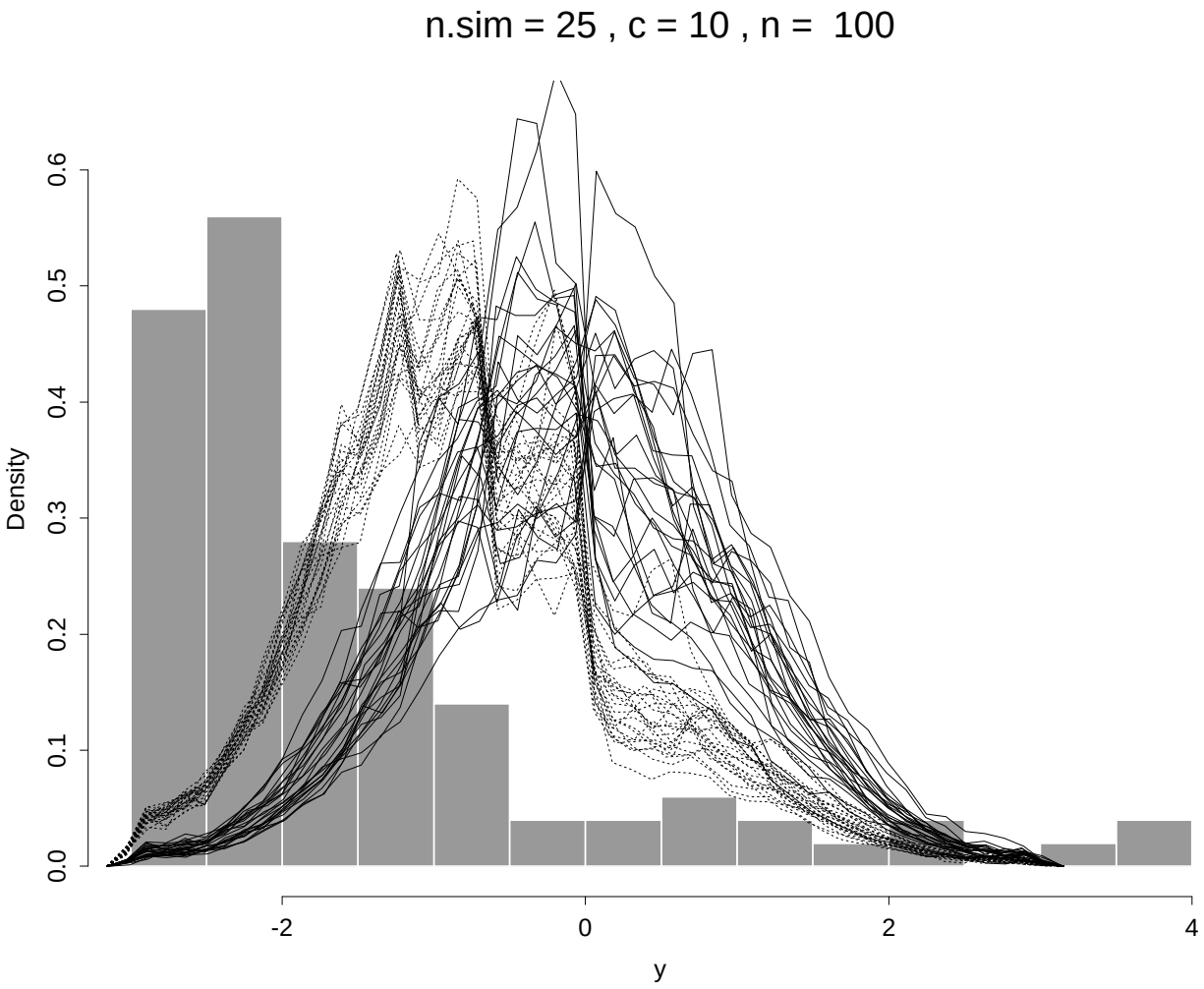
Draws from the **prior (solid lines)**; **data** (**histogram**, $n = 100$); and **draws** from the **posterior (dotted lines)**, with $\alpha = 0.1$; for α **close to 0** the **posterior almost coincides with the data**.

PT Prior-To-Posterior Updating (continued)



Draws from the **prior (solid lines)**; **data** (**histogram**, $n = 100$); and **draws** from the **posterior (dotted lines)**, with $\alpha = 1$; the **posterior** is now a **compromise** between the **prior** and the **data**.

PT Prior-To-Posterior Updating (continued)



Draws from the **prior (solid lines)**; **data** (**histogram**, $n = 100$); and **draws** from the **posterior (dotted lines)**, with $\alpha = 10$; now the **posterior** is **much closer to the prior**.

Pólya Tree Analysis of the NB10 Data

I used the **functions** `PTlm` and `PTdensity` in the `CRAN` **package** `DPpackage` to **perform** a **number** of **analyses** of the **NB10 data**.

`PTdensity` fits the **model**

$$\begin{aligned} (G|\mathcal{B}) &\sim PT(\Pi, \mathcal{A}_\alpha), \quad N(\mu, \sigma^2) \textbf{ determines } \Pi \\ (y_i|G\,\mathcal{B}) &\overset{\text{IID}}{\sim} G \\ (\mu, \sigma^2|\mathcal{B}) &\sim p(\mu, \sigma^2|\mathcal{B}) \\ (\alpha|\mathcal{B}) &\sim p(\alpha|\mathcal{B})\,. \end{aligned} \tag{34}$$

You have a **number of choices**: You can fix μ, σ and/or α at **single numbers** rather than **giving them non-point-mass prior distributions**, You can set m (the **number** of **levels** of the **Pólya tree**; the **symbol** for this in the **code** is `M`), You can **fiddle** with the **MCMC tuning constants** — I'll **cover all of this** in a **real-time demonstration** during the **short course**.

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 0	1	2	3	4	5	6	7	n	Mean	SD
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 $\mathcal{P} = \{$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 $\frac{1.01^2}{0.768} \doteq \mathbf{1.33}$.

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

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

$$\begin{aligned} (y_i|\lambda_i\, M_3\, \mathcal{B}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \gamma_0 + \gamma_1 x_i + \epsilon_i \\ (\epsilon_i|\sigma_\epsilon^2\, M_3\, \mathcal{B}) &\overset{\text{IID}}{\sim} N\left(0, \sigma_\epsilon^2\right) \\ \left(\gamma_0\, \gamma_1\, \sigma_\epsilon^2|M_3\, \mathcal{B}\right) &\sim \text{diffuse.} \end{aligned} \tag{35}$$

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

BNP Example

$\theta = \frac{\text{population } \bar{T}}{\text{population } \bar{C}} = e^{\gamma_1}$; 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 (35) 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 $N\left(0, \sigma_\epsilon^2\right)$ **distribution** but that expresses **substantial prior uncertainty** about the **Gaussian assumption**:

$$
\begin{aligned}
(y_i | \lambda_i \, M_4 \, \mathcal{B}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \gamma_0 + \gamma_1 x_i + \epsilon_i \\
(\epsilon_i | F \, M_4 \, \mathcal{B}) &\overset{\text{IID}}{\sim} F \\
(F | \alpha \, \sigma_\epsilon^2 \, M_4 \, \mathcal{B}) &\sim DP(\alpha, F_0), \;\; F_0 = N(0, \sigma_\epsilon^2) \\
\left(\gamma_0 \, \gamma_1 \, \sigma_\epsilon^2 | M_4 \, \mathcal{B}\right) &\sim \text{diffuse}; \;\; (\alpha | M_4) \sim \text{small positive}\,.
\end{aligned}
\tag{36}
$$

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 (36), $DP(\alpha, F_0)$ is a **Dirichlet-process prior** on F with **prior estimate** $F_0 = N(0, \sigma_\epsilon^2)$ and a **quantity** (α) that behaves something like a **prior sample size**; this is referred to as **Dirichlet-process mixture modeling**, because (36) 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** $\mathcal{M} = \{M_1, \ldots, M_m\}$,

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

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

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

CCV (continued)

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

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

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

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

This is **related** to the **machine-learning** practice (e.g., **Hastie, Tibshirani, Friedman** [HTF] 2009) of **Train/Validation/Test** partitioning, with one **improvement** (**decision theory** provides an **optimal way** to choose the **data subset sizes**); I **don't agree** with HTF that this can **only be done with large data sets**: it's even **more important** to do it with **small** and **medium-size data sets** (You just need to work with **multiple (M, V, C) partitions** and **average**).

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):

$\boxed{Q_1}$: Is M_1 **better** than M_2? $\boxed{Q_2}$: 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**,

Bayes Factors

$$\frac{p(M_2|D\,\mathcal{B})}{p(M_1|D\,\mathcal{B})} = \left[\frac{p(M_2|\mathcal{B})}{p(M_1|\mathcal{B})}\right] \cdot \left[\frac{p(D|M_2\,\mathcal{B})}{p(D|M_1\,\mathcal{B})}\right]; \tag{37}$$

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 (37) says

$$\left(\begin{array}{c}\textbf{posterior}\\ \textbf{odds}\\ \text{for } M_2\\ \text{over } M_1\end{array}\right) = \left(\begin{array}{c}\textbf{prior odds}\\ \text{for } M_2\\ \text{over } M_1\end{array}\right) \cdot \left(\begin{array}{c}\textbf{Bayes factor}\\ \text{for } M_2\\ \text{over } M_1\end{array}\right). \tag{38}$$

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

Odds o are related to **probabilities** p via $o = \frac{p}{1-p}$ and $p = \frac{o}{1+o}$; these are **monotone increasing transformations**, so the **decision rules** {choose M_2 over M_1 if the **posterior odds** for M_2 are greater} and {choose M_2 over M_1 if $p(M_2|D\,\mathcal{B}) > p(M_1|D\,\mathcal{B})$} 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 $\mathcal{M} = \{M_1, \ldots, M_m\}$ for finite m, and You pretend that one of the models in $\mathcal{M}$ must be the **true data-generating mechanism** M_{DG}, and You pretend that the **utility function**

$$U(M, M_{DG}) = \left\{ \begin{array}{cc} 1 & \text{if } M = M_{DG} \\ 0 & \text{otherwise} \end{array} \right\} \tag{39}$$

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,

$$BF(M_2 \text{ over } M_1|D\,\mathcal{B}) = \frac{p(D|M_2\,\mathcal{B})}{p(D|M_1\,\mathcal{B})}, \tag{40}$$

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

$$M_j : \left\{ \begin{array}{c} (\gamma_j|M_j\,\mathcal{B}) \sim p(\gamma_j|M_j\,\mathcal{B}) \\ (D|\gamma_j\,M_j\,\mathcal{B}) \sim p(D|\gamma_j\,M_j\,\mathcal{B}) \end{array} \right\} . \tag{41}$$

Here the quantity $p(D|M_j\,\mathcal{B})$ that **defines the Bayes factor** is

Integrated Likelihoods

$$p(D|M_j\,\mathcal{B}) = \int p(D|\gamma_j\,M_j\,\mathcal{B})\,p(\gamma_j|M_j\,\mathcal{B})\,d\gamma_j\,; \tag{42}$$

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 (42) can be **difficult to compute**, and may not even be easy to **approximate**.

The second thing to **notice** is that (42) can be **rewritten** as

$$p(D|M_j\,\mathcal{B}) = E_{(\gamma_j|M_j\,\mathcal{B})}\;p(D|\gamma_j\,M_j\,\mathcal{B})\,. \tag{43}$$

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**

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

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

$\{$**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 $D = (y_1\ \ldots\ y_n)$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$;

Instability of Bayes Factors (continued)

$M_1 =$ **Geometric**(θ_1) likelihood with a **Beta**(α_1, β_1) prior on θ_1;

$M_2 =$ **Poisson**(θ_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

$$\frac{\Gamma(\alpha_1+\beta_1)\,\Gamma(n+\alpha_1)\,\Gamma(n\bar{y}+\beta_1)\,\Gamma(\alpha_2)\,(n+\beta_2)^{n\bar{y}+\alpha_2}\left(\prod_{i=1}^n y_i!\right)}{\Gamma(\alpha_1)\,\Gamma(\beta_1)\,\Gamma(n+n\bar{y}+\alpha_1+\beta_1)\,\Gamma(n\bar{y}+\alpha_2)\,\beta_2^{\alpha_2}}. \quad (44)$$

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

$$\frac{\Gamma(n+1)\,\Gamma(n\bar{y}+1)\,\Gamma(\epsilon)\,(n+\epsilon)^{n\bar{y}+\epsilon}\left(\prod_{i=1}^n y_i!\right)}{\Gamma(n+n\bar{y}+2)\,\Gamma(n\bar{y}+\epsilon)\,\epsilon^{\epsilon}}. \quad (45)$$

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

$$p(D|M_j\,\mathcal{B}) = \int p(D|\gamma_j\,M_j\,\mathcal{B})\,p(\gamma_j|M_j\,\mathcal{B})\,d\gamma_j\,; \tag{46}$$

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

$$\begin{aligned} \log p(D|M_j\, \mathcal{B}) \quad = \quad & \log p(D|\hat{\gamma}_j\, M_j\, \mathcal{B}) + \log p(\hat{\gamma}_j|M_j\, \mathcal{B}) \\ & + \frac{k_j}{2} \log 2\pi - \frac{1}{2} \log |\hat{I}_j| + O\left(\frac{1}{n}\right) \; ; \end{aligned} \tag{47}$$

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.

BIC

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)**:

$$\log p(y | M_j \mathcal{B}) = \log p(y | \hat{\gamma}_j M_j \mathcal{B}) - \frac{k_j}{2} \log n + O(1). \tag{48}$$

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

$$BIC(M_j | D \mathcal{B}) = -2 \log p(D | \hat{\gamma}_j M_j \mathcal{B}) + k_j \log n \tag{49}$$

(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:

$$AIC(M_j|D\,\mathcal{B}) = -2\log p(D|\hat{\gamma}_j\,M_j\,\mathcal{B}) + 2\,k_j\,. \tag{50}$$

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

$$(\gamma_j|M_j\,\mathcal{B}) \sim N_{k_j}(\hat{\gamma}_j, n\hat{I}_j^{-1}). \tag{51}$$

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

- **Log scores** are **based on** the

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

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

Log Scores

$$M_j\text{:} \left\{ \begin{array}{c} (\gamma_j | M_j \, \mathcal{B}) \sim p(\gamma_j | M_j \, \mathcal{B}) \\ (y | \gamma_j \, M_j \, \mathcal{B}) \sim p(y | \gamma_j \, M_j \, \mathcal{B}) \end{array} \right\} . \tag{52}$$

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

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

$$LS_{CV}(M_j | y \, \mathcal{B}) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | y_{-i} \, M_j \, \mathcal{B}) \, , \tag{53}$$

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

$$LS_{FS}(M_j|y\,\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y\,M_j\,\mathcal{B})\,, \tag{54}$$

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**:

$$\begin{aligned} KL[M_{DG}||p(\cdot|y\,M_j\,\mathcal{B})] &= E_{M_{DG}} \log M_{DG} - E_{M_{DG}} \log p(\cdot|y\,M_j\,\mathcal{B}) \\ &\doteq E_{M_{DG}} \log M_{DG} - LS_{FS}(M_j|y\,\mathcal{B})\,; \end{aligned} \tag{55}$$

the **first term** on the **right side** of (55) is **constant** in $p(\cdot|y\,M_j\,\mathcal{B})$, so **minimizing** $KL[M_{DG}||p(\cdot|y\,M_j\,\mathcal{B})]$ is **approximately the same** as **maximizing** LS_{FS}.

Bayes Factors/BIC Versus Log Scores

What follows is a **sketch** of **recent results** (Draper, 2011) based on **simulation experiments** with **realistic sample sizes**; in my view **standard asymptotic calculations — choosing between** the **models** in $\mathcal{M} = \{M_1, M_2\}$ as $n \to \infty$ with $\mathcal{M}$ **remaining fixed —** are **essentially irrelevant** in **calibration studies**, for **two reasons**:

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

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

• **Versions** of **Bayes factors** that **behave sensibly** with **diffuse priors** on the **model parameters** (e.g., **intrinsic Bayes factors**: Berger and Pericchi, 1996, and **more recent** cousins) tend to have **model discrimination performance similar** to that of **BIC** in **calibration** (**repeated-sampling** with **known** M_{DG}) **environments**; I'll show **results** for **BIC** here.

Example: Consider **assessing** the **performance** of a **drug**, for **lowering**

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

The **real-world purpose** of this **experiment** is to **decide** whether to **take the drug forward** to **phase III**; under the **weight** of **20th-century**

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**:

$$M_1\text{: } \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } \theta \leq \Delta \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{56}$$

$$M_2\text{: } \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } \theta > \Delta \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{57}$$

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 $LS_{FS}(M_2|y\,\mathcal{B}) > LS_{FS}(M_1|y\,\mathcal{B})$.

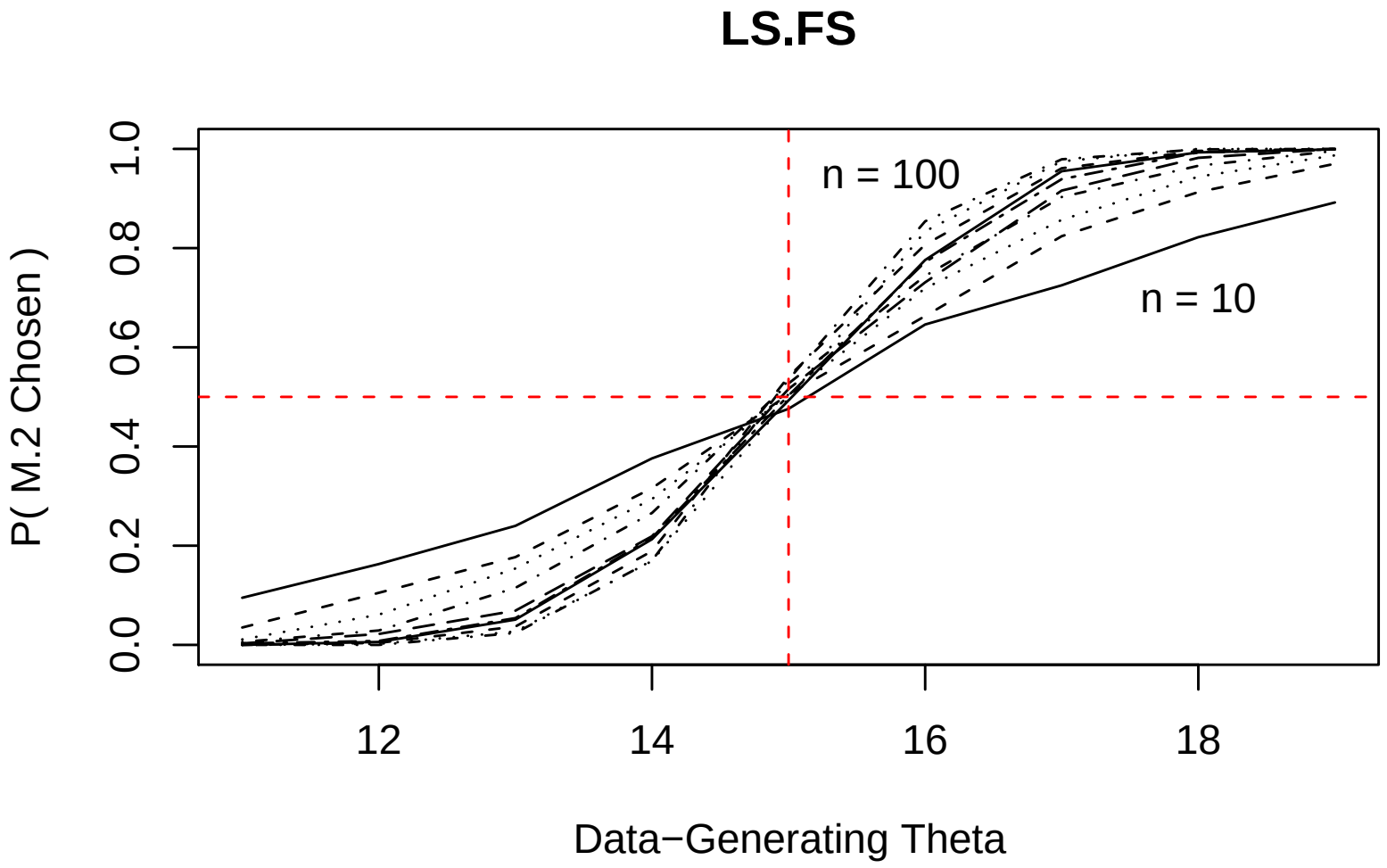
- **Posterior probability**: let $M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \Re, (y_i|\theta\,\mathcal{B}) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)\}$ and **choose** M_2 if $p(\theta > \Delta|y\,M^*\,\mathcal{B}) > 0.5$.

- **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**: $\Delta = 15$; $\sigma = 10$; $n = 10, 20, \ldots, 100$; **data-generating** $\theta_{DG} = 11, 12, \ldots, 19$; $\alpha = 0.05$; 1,000 **simulation replications**; **Monte-Carlo approximations** of the **predictive ordinates** in LS_{FS} based on **10,000 posterior draws**.

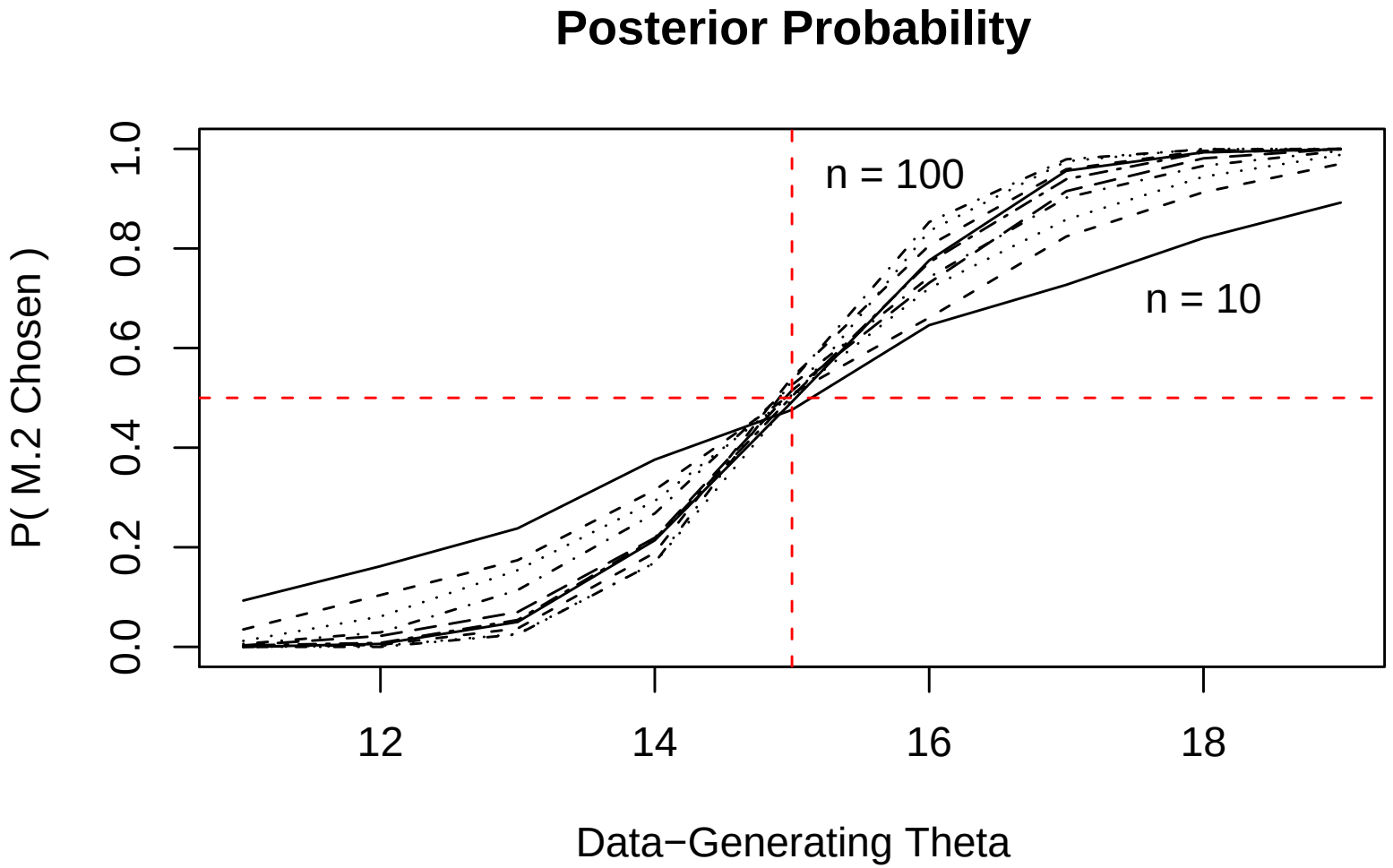
The **figures** below give **Monte-Carlo estimates** of the **probability that M_2 is chosen**.

LS_{FS} Results: Quantifying Improvement



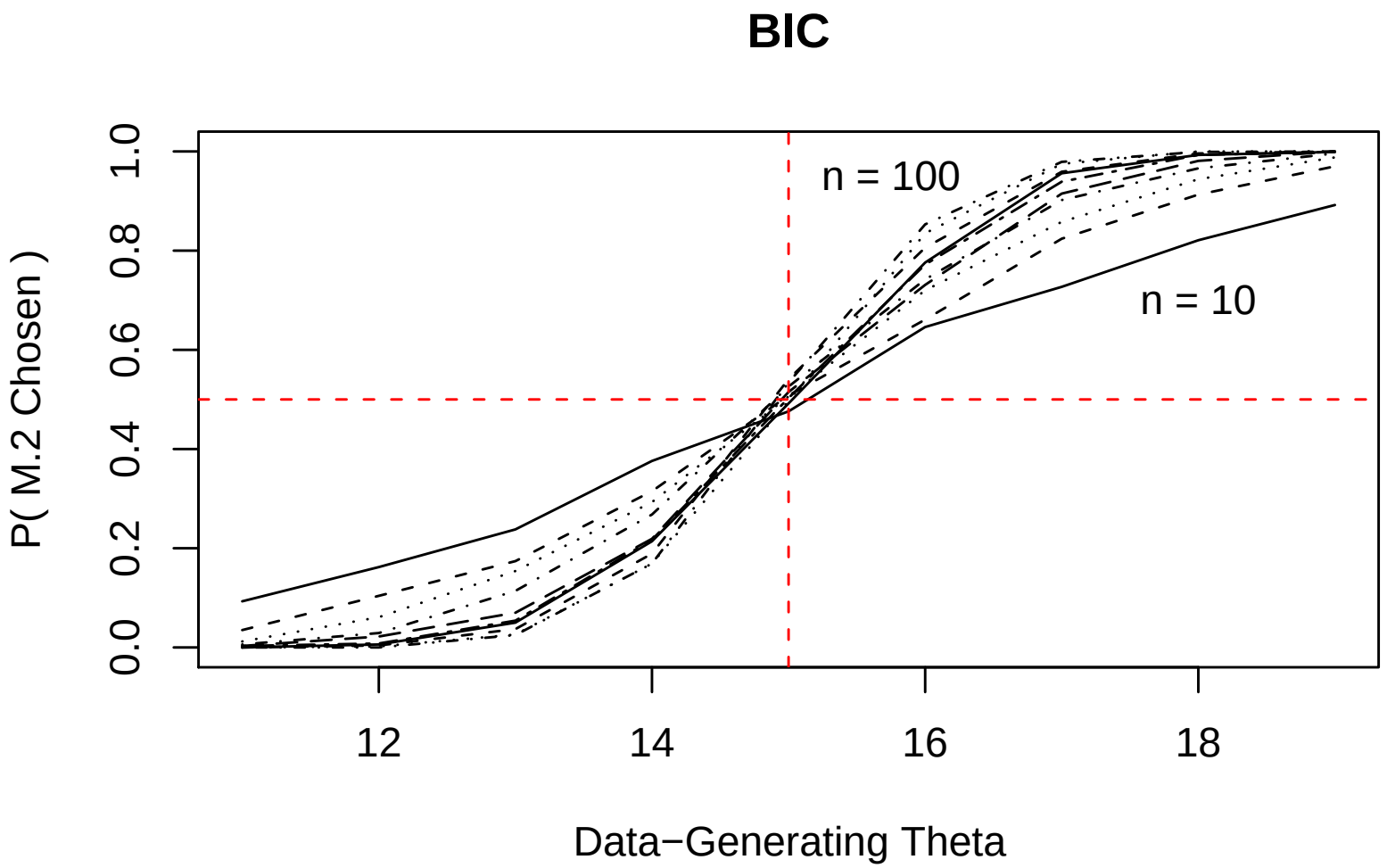
This **exhibits all** the **monotonicities** that it **should**, and **correctly yields 0.5** for all n with $\theta_{DG} = 15$.

Posterior Probability Results: Quantifying Improvement



Even though the LS_{FS} and **posterior-probability methods** are **quite different**, their **information-processing** in **discriminating** between M_1 and M_2 is **identical** to within $\pm$**0.003** (**well within simulation noise** with **1,000 replications**).

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**

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \qquad (58)$$

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

Again in this **setting** there's **nothing special** about $\theta = 0$, and **as before** You **know scientifically** that 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

$$M_3: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{59}$$

$$M_4: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{60}$$

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 $M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \Re, (y_i|\theta\,\mathcal{B}) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)\}$, but this time **favor** M_4 over M_3 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

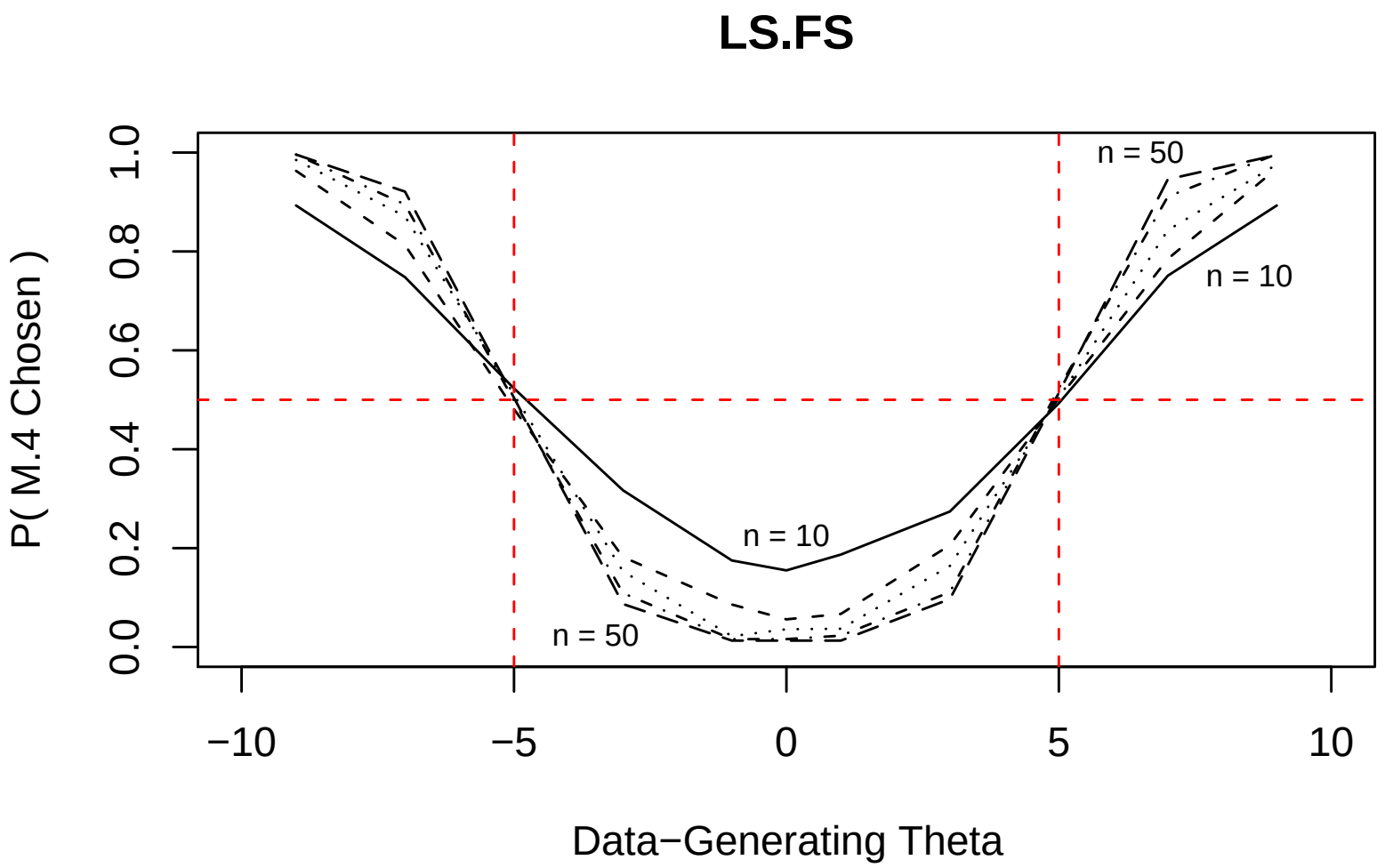
{**claiming** the two drugs were **bio-equivalent** when they **were**,
concluding that they were **bio-equivalent** when they **were not**,
deciding that they were **not bio-equivalent** when they **were**,
judging that they were **not bio-equivalent** when they were **not**},

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**: $\lambda = 5$; $\sigma = 10$; $n = 10, 20, \ldots, 100$; **data-generating** $\theta_{DG} = \{-9, -7, -5, -3, -1, 0, 1, 3, 5, 7, 9\}$; $\alpha = 0.05$; **1,000 simulation replications**, $M =$ **10,000 Monte-Carlo draws** for LS_{FS}.

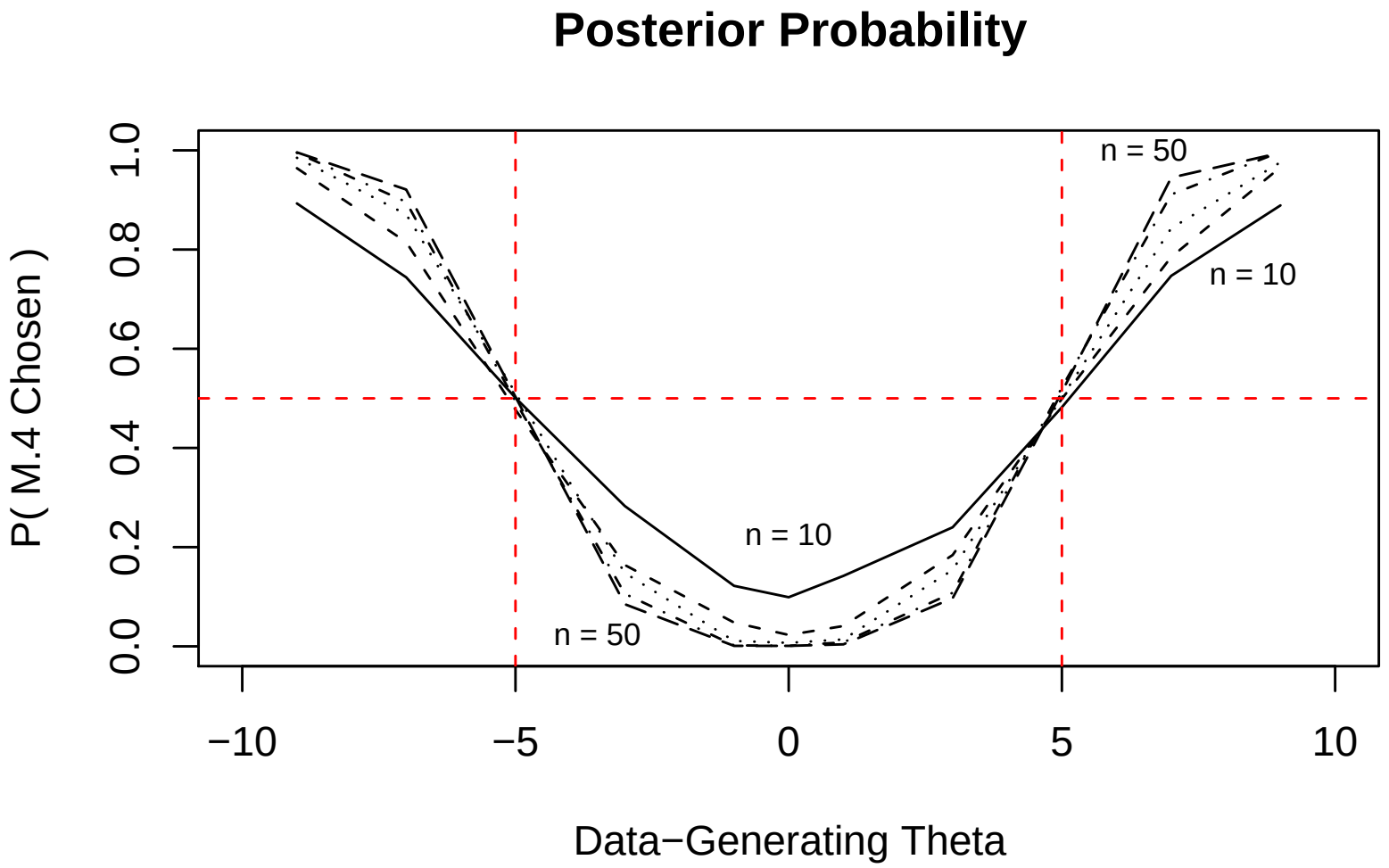
NB It has **previously been established** that when **making** the **(unrealistic) sharp-null comparison** $\theta = 0$ versus $\theta \neq 0$ in the **context** of $(y_i|\theta\, \mathcal{B}) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)$, as $n \to \infty$ LS_{FS} **selects** the $\theta \neq 0$ **model** with **probability** $\to$ **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} Results: Bio-Equivalence



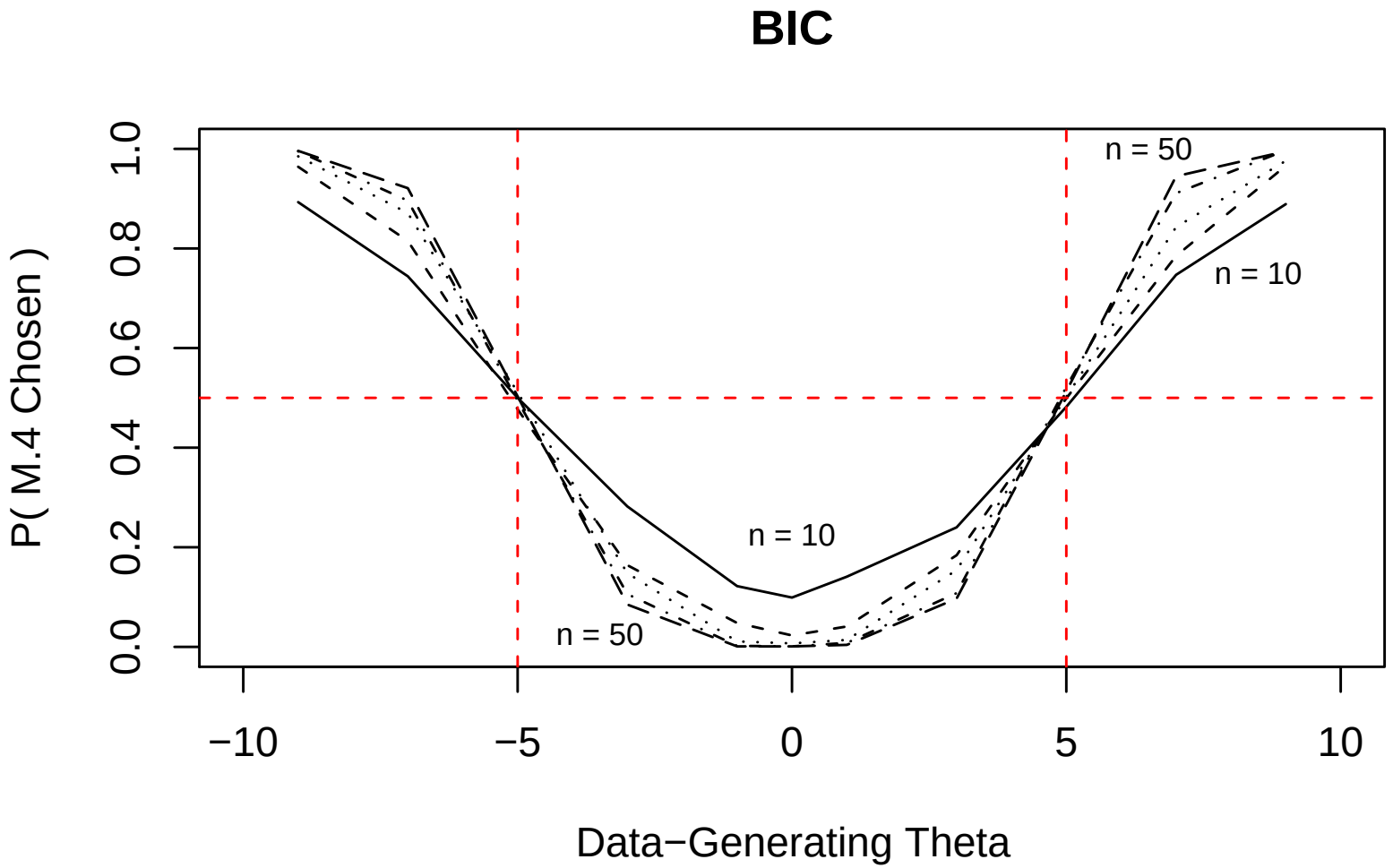
In this **more realistic setting**, comparing $|\theta| \leq \lambda$ versus $|\theta| > \lambda$ with $\lambda > 0$, LS_{FS} has the **correct large-sample behavior**, **both** when $|\theta_{DG}| \leq \lambda$ and when $|\theta_{DG}| > \lambda$.

Posterior Probability Results: Bio-Equivalence



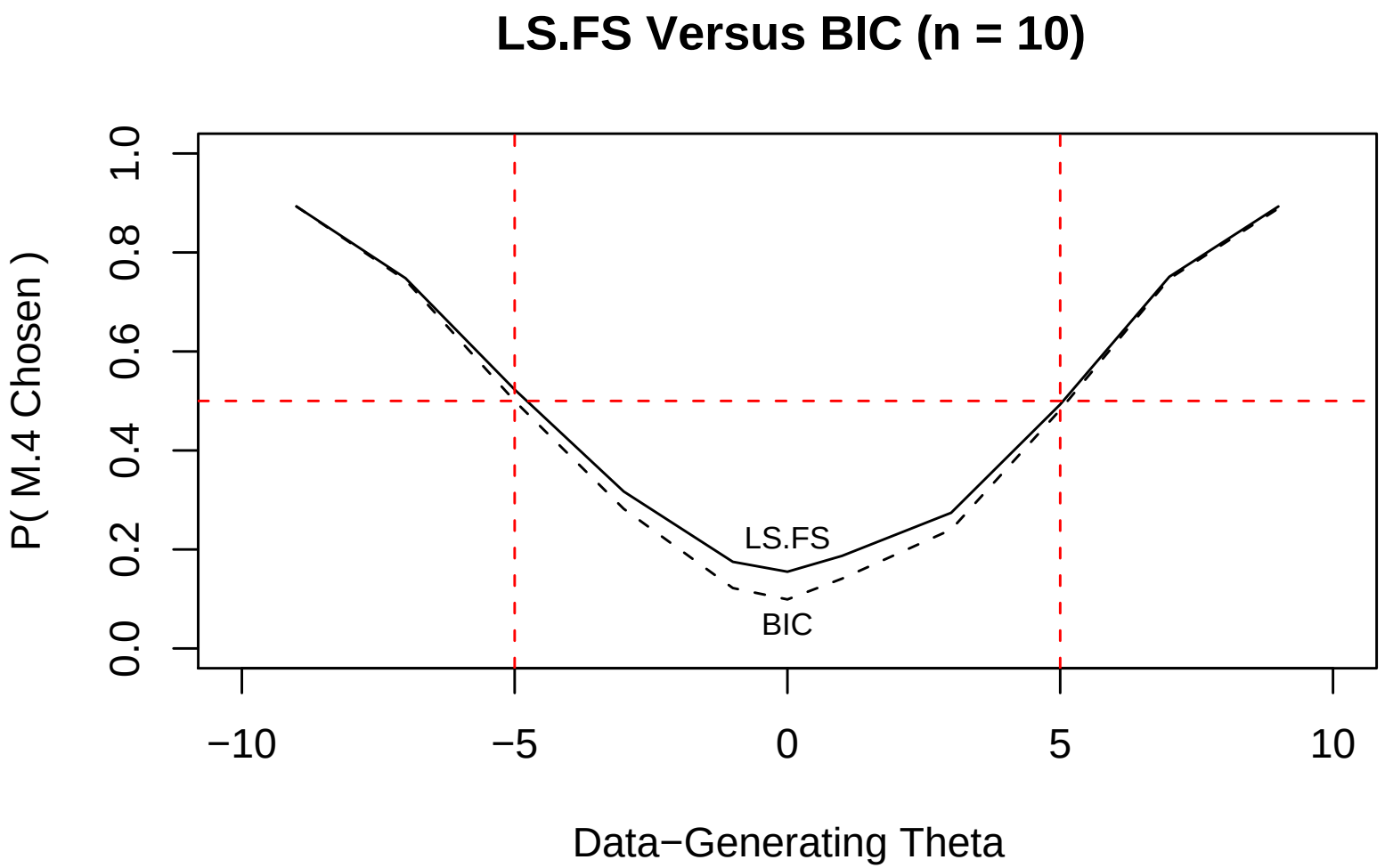
The **qualitative behavior** of the LS_{FS} and **posterior-probability methods** is **identical**, although there are some **numerical differences** (**highlighted** later).

BIC Results: Bio-Equivalence



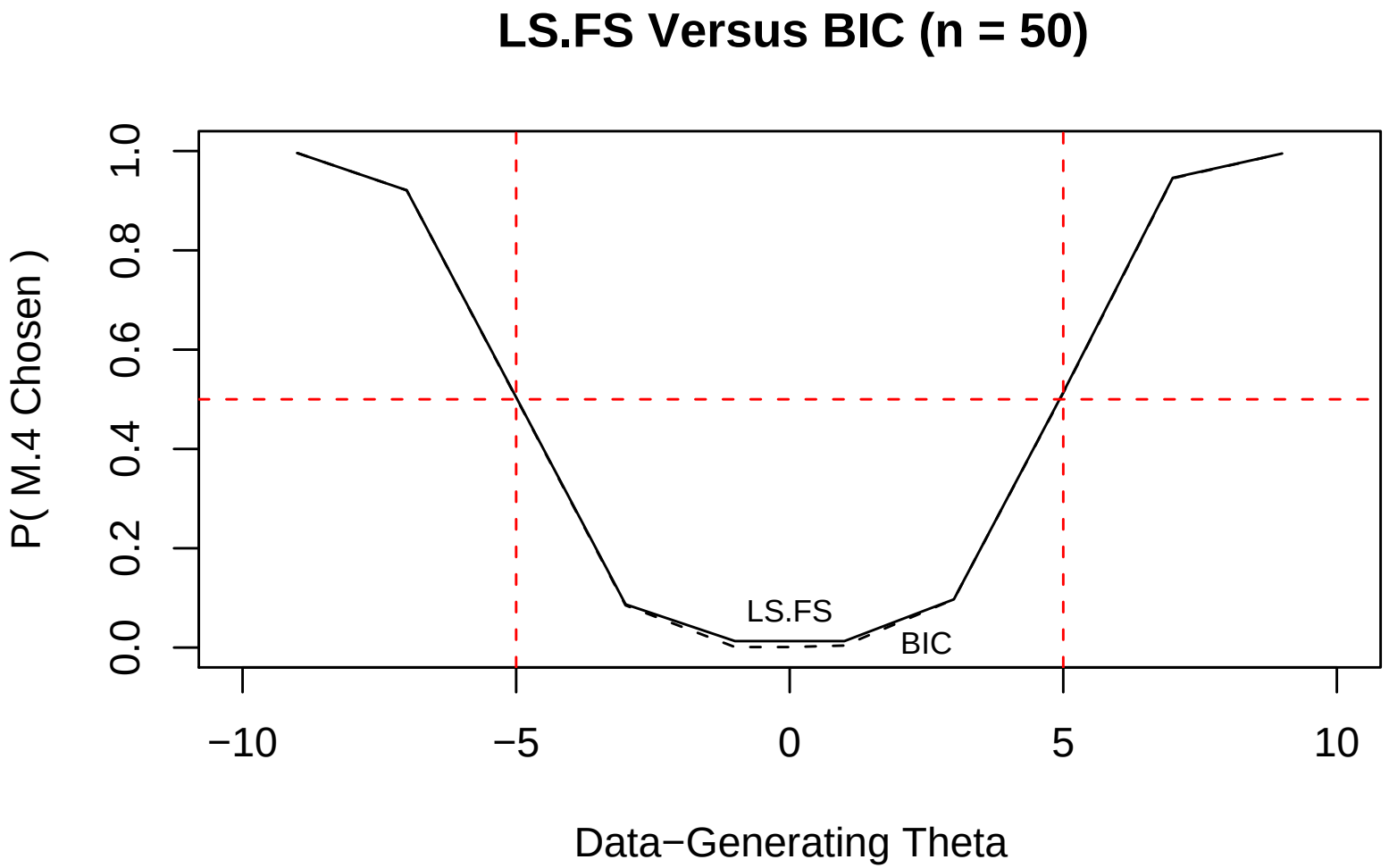
In the **quantifying-improvement** case, the **BIC** and **posterior-probability methods** were **algebraically identical**; here they **nearly coincide** (**differences** of $\pm$ **0.001** with **1,000 simulation repetitions**).

LS_{FS} Versus BIC Results: Bio-Equivalence



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



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

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 {**these two genes** are on **different chromosomes** (the **strength** θ of their **genetic linkage** is $\theta_0 = 0$)} and {**these two genes** are on the **same chromosome** ($\theta > 0$)}.

Sharp-null testing without **structural singletons** is **always unwise** because

(a) **You already know** from **scientific context**, when the **outcome variable** is **continuous**, that H_0 is **false**, and **(relatedly)**

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

$$M_5: \left\{ \begin{array}{ccc} (\sigma^2|\mathcal{B}) & \sim & \text{diffuse on } (0, \text{large}) \\ (y_i|\sigma^2\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(0, \sigma^2) \end{array} \right\} \quad \text{and} \qquad (61)$$

$$M_6: \left\{ \begin{array}{ccc} (\theta\,\sigma^2|\mathcal{B}) & \sim & \text{diffuse on } (-\text{large}, \text{large}) \times (0, \text{large}) \\ (y_i|\theta\,\sigma^2\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \qquad (62)$$

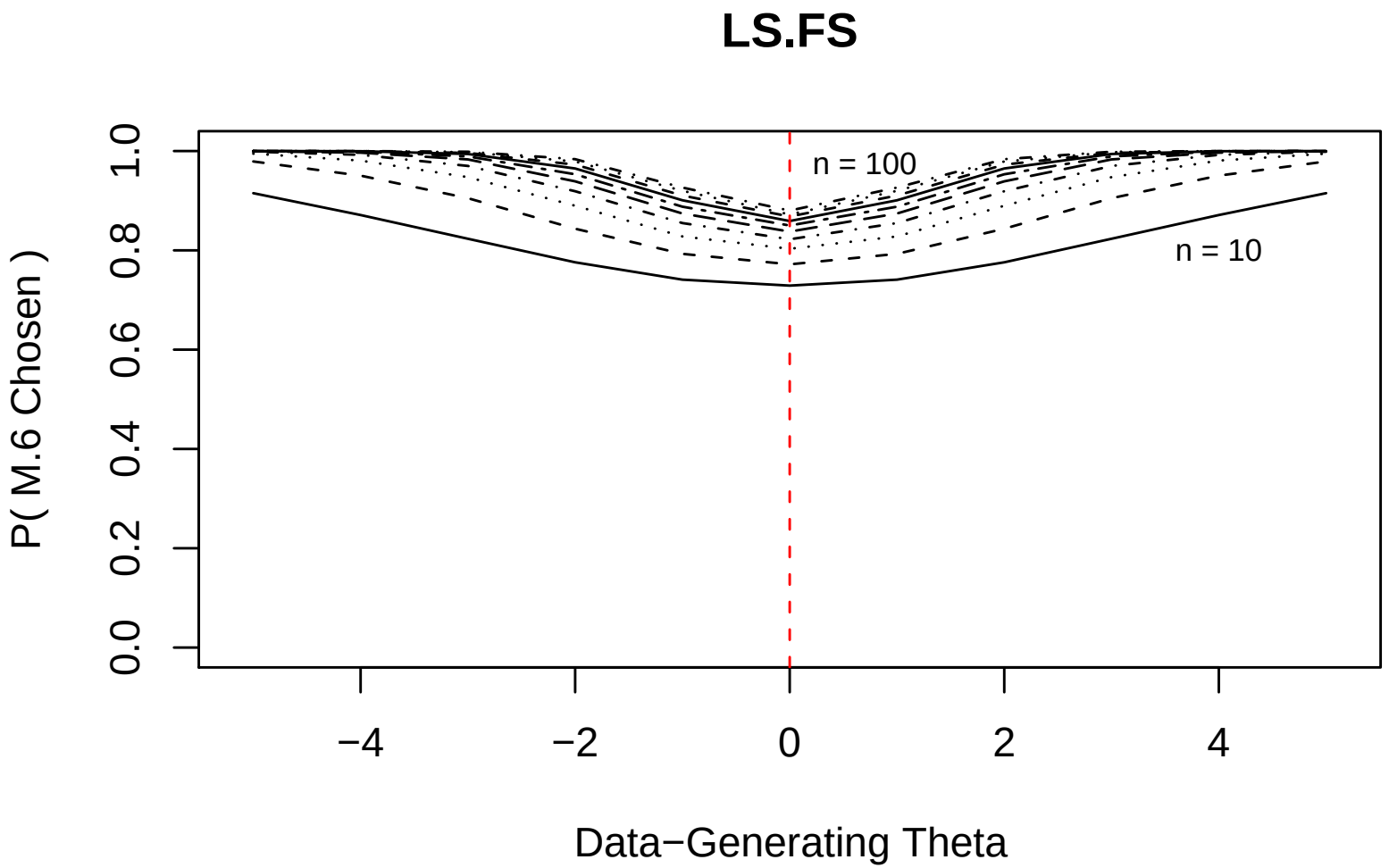
In **this case** a **natural Bayesian competitor** to **BIC** and LS_{FS} would be to **construct** the **central** $100(1-\alpha)\%$ **posterior interval** for θ under M_6 and **choose** M_6 if **this interval doesn't contain 0**.

Testing Sharp-Null Hypotheses (continued)

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

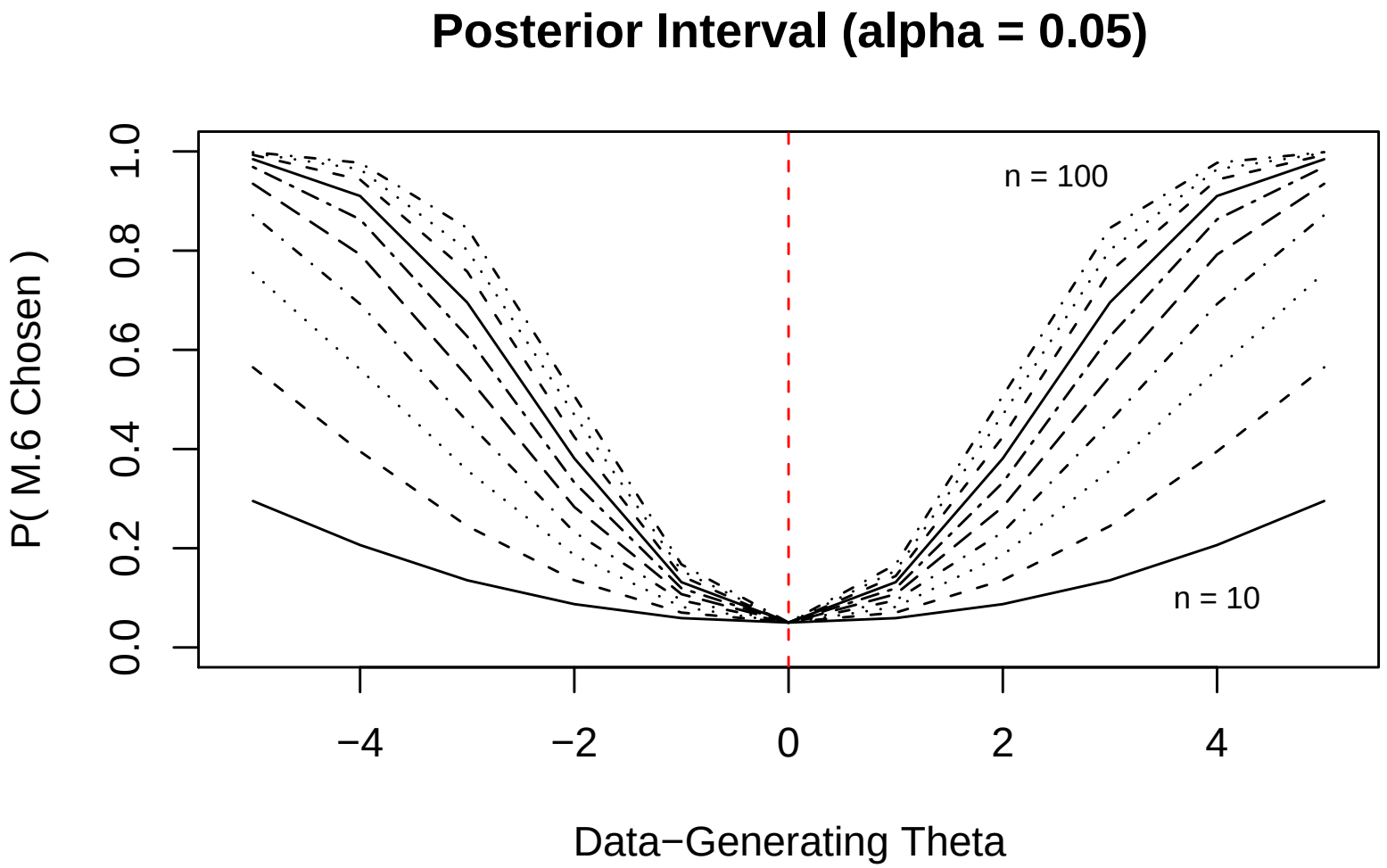
As before, let's call **choosing** M_6: $\theta \neq 0$ when $\theta_{DG} = 0$ a **false-positive** error and **choosing** M_5: $\theta = 0$ when $\theta_{DG} \neq 0$ a **false-negative** mistake.

LS_{FS} Results: Sharp-Null Testing



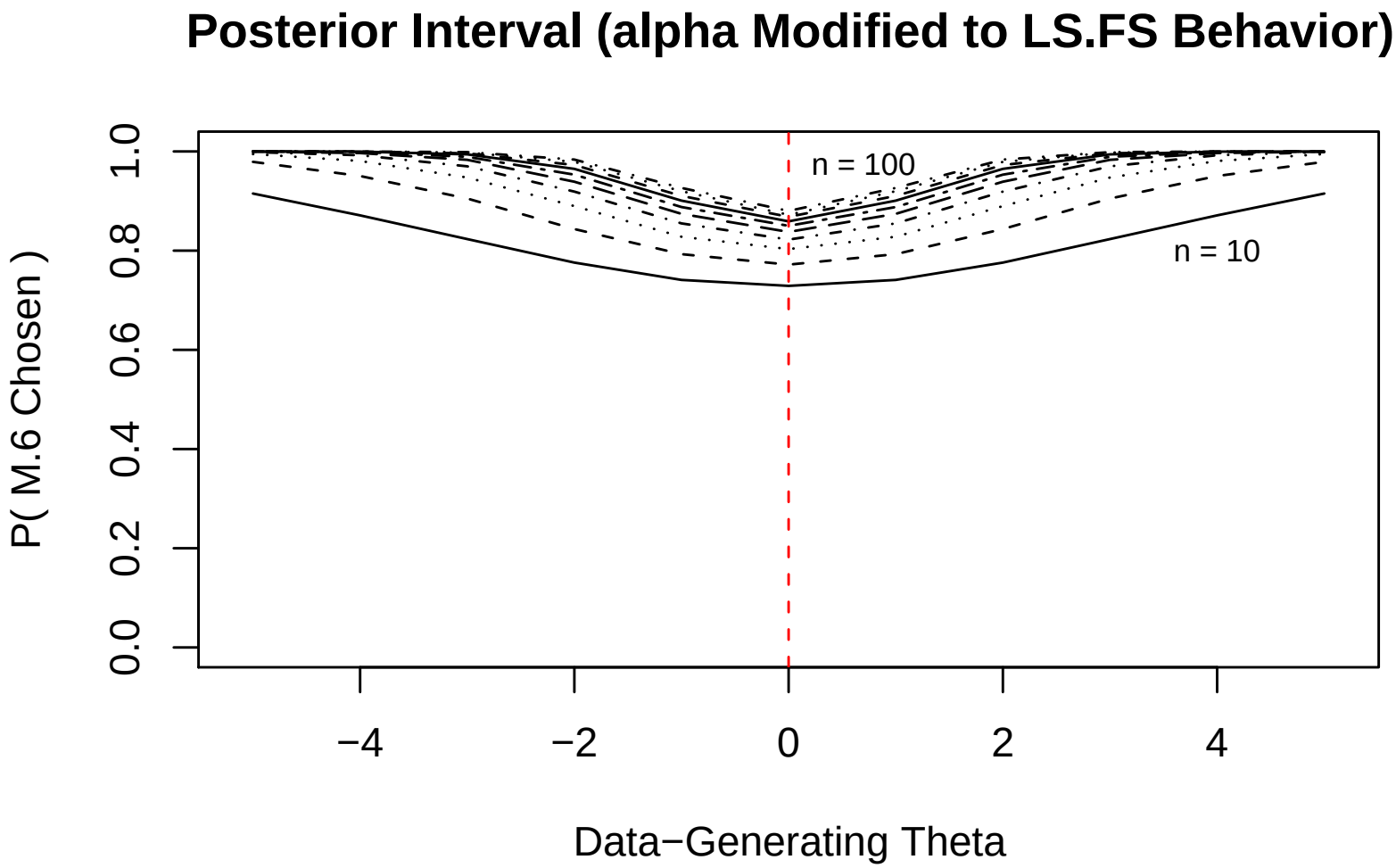
In the **limit** as $\lambda \downarrow 0$, the LS_{FS} **approach** makes **hardly any false-negative errors** but **quite a lot of false-positive mistakes**.

Interval ($\alpha = 0.05$) Results: Sharp-Null Testing



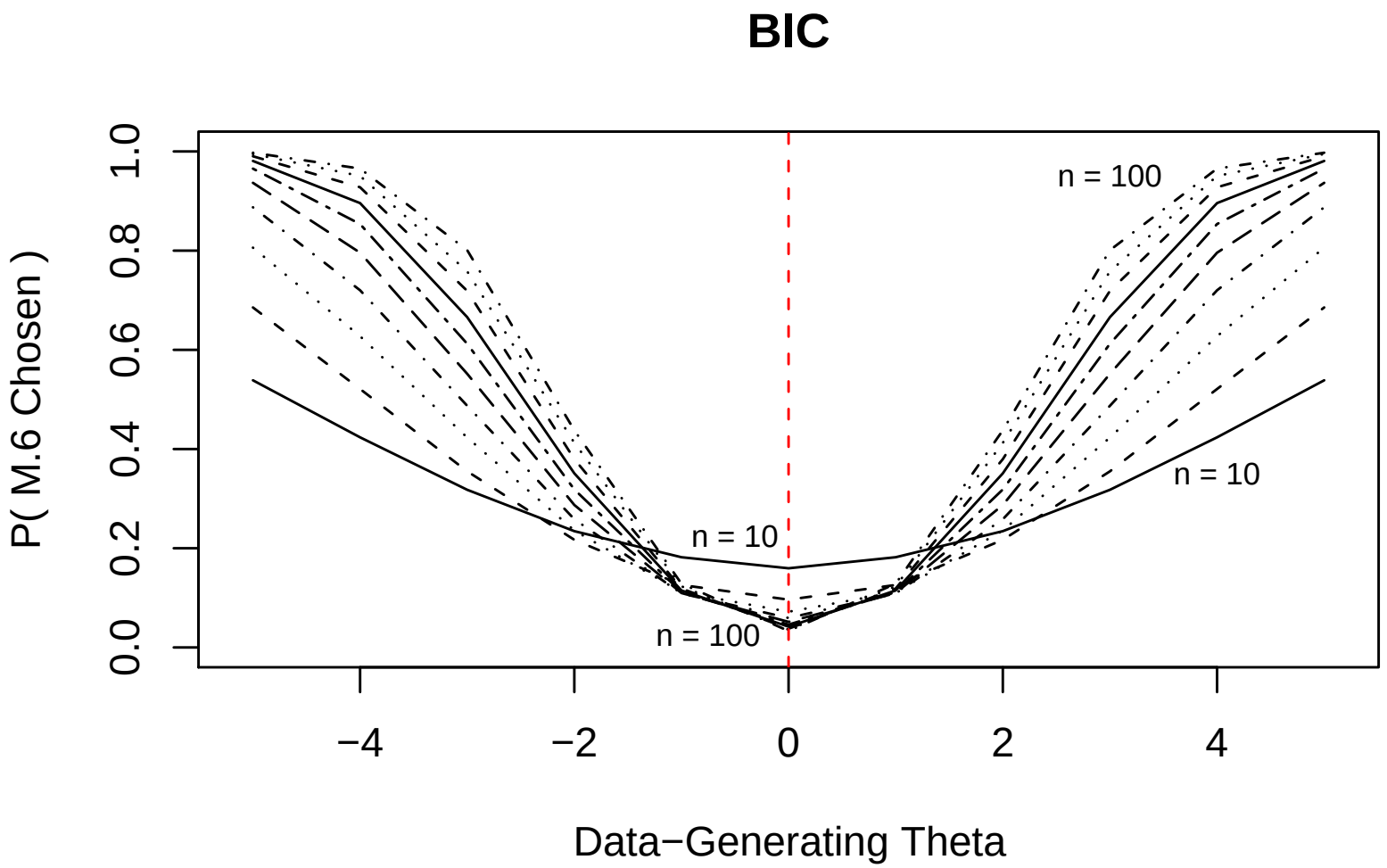
The **behavior** of the **posterior interval approach** is of course **quite different**: it makes **many false-negative errors** because its **rate of false-positive mistakes** is **fixed at 0.05**.

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



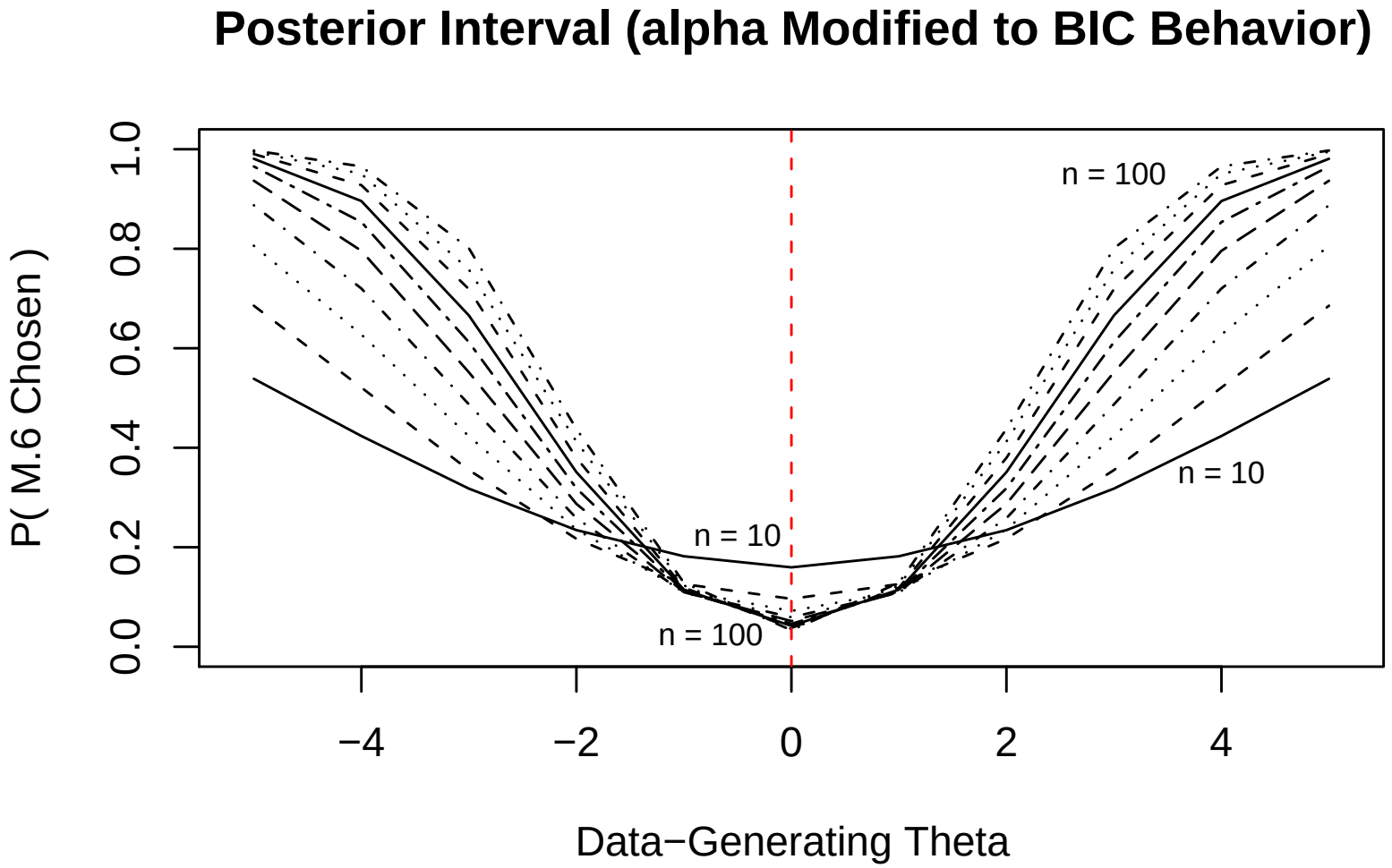
When the **interval method** is **modified** so that α **matches** the LS_{FS} **behavior** at $\theta_{DG} = 0$ (letting α **vary** with n), the **two approaches** have **identical model-discrimination ability**.

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



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 $D = y = (y_1 \ldots y_n)$ 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 $s = \sum_{i=1} y_i$ and $\hat{\theta}_1 = \frac{\alpha_1 + n}{\alpha_1 + \beta_1 + s + n}$,

$$\begin{aligned} LS_{FS}(M_7|y\,\mathcal{B}) \quad = \quad & \log\Gamma(\alpha_1 + n + \beta_1 + s) + \log\Gamma(\alpha_1 + n + 1) \\ & - \log\Gamma(\alpha_1 + n) - \log\Gamma(\beta_1 + s) \\ & + \frac{1}{n}\sum_{i=1}^{n}[\log\Gamma(\beta_1 + s + y_i) \\ & - \log\Gamma(\alpha_1 + n + \beta_1 + s + y_i + 1)]\,, \end{aligned} \tag{63}$$

$$BIC(M_7|y\,\mathcal{B}) = -2[n\log\hat{\theta}_1 + s\log(1 - \hat{\theta}_1)] + \log n\,, \tag{64}$$

Geometric Versus Poisson (continued)

$$
\begin{aligned}
LS_{FS}(M_8|y\,\mathcal{B}) \;=\; & (\alpha_2+s)\log(\beta_2+n) - \log\Gamma(\alpha_2+s) \\
& -(\alpha_2+s)\log(\beta_2+n+1) \qquad (65) \\
& +\frac{1}{n}\sum_{i=1}^{n}[\log\Gamma(\alpha_2+s+y_i) - y_i\log(\beta_2+n+1) \\
& -\log\Gamma(y_i+1)]\,, \text{ and}
\end{aligned}
$$

$$
BIC(M_8|y\,\mathcal{B}) = -2[s\log\hat{\theta}_2 - n\,\hat{\theta}_2 - \sum_{i=1}^{n}\log(y_i!)] + \log n\,, \qquad (66)
$$

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

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

n	LS.FS	BIC
10	0.8967	0.8661
20	0.9185	0.8906
40	0.9515	0.9363
80	0.9846	0.9813

M.DG = Geometric

n	LS.FS	BIC
10	0.4857	0.4341
20	0.3152	0.2671
40	0.1537	0.1314
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}).

Properties of 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

$$\begin{aligned} p(y_i|y\,M_j\,\mathcal{B}) &= \int p(y_i|\gamma_j\,M_j\,\mathcal{B})\,p(\gamma_j|y\,M_j\,\mathcal{B})\,d\gamma_j \qquad (67)\\ &= E_{(\gamma_j|y\,M_j\,\mathcal{B})}p(y_i|\gamma_j\,M_j\,\mathcal{B})\,; \end{aligned}$$

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

$$n\, LS_{FS}(M_j|y, \mathcal{B}) > n\, LS_{FS}(M_{j'}|y, \mathcal{B}). \tag{68}$$

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,

Summary

$$nLS_{FS}(M_j|y,\mathcal{B}) = \log \prod_{i=1}^{n} \left[\int p(y_i|\eta_j, M_j, \mathcal{B})\, p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j \right], \quad (69)$$

and this is **not the same as**

$$\log \int \left[\prod_{i=1}^{n} p(y_i|\eta_j, M_j, \mathcal{B}) \right] p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j = \bar{L}_j^{PBF} \quad (70)$$

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 $k =$ **10–100**.

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

$$\begin{aligned} (y_i|\lambda_i, \mathcal{B}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log \lambda_i &= \beta_0 + \beta_1 x_i + e_i \\ (e_i|\sigma_e^2, \mathcal{B}) &\overset{\text{IID}}{\sim} N(0, \sigma_e^2)\,, \quad (\beta_0, \beta_1, \sigma_e^2) \sim \text{diffuse}\,, \end{aligned} \tag{71}$$

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

it's **also true** that

New Theorem (Draper, 2011): for any $\lambda > 0$,
$P_{|\theta_{DG}| \le \lambda}(LS_{FS} \text{ chooses } |\theta| \le \lambda) \to 1 \text{ as } n \to \infty,$

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

- As a **profession**, we need to **strengthen** the progression

Principles → Axioms → 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 $\boxed{Q_2}$: **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**

$\boxed{Q_{2'}}$: **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 $P = \mathbf{0.4}$ from Gelman et al. (indicating that **Your model is fine**) when a **well-calibrated** version of **their idea** would have $P = \mathbf{0.04}$ (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**).

Summary

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

• **Cox's Theorem** lays out a **progression** from

Principles → **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 the **data-generating model** M_{DG} is **one** or the **other** of $\mathcal{M} = \{M_1, M_2\}$ (apart from **parameter estimation**), and call $\{$choosing M_2 when $M_{DG} = M_1\}$ a **false positive** and $\{$choosing M_1 when $M_{DG} = M_2\}$ a **false negative**; then

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

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

and **maximize expected utility**.

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 (2010) for a **method** for **answering the question** $Q_{2'}$**: Could the data have arisen from model** M_j**?** in a **well-calibrated way**.

• **CCV** provides an **approach** to finding a **good ensemble** $\mathcal{M}$ **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**.)