Bayesian Model Specification: Towards 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

Hierarchical Models and Markov Chain Monte Carlo: Conference in Honour of Adrian FM Smith

Hersonissos, Crete, Greece: *3 June 2011*

(There's a **longer version** of **this talk** at my **web site**.)

Outline

(1) **Foundations** of **probability** are **secure**:
(RT Cox, 1946) **Principles** $\rightarrow$ **Axioms** $\rightarrow$ **Theorem**:

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

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

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

(3) A **Modeling-As-Decision Principle** solves **2 (a)**, but this is **hard work**; **approximate solutions** are **helpful**; this is where **log scores** (based on a **Prediction Principle**) and **Bayes factors** come in.

(4) A **Calibration Principle** fixes **2 (b)** via **decision theory**.

Ingredients

- **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**.

- 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**.

- $\mathcal{B}$, a **set** of **true/false propositions** summarizing Your **background assumptions and judgments** about **how the world works** vis à vis θ and D.

- **Statistics** concerns itself **principally** with **five things** (omitted: **description**, **data integrity**, ...):

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

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

Axiomatization of Probability

(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**.

These **tasks require** a **probability framework — three main attempts** so far to **axiomatize probability**:

- **Kolmogorov** (1933), based on **sets**; **many types of uncertainty cannot (uniquely, comfortably) be fit into this framework**.

- **de Finetti** (1937), based on **betting odds** on the **truth** of **true/false propositions**: **more general** than **Kolmogorov**, but **betting odds** are **not fundamental to science**.

Cox's Theorem

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

Cox put forward **three principles** that $pl(A|B)$ — the **plausibility** that the **proposition** A is **true**, given that the **proposition** B is **known to be true** — should **follow** so that $pl(A|B)$ behaves **sensibly**; he then **derived three axioms** from the **principles**, and **proved** a

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)**:

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

Corollaries from Cox's Theorem

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.

The **advance** made by this **theorem** is that **all of the usual probability rules** emerge from a **basis of propositions, not sets**, so that **direct quantification** of **uncertainty/information** about **(Bayesian) statements** such as $(\theta \leq q)$ is **justified**, without appeal to **betting odds**.

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

• 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: **CDFs** and **densities arise naturally** when $\Theta = \Re^k$ for $1 \leq k < \infty$,

• 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**:

Optimal Inference, Prediction and Decision

(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**:

$$p(\theta|D\,\mathcal{B}) = c\,p(\theta|\mathcal{B})\,p(D|\theta\,\mathcal{B})\,, \qquad \textbf{(inference)} \qquad (1)$$

where $c > 0$ is a **normalizing constant** chosen so that the **left-hand side** of (1) **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:

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

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

$$p(D^*|D\,\mathcal{B}) = \int_\Theta p(D^*|\theta\,\mathcal{B})\,p(\theta|D\,\mathcal{B})d\theta\,; \qquad \textbf{(prediction)} \qquad (2)$$

(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 $p(\theta|D\,\mathcal{B})$:

$$a^* = \operatorname*{argmax}_{a\in\mathcal{A}} E_{(\theta|D\,\mathcal{B})}\,U(a,\theta) = \operatorname*{argmax}_{a\in\mathcal{A}} \int_\Theta U(a,\theta)\,p(\theta|D\,\mathcal{B})\,d\theta\,. \qquad (3)$$

The Specification Burden

These **corollaries** to **Cox's theorem** leave **no ambiguity** about how to draw **inferences**, and make **predictions** and **decisions**, in the presence of **uncertainty** — but 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**);

- $p(D|\theta\,\mathcal{B})$, 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

The Specification Burden (continued)

(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**:

Theory of Applied Statistics

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 **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.

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 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**.

The Modeling-As-Decision Principle

This **acknowledgment** of Your **model uncertainty** implies a **willingness** by You to **consider two or more models** in an **ensemble** $\mathcal{M} = \{M_1, M_2, \ldots\}$, and this in turn **creates** a **need** to **answer two types** of **questions**:

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

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

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

This solves the model specification problem with no ad-hockery (**Modeling-As-Decision Principle** $\rightarrow$ **Theorem**: **optimal model specification** via **MEU**), and there are **examples** of its use (e.g., Draper and Fouskakis, *JASA*, 2008: **variable selection** in

generalized linear models under **cost constraints**), but this is **hard work**; there's a **powerful desire** for **generic model-comparison methods** whose **utility structure** may provide a **decent approximation** to **problem-specific utility elicitation**.

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

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

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

the **first term** on the right is the **prior odds** in favor of M_2 over M_1, and the **second term** on the right is the **Bayes factor**.

This **approach** does have a **decision-theoretic basis**, but it **may not be a good approximation** to Your **problem-specific utility elicitation**: if You **take the view** that there is an **underlying data-generating mechanism** M_{DG} (this connects with **calibration**: see below), and You **pretend** that one of the **models** in Your **ensemble** $\mathcal{M} = \{M_1, M_2, \ldots,\}$ must be M_{DG}, and You **pretend** that the **utility function**

Integrated Likelihood

$$U(M, M_{DG}) = \left\{ \begin{array}{cc} 1 & \text{if } M = M_{DG} \\ 0 & \text{otherwise} \end{array} \right\} \tag{5}$$

reflects Your **real-world values**, then it's **decision-theoretically optimal** to **choose** the **model** in $\mathcal{M}$ with the **highest posterior probability**, and if (e.g.) it's **scientifically appropriate** to take the **prior model probabilities** $p(M_j|\mathcal{B})$ to be **equal**, this involves **maximizing Bayes factors** over the **models** in $\mathcal{M}$.

Moreover, as is **well known**, in **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

$$M_j : \left\{ \begin{array}{c} (\gamma_j|M_j\,\mathcal{B}) \sim p(\gamma_j|M_j\,\mathcal{B}) \\ (D|\gamma_j\,M_j\,\mathcal{B}) \sim p(D|\gamma_j\,M_j\,\mathcal{B}) \end{array} \right\}, \tag{6}$$

the **integrated likelihood**

$$p(D|M_j\,\mathcal{B}) = \int p(D|\gamma_j\,M_j\,\mathcal{B})\,p(\gamma_j|M_j\,\mathcal{B})\,d\gamma_j\,; \tag{7}$$

at the **heart** of the **Bayes factor** can be **written**

BIC

$$p(D|M_j\,\mathcal{B}) = E_{(\gamma_j|M_j\,\mathcal{B})}\; p(D|\gamma_j\,M_j\,\mathcal{B})\,, \tag{8}$$

and is thereby seen to be the expectation of the sampling distribution over the **prior** for γ_j in model M_j (evaluated at the **observed data set** D); when there is **little information** about the γ_j **external** to D, motivating **diffuse priors** on the **parameter vectors**, this makes **Bayes factors extremely sensitive** to **small details** in how the **diffuseness** is **specified**.

This has **given rise** to an **amazing amount** of **ingenuity** and **ad hockery**, in **equal measure**, as people have tried to **fix the problem**; my favorite **Bayes factor fix** is **BIC**, which has a **sensible implicit diffuse prior**, as follows: with **sample size** n the usual **Laplace approximation** to the **log integrated likelihood** is

$$\begin{aligned}\log p(D|M_j\,\mathcal{B}) \;=\; & \log p(D|\hat{\gamma}_j\,M_j\,\mathcal{B}) + \log p(\hat{\gamma}_j|M_j\,\mathcal{B}) \\ & + \frac{k_j}{2}\log 2\pi - \frac{1}{2}\log|\hat{I}_j| + O\left(\frac{1}{n}\right)\,,\end{aligned} \tag{9}$$

where $\hat{\gamma}_j$ 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.

Unit-Information Prior

Schwarz (1978) used a **less precise Taylor expansion** to obtain

$$\log p(y|M_j\,\mathcal{B}) = \log p(y|\hat{\gamma}_j\,M_j\,\mathcal{B}) - \frac{k_j}{2}\log n + O(1)\,; \tag{10}$$

as usual a **multiple** of this is **often used** for **model comparison**:

$$BIC(M_j|D\,\mathcal{B}) = -2\log p(D|\hat{\gamma}_j\,M_j\,\mathcal{B}) + k_j \log n\,, \tag{11}$$

in which **models** with **small BIC values** are **preferred**.

You can now **work out** what **implied prior BIC is using**, from the point of view of the **Laplace approximation**; the result is

$$(\gamma_j|M_j\,\mathcal{B}) \sim N_{k_j}(\hat{\gamma}_j, n\hat{I}_j^{-1}). \tag{12}$$

This is 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).

Log Scores

- **Log scores** are **based on** a

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)

$$M_j\text{:} \left\{ \begin{array}{c} (\gamma_j | M_j\, \mathcal{B}) \sim p(\gamma_j | M_j\, \mathcal{B}) \\ (y | \gamma_j\, M_j\, \mathcal{B}) \sim p(y | \gamma_j\, M_j\, \mathcal{B}) \end{array} \right\} . \tag{13}$$

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

Full-Sample Log Score

composite measure of **predictive accuracy** of M_j is a **cross-validated** version (e.g., Gelfand and Dey, 1994) of the resulting **log score**,

$$LS_{CV}(M_j|y\,\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y_{-i}\,M_j\,\mathcal{B})\,, \tag{14}$$

in which y_{-i} is the y **vector** with observation i **omitted**.

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

$$LS_{FS}(M_j|y\,\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y\,M_j\,\mathcal{B})\,, \tag{15}$$

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).

Full-sample log scores have a **direct decision-theoretic basis**, and if, in the spirit of **calibration** (see below), You're prepared to **think about** an **underlying data-generating model** M_{DG}, LS_{FS} also has a

Calibration Principle

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**:

$$\begin{aligned} KL[M_{DG}||p(\cdot|y\, M_j\, \mathcal{B})] &= E_{M_{DG}} \log M_{DG} - E_{M_{DG}} \log p(\cdot|y\, M_j\, \mathcal{B}) \\ &\doteq E_{M_{DG}} \log M_{DG} - LS_{FS}(M_j|y\, \mathcal{B})\,; \qquad (16) \end{aligned}$$

the **first term** on the **right side** of (16) 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}.

- Earlier I mentioned that **Cox's Theorem** also **fails to address** another **important aspect** of **scientific Bayesian modeling**; it needs to be **supplemented** by a

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:

Reasoning Behind the Calibration Principle

(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.

(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**.

Bayes Factors/BIC Versus Log Scores

There's a **new idea** here: **logical consistency** justifies **Bayesian uncertainty assessment** but **does not provide guidance** on **model specification**; under 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**.

- What follows is a **sketch** of **recent results** (Draper, 2011) based on **calibration 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**.

Clinical Trial to Quantify Improvement

• **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**.

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

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

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

Decision, Not Inference

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 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.

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**).

Models For Quantifying Improvement

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, **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**.

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

$$M_1: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } \theta \leq \Delta \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{17}$$

$$M_2: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } \theta > \Delta \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{18}$$

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

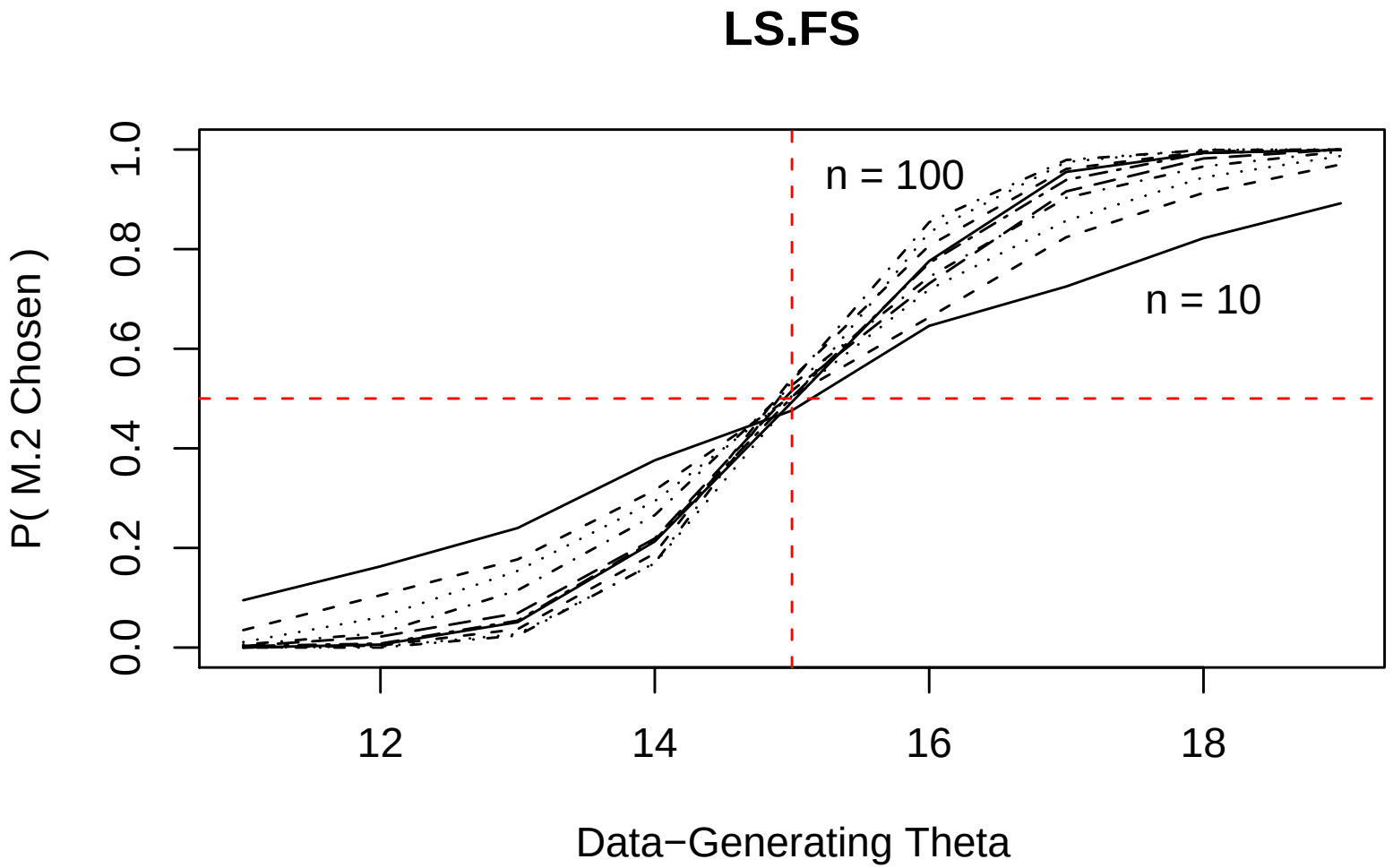
- **Posterior probability**: let $M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \Re, (y_i|\theta\,\mathcal{B}) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)\}$ and **choose** M_2 if $p(\theta > \Delta|y\,M^*\,\mathcal{B}) > 0.5$.

- **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**.

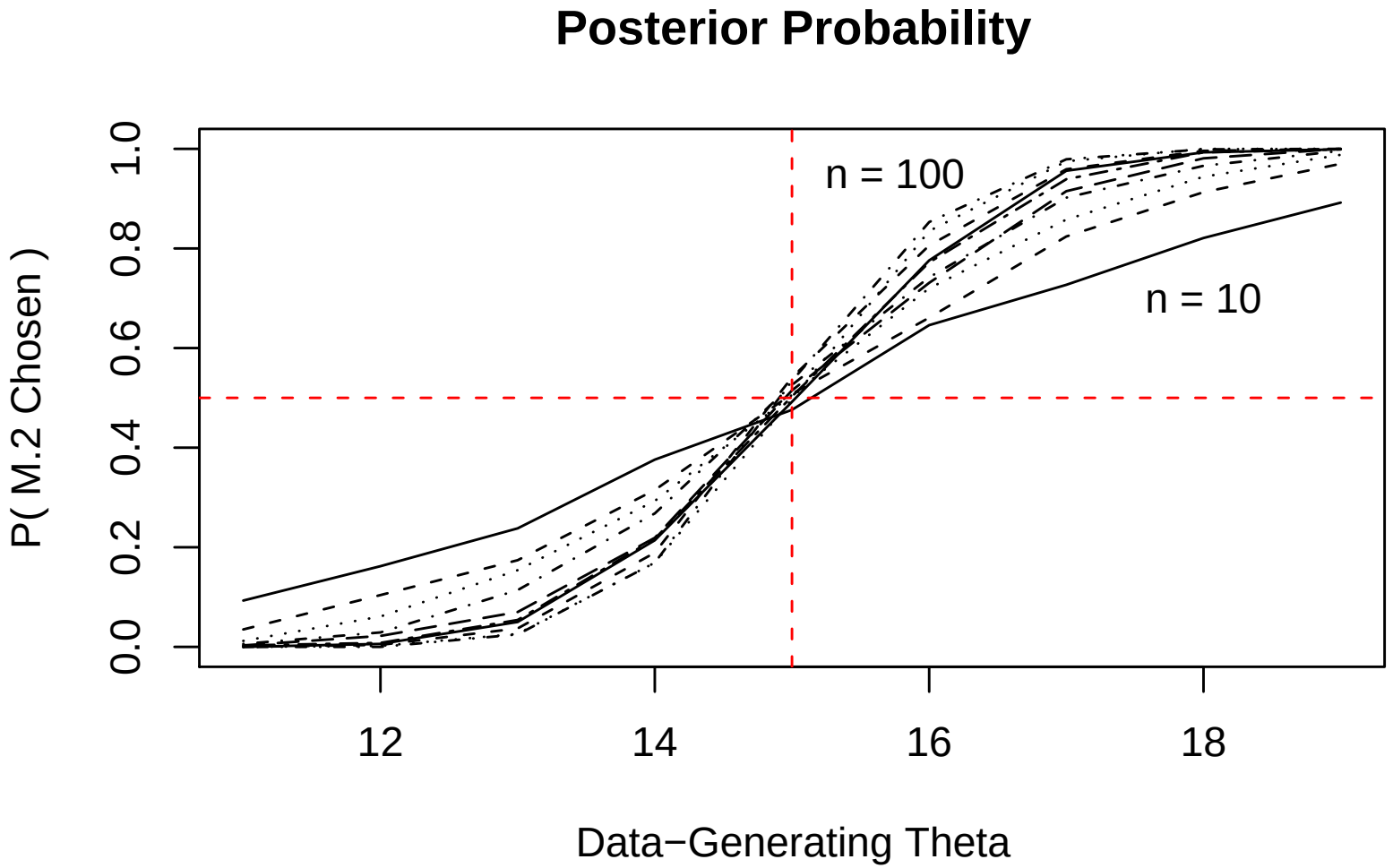
The **figures** below give **Monte-Carlo estimates** of the **probability that M_2 is chosen**.

LS_{FS} Results: Quantifying Improvement



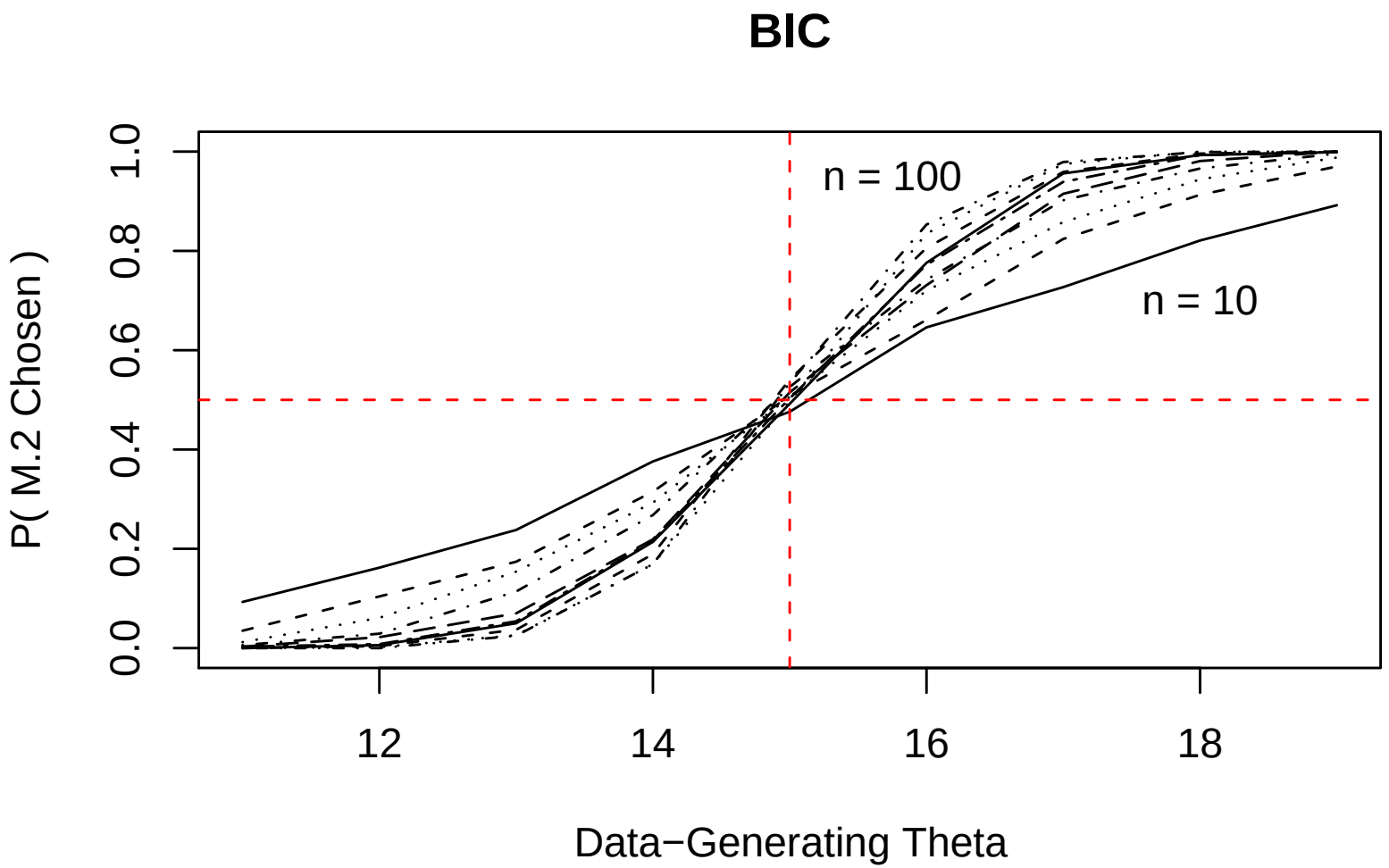
This **exhibits all** the **monotonicities** that it **should**, and **correctly yields 0.5** for all n with $\theta_{DG} = 15$.

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**

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \qquad (19)$$

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

$$M_3: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{20}$$

$$M_4: \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{21}$$

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

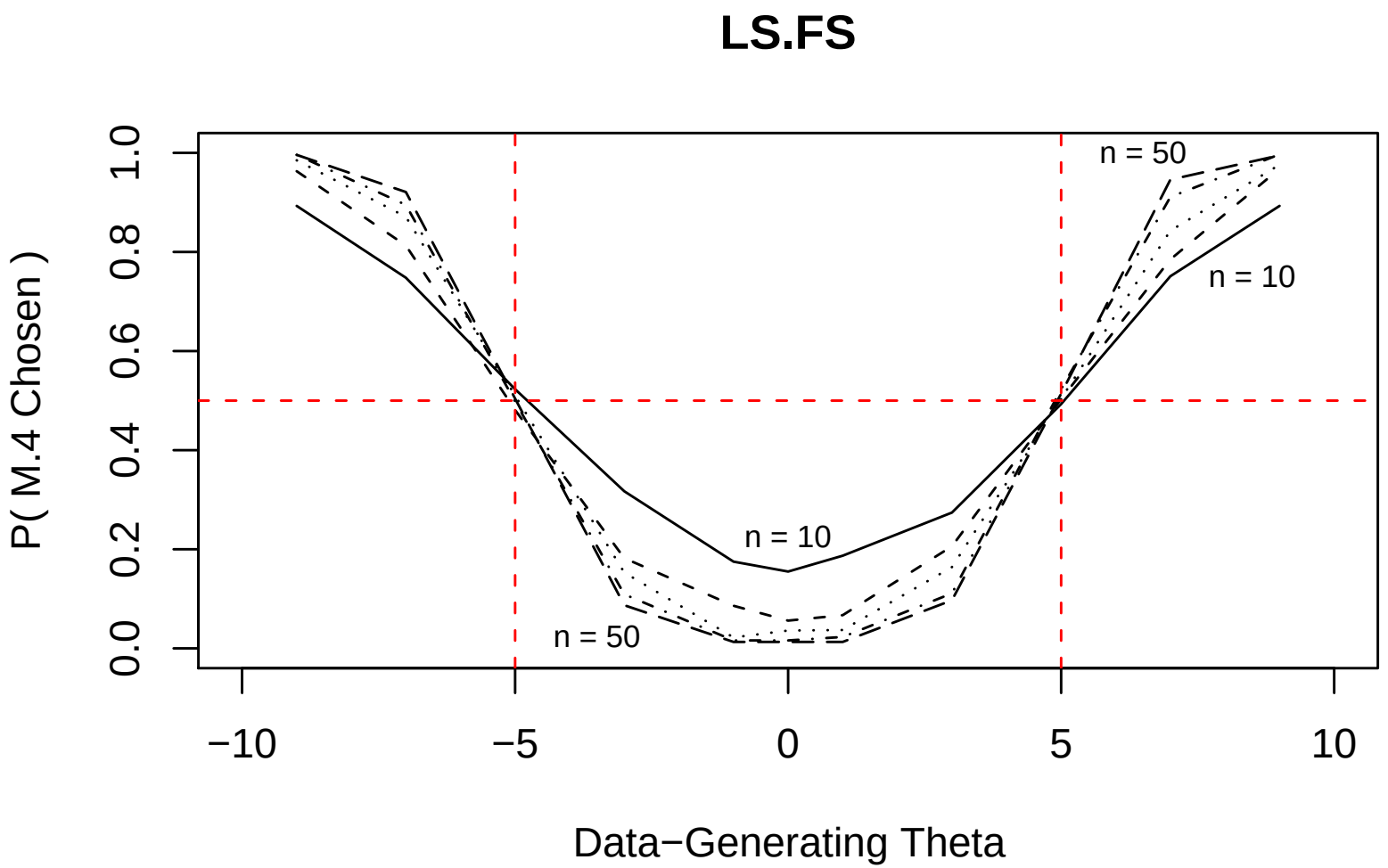
{**claiming** the two drugs were **bio-equivalent** when they **were**,
concluding that they were **bio-equivalent** when they **were not**,
deciding that they were **not bio-equivalent** when they **were**,
judging that they were **not bio-equivalent** when they were **not**},

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}.

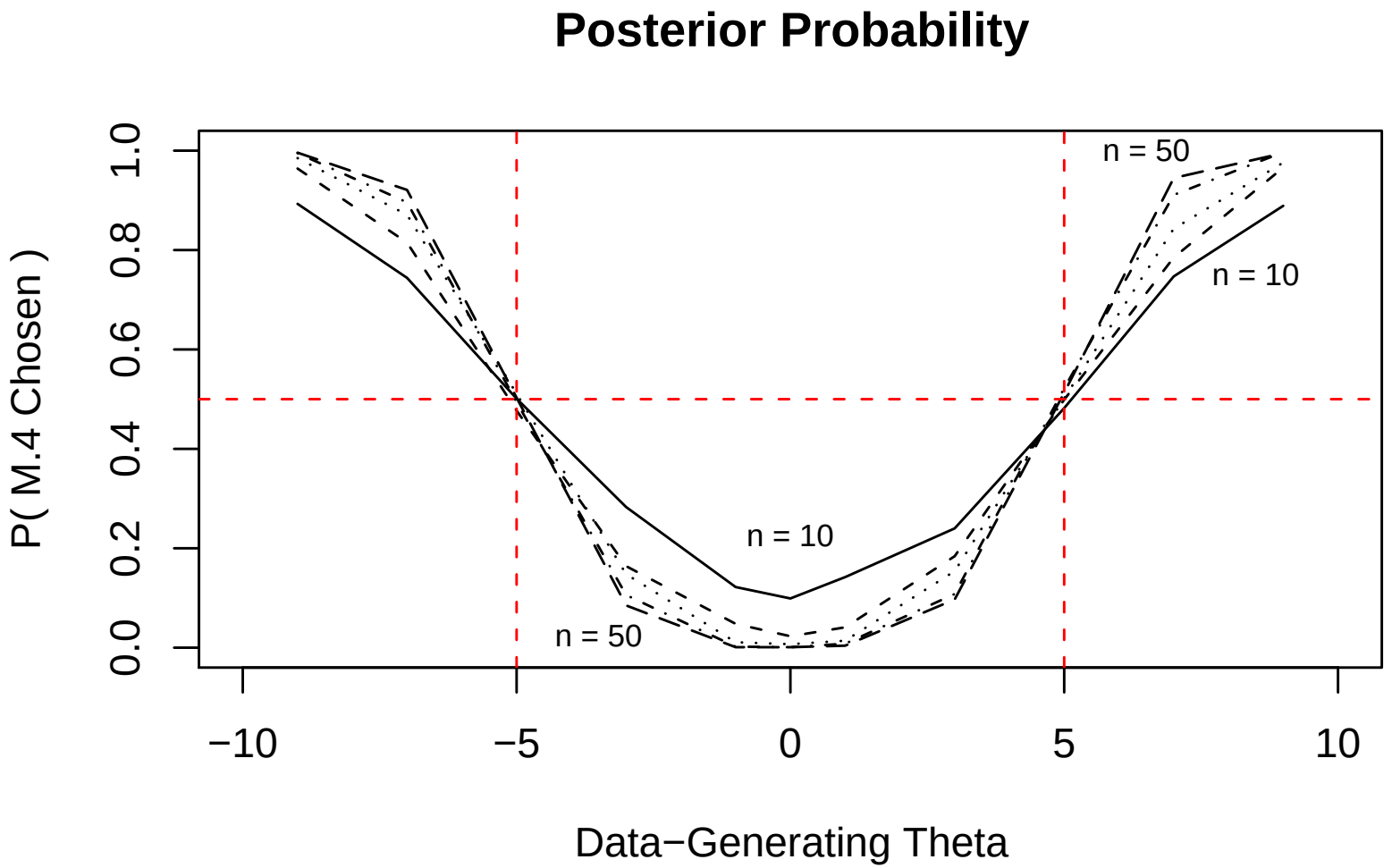
NB It has **previously been established** that when **making** the **(unrealistic) sharp-null comparison** $\theta = 0$ versus $\theta \neq 0$ in the **context** of $(y_i|\theta\, \mathcal{B}) \overset{\text{IID}}{\sim} N(\theta, \sigma^2)$, as $n \to \infty$ LS_{FS} **selects** the $\theta \neq 0$ **model**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



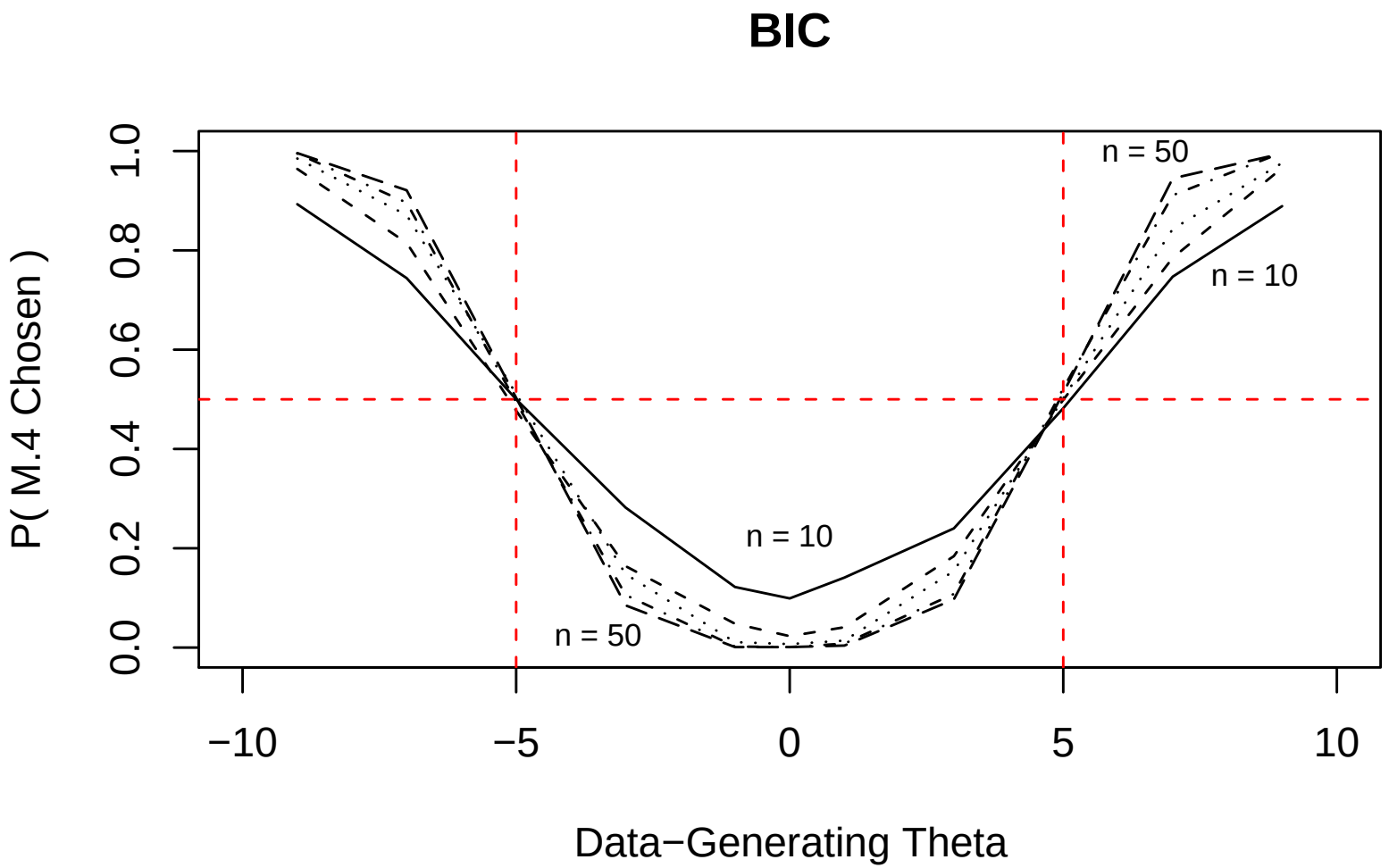
In this **more realistic setting**, comparing $|\theta| \leq \lambda$ versus $|\theta| > \lambda$ with $\lambda > 0$, LS_{FS} has the **correct large-sample behavior**, **both** when $|\theta_{DG}| \leq \lambda$ and when $|\theta_{DG}| > \lambda$.

Posterior Probability Results: Bio-Equivalence



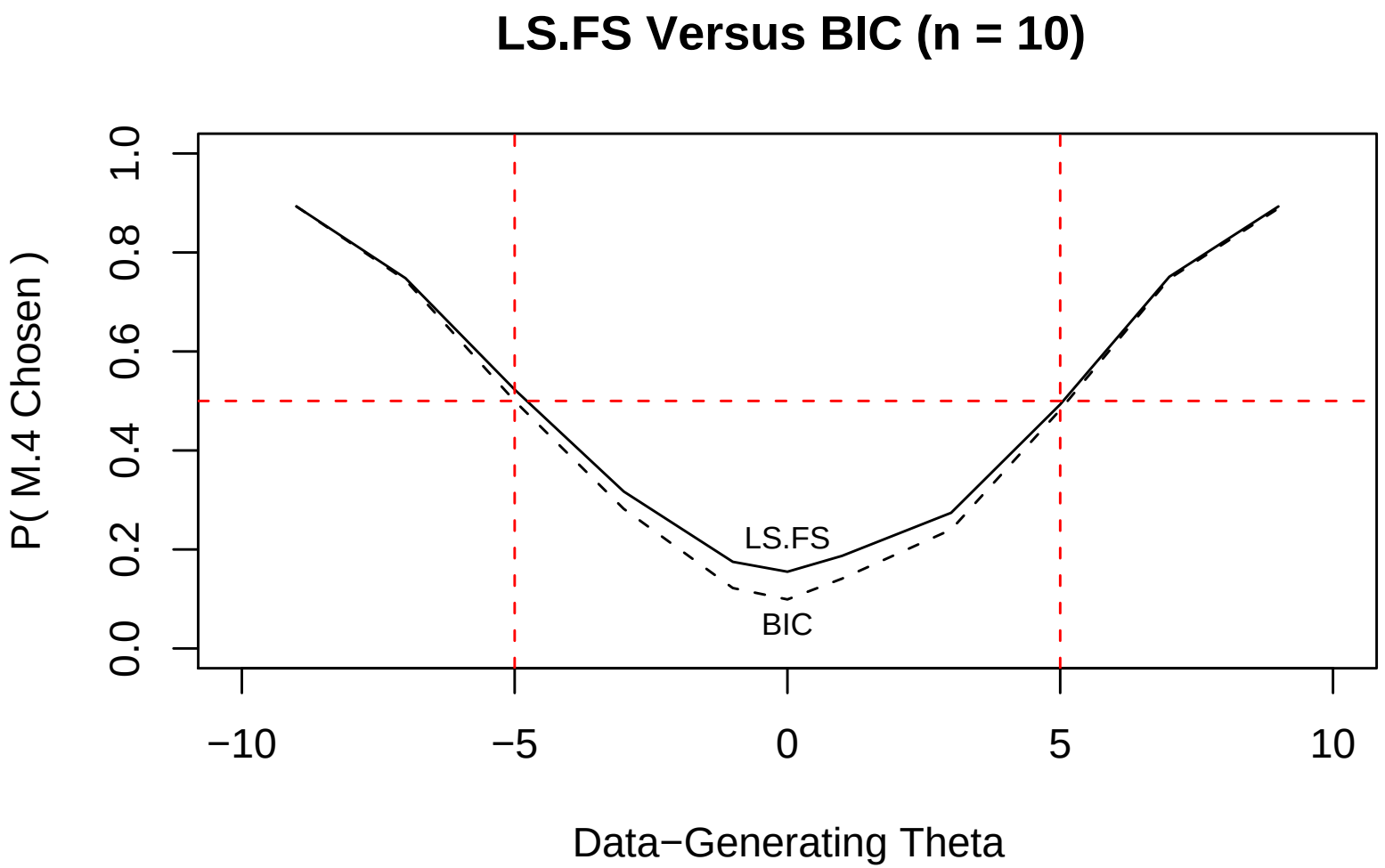
The **qualitative behavior** of the LS_{FS} and **posterior-probability methods** is **identical**, although there are some **numerical differences** (**highlighted** later).

BIC Results: Bio-Equivalence



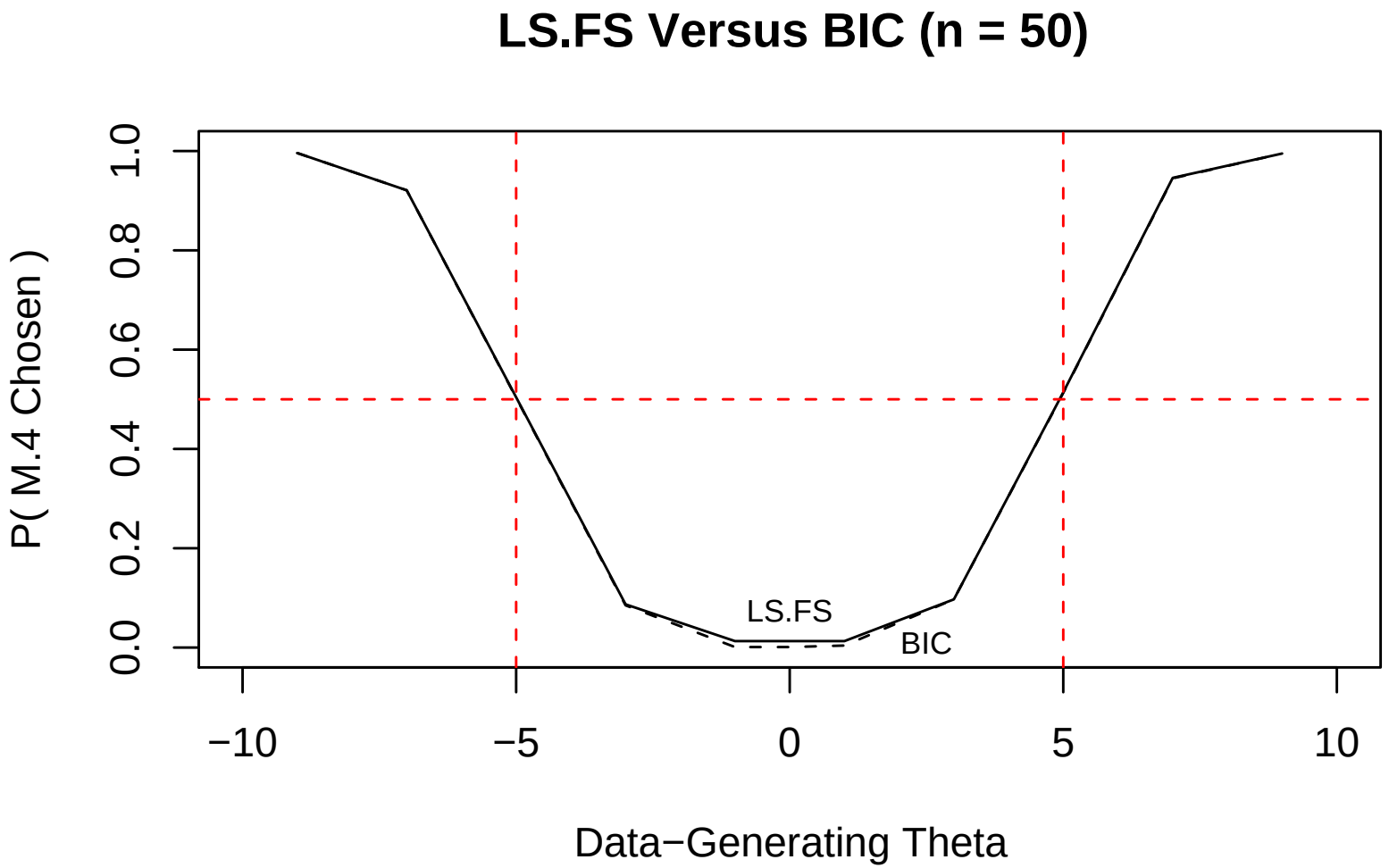
In the **quantifying-improvement** case, the **BIC** and **posterior-probability methods** were **algebraically identical**; here they **nearly coincide** (**differences** of $\pm$**0.001** with **1,000 simulation repetitions**).

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: $\theta \neq 0$.

Sharp-null testing is **frequently unwise** because

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

(b) it's **silly** from a **measurement point of view**: with a **(conditionally) IID** $N(\theta, \sigma^2)$ **sample** of size n, your **measuring instrument** $\bar{y}$ is only **accurate** to **resolution** $\frac{\sigma}{\sqrt{n}} > 0$; **claiming** to be **able** to **discriminate** between $\theta = 0$ and $\theta \neq 0$ — 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.

Testing Sharp-Null Hypotheses (continued)

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

$$M_5\text{:} \left\{ \begin{array}{ccc} (\sigma^2|\mathcal{B}) & \sim & \text{diffuse on } (0, \text{large}) \\ (y_i|\sigma^2\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(0, \sigma^2) \end{array} \right\} \quad \text{and} \qquad (22)$$

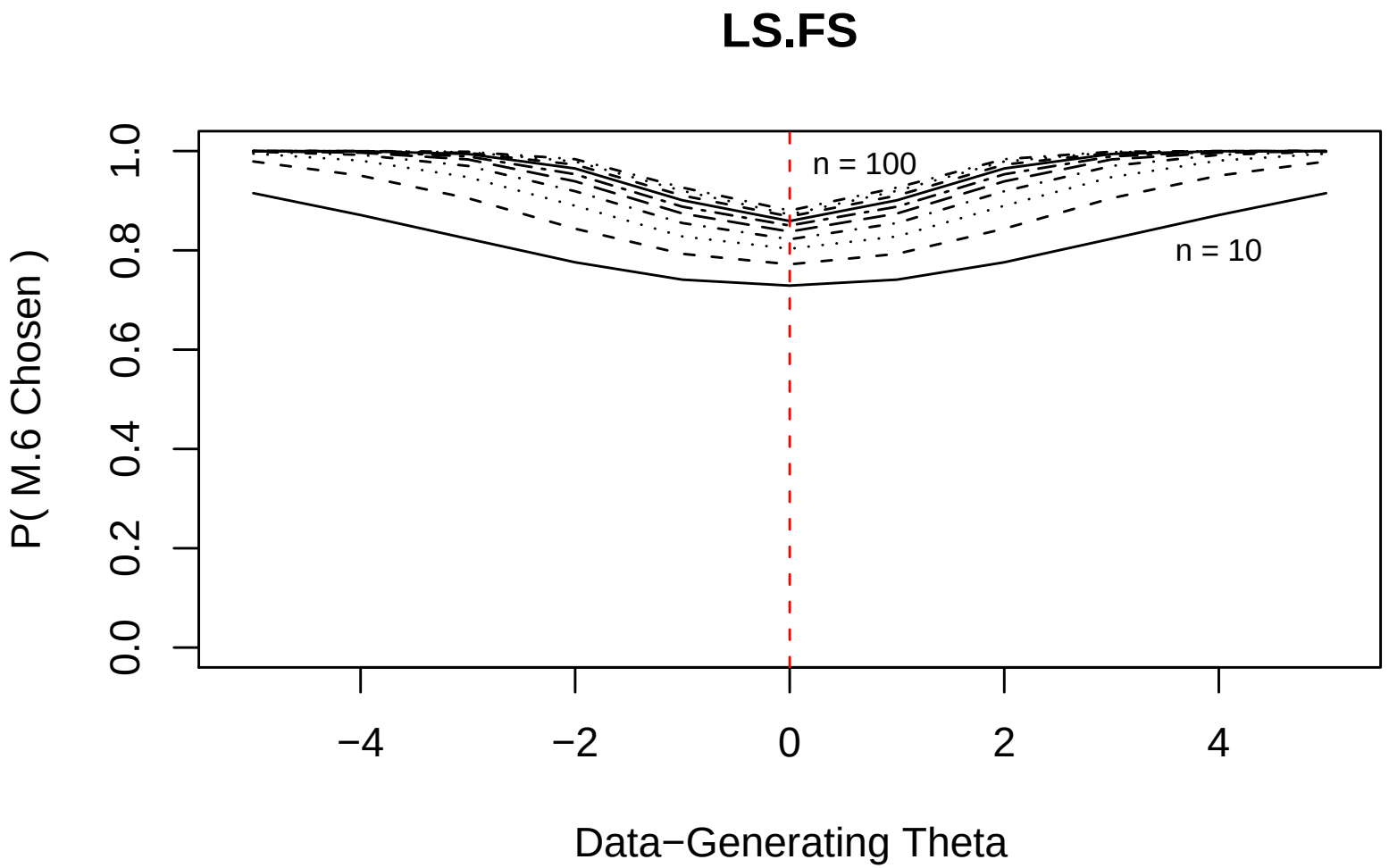
$$M_6\text{:} \left\{ \begin{array}{ccc} (\theta\,\sigma^2|\mathcal{B}) & \sim & \text{diffuse on } (-\text{large}, \text{large}) \times (0, \text{large}) \\ (y_i|\theta\,\sigma^2\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \quad (23)$$

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**.

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**.

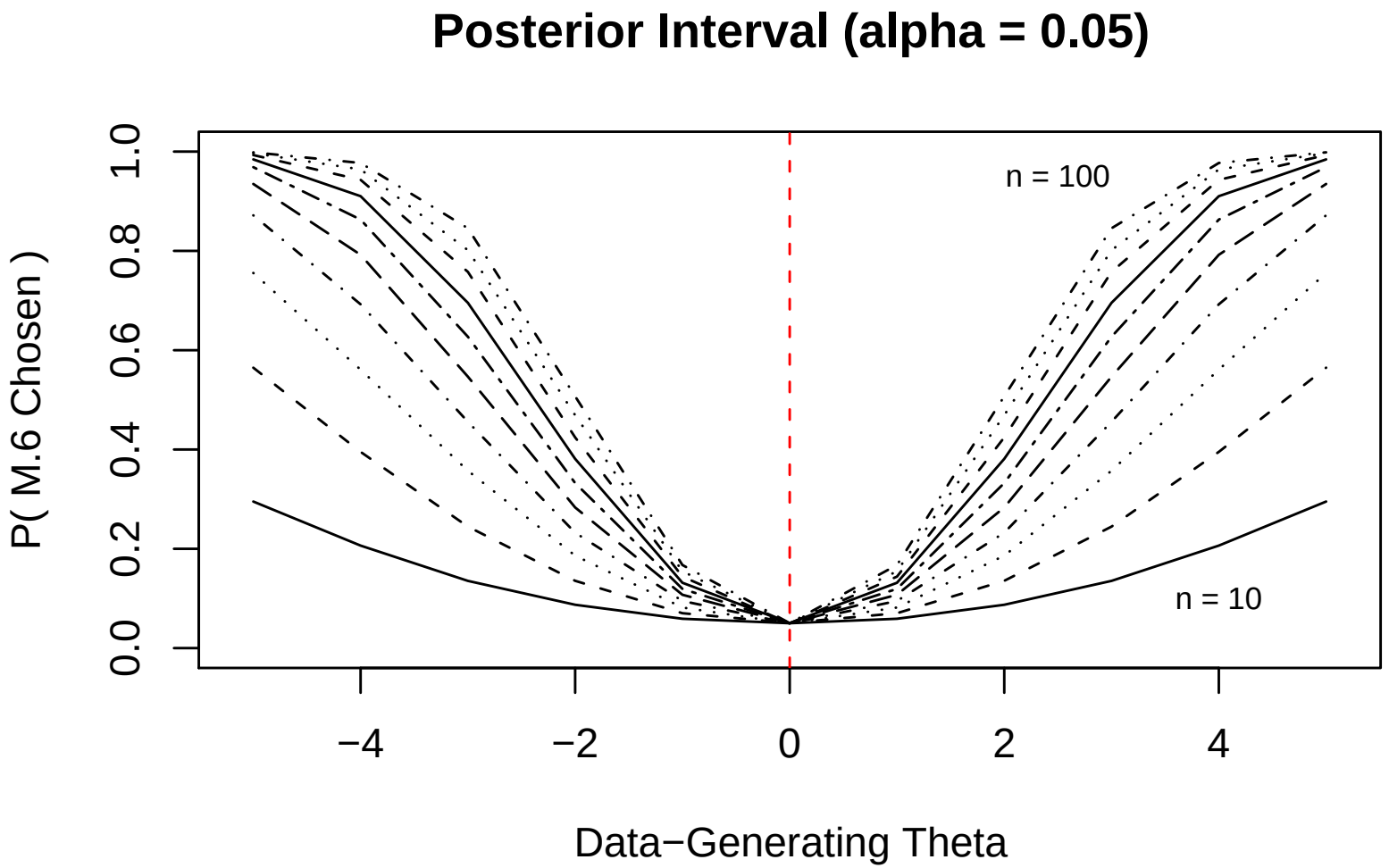
As before, let's call **choosing** M_6: $\theta \neq 0$ when $\theta_{DG} = 0$ a **false-positive** error and **choosing** M_5: $\theta = 0$ when $\theta_{DG} \neq 0$ a **false-negative** mistake.

LS_{FS} Results: Sharp-Null Testing



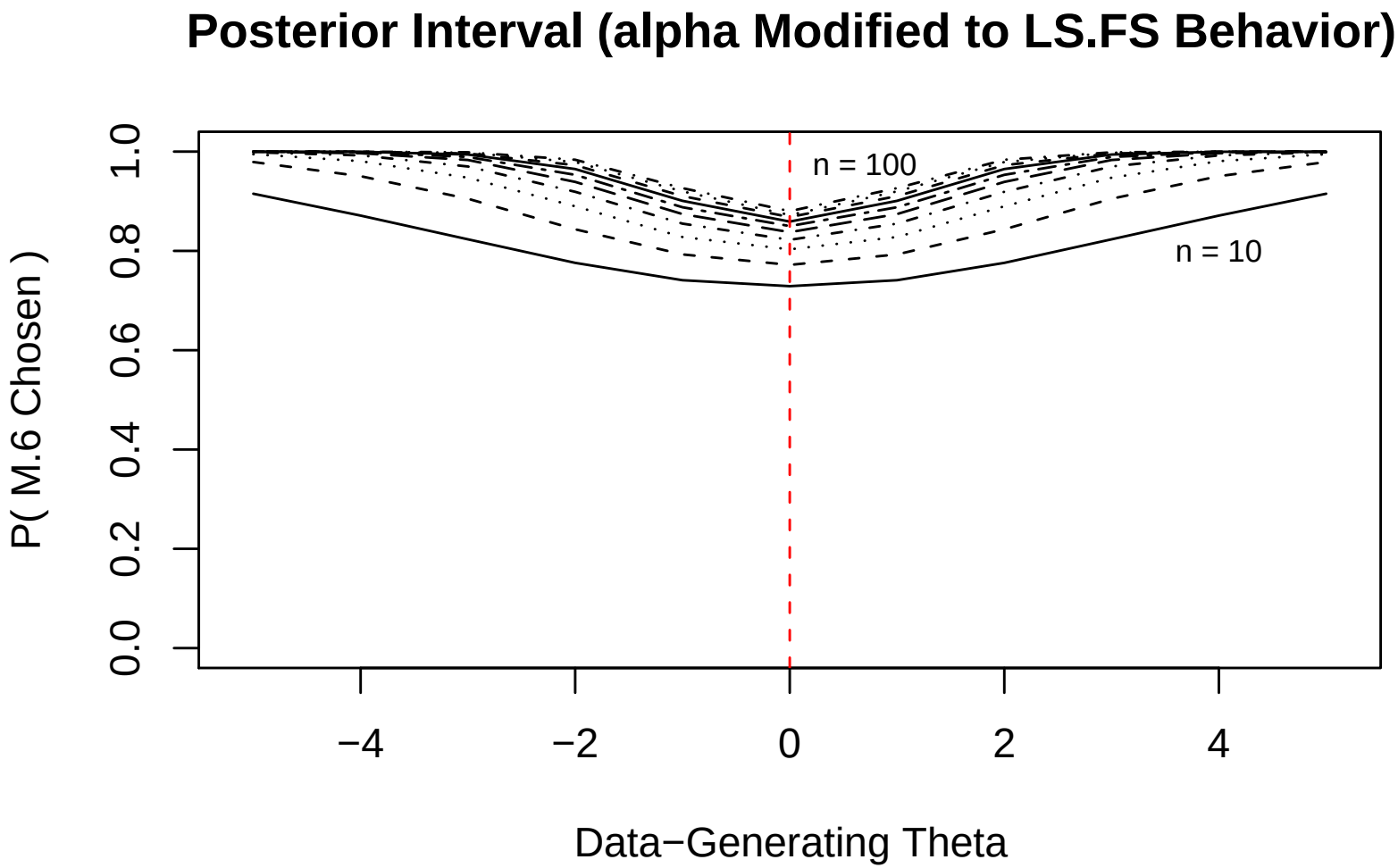
In the **limit** as $\lambda \downarrow 0$, the LS_{FS} **approach** makes **hardly any false-negative errors** but **quite a lot of false-positive mistakes**.

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



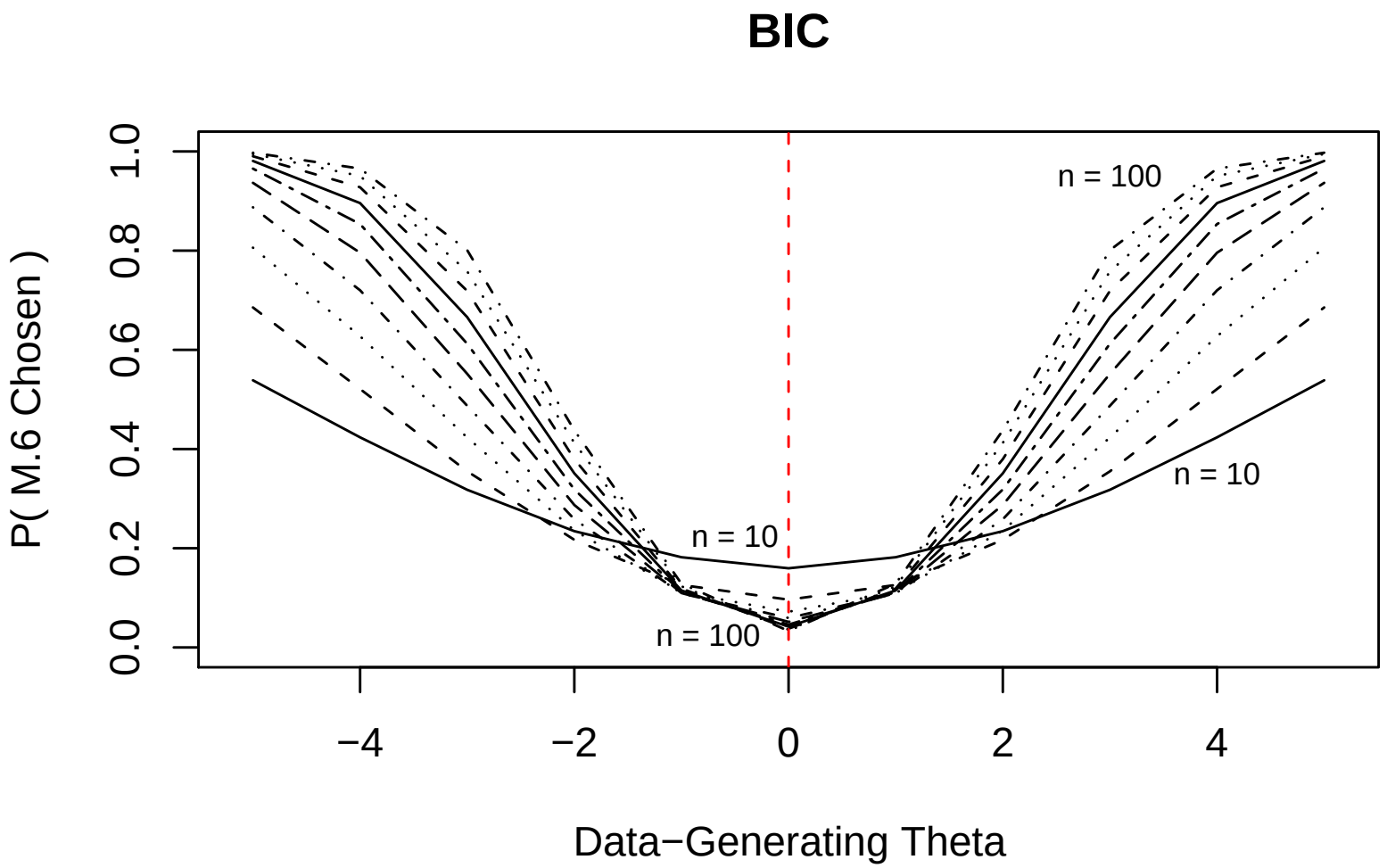
The **behavior** of the **posterior interval approach** is of course **quite different**: it makes **many false-negative errors** because its **rate of false-positive mistakes** is **fixed at 0.05**.

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



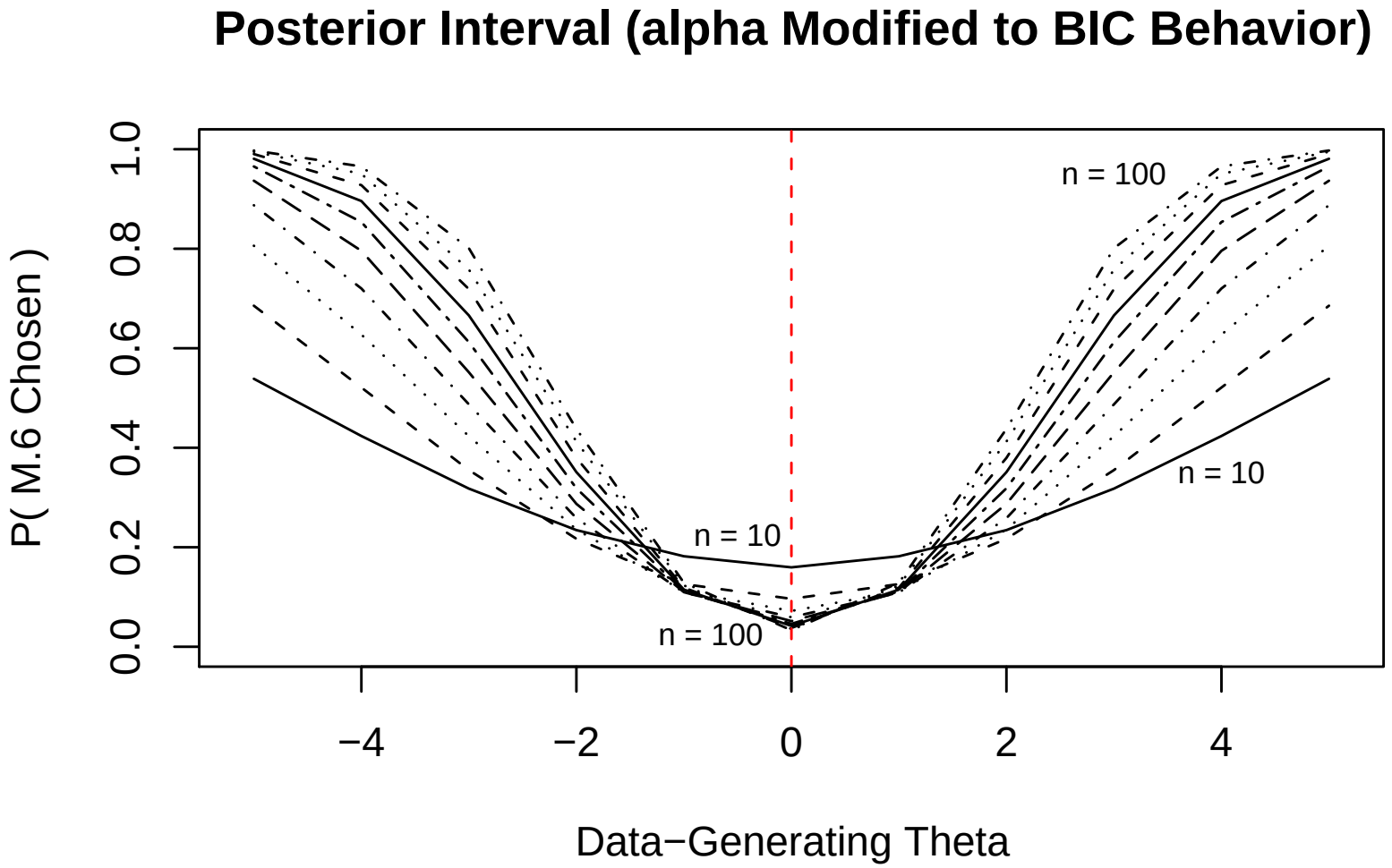
When the **interval method** is **modified** so that α **matches** the LS_{FS} **behavior** at $\theta_{DG} = 0$ (letting α **vary** with n), the **two approaches** have **identical model-discrimination ability**.

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 $\hat{\theta}_1 = \frac{\alpha_1 + n}{\alpha_1 + \beta_1 + s + n}$,

$$\begin{aligned} LS_{FS}(M_7|y\,\mathcal{B}) \quad = \quad & \log\Gamma(\alpha_1 + n + \beta_1 + s) + \log\Gamma(\alpha_1 + n + 1) \\ & - \log\Gamma(\alpha_1 + n) - \log\Gamma(\beta_1 + s) \qquad (24) \\ & + \frac{1}{n}\sum_{i=1}^{n}[\log\Gamma(\beta_1 + s + y_i) \\ & - \log\Gamma(\alpha_1 + n + \beta_1 + s + y_i + 1)]\,, \end{aligned}$$

$$BIC(M_7|y\,\mathcal{B}) = -2[n\log\hat{\theta}_1 + s\log(1 - \hat{\theta}_1)] + \log n\,, \qquad (25)$$

Geometric Versus Poisson (continued)

$$
\begin{aligned}
LS_{FS}(M_8|y\,\mathcal{B}) \;=\; & (\alpha_2+s)\log(\beta_2+n) - \log\Gamma(\alpha_2+s) \\
& -(\alpha_2+s)\log(\beta_2+n+1) \qquad (26) \\
& +\frac{1}{n}\sum_{i=1}^{n}[\log\Gamma(\alpha_2+s+y_i) - y_i\log(\beta_2+n+1) \\
& -\log\Gamma(y_i+1)]\,, \text{ and}
\end{aligned}
$$

$$
BIC(M_8|y\,\mathcal{B}) = -2[s\log\hat{\theta}_2 - n\,\hat{\theta}_2 - \sum_{i=1}^{n}\log(y_i!)] + \log n\,, \qquad (27)
$$

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

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

$$\begin{aligned} p(y_i|y\,M_j\,\mathcal{B}) &= \int p(y_i|\gamma_j\,M_j\,\mathcal{B})\,p(\gamma_j|y\,M_j\,\mathcal{B})\,d\gamma_j \qquad (28)\\ &= E_{(\gamma_j|y\,M_j\,\mathcal{B})}p(y_i|\gamma_j\,M_j\,\mathcal{B})\,; \end{aligned}$$

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

$$n\, LS_{FS}(M_j|y, \mathcal{B}) > n\, LS_{FS}(M_{j'}|y, \mathcal{B}). \tag{29}$$

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

$$nLS_{FS}(M_j|y,\mathcal{B}) = \log \prod_{i=1}^{n} \left[\int p(y_i|\eta_j, M_j, \mathcal{B})\, p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j \right], \quad (30)$$

and this is **not the same as**

$$\log \int \left[\prod_{i=1}^{n} p(y_i|\eta_j, M_j, \mathcal{B}) \right] p(\eta_j|y, M_j, \mathcal{B})\, d\eta_j = \bar{L}_j^{PBF} \quad (31)$$

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**

$$\begin{aligned} (y_i|\lambda_i, \mathcal{B}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log \lambda_i &= \beta_0 + \beta_1 x_i + e_i \\ (e_i|\sigma_e^2, \mathcal{B}) &\overset{\text{IID}}{\sim} N(0, \sigma_e^2)\,, \quad (\beta_0, \beta_1, \sigma_e^2) \sim \text{diffuse}\,, \end{aligned} \tag{32}$$

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: $P_{\theta_{DG}=0}(LS_{FS}$ chooses $\theta = 0) \to 0$ as $n \to \infty$, it's **also true** that

New Theorem (Draper, 2011): for any $\lambda > 0$,
$P_{|\theta_{DG}| \leq \lambda}(LS_{FS}$ chooses $|\theta| \leq \lambda) \to 1$ 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**, and the **Prediction Principle** seem **helpful** in **moving toward this goal**.